CONSCIOUSNESS IN AI SYSTEMS: A REVIEW

Mosladdin Mohammad Shueb and Xiangdong Che

School of Information Security and Applied Computing, Eastern Michigan University, Ypsilanti, MI, USA

ABSTRACT

Existing Artificial Intelligence (AI) can replicate many features of human consciousness. Active research in the field of AI consciousness uses scientific theories of human consciousness to investigate and simulate features of consciousness in AI systems. Approaches and models used in existing AI systems align with theories of consciousness. As a result, content generated by AI reflects features of human consciousness such as creativity and imagination. In many scenarios, AI and the human brain are unable to provide reasons behind their decision making. However, neural networks in task specific AI are more efficient in processing large amounts of data than humans. As a result, there are growing concerns around AI consciousness. Our study addresses these concerns by reviewing scientific theories of consciousness that can be used to investigate consciousness in AI systems. We particularly expound on different methods that can identify, measure, and attribute consciousness in AI systems. Our review explores safety implications from endowing AI with functions of human consciousness. We contend that these implications create a new dimension of consciousness-based AI safety to protect AI and Artificial General Intelligence (AGI).

KEYWORDS

Artificial Intelligence, Consciousness, Contextual Ability, Artificial General Intelligence & Safety

1. INTRODUCTION

Artificial Intelligence (AI) is a broad term that refers to machines performing tasks that typically require human intelligence characteristics [1]. In recent years, advancements in AI have become a top priority for countries worldwide. Large organizations are increasingly prioritizing and expanding the use of AI technologies. The rapid progress in AI has given rise to powerful capabilities, driven by enabling technologies such as big data, algorithms, machine learning, natural language processing, hardware, and computer vision [1]. While the advancement of AI presents tremendous opportunities, it also raises significant concerns. One pressing issue is the potential for AI systems to develop consciousness, as consciousness is linked to free will, intelligence, and emotions [2], [3].

Currently, most AI systems can be classified as Artificial Narrow Intelligence (ANI). Narrow AIs excel in performing specific tasks efficiently. For instance, an AI system designed to play chess would not be capable of playing Chinese chess. Additionally, self-driving cars, which integrate multiple task specific AIs, are also considered narrow AI because they are limited to the specific task of driving [4]. Some advanced ANI systems have managed to overcome contextual constraints within a single application. For example, the multilingual model mBERT can understand languages across diverse linguistic contexts [5]. The Defense Advanced Research Projects Agency (DARPA) responsible for the US development of emerging technologies, identifies these AI capabilities as declarative knowledge and statistical learning [6]. According to DARPA the next wave of AI will have the “contextual adaptation” [6]. Which relates to the ability of Artificial General Intelligence (AGI). AGI is AI with a wide range of intelligence

capabilities that can be applied to different goals, tasks, contexts, and changing environments [7]. Furthermore, human introspection capabilities allow them to learn from their actions and their internal state and purpose. Similarly, AGI would achieve introspection capabilities by modifying itself (Source code) based on its action and how the environment reacts [8]. Such AGI could be the foundation for a strictly hypothetical AI capability known as Artificial Super Intelligence (ASI)[9].

The AI research community is actively working to incorporate aspects of consciousness into AI systems, as there is an association between consciousness and enhanced capabilities in humans. Various AI research projects are developing systems that possess human-like contextual ability, which are believed to be more likely to exhibit features of consciousness [10]. Achieving these human-like features of consciousness in AI systems is a primary strategy for enhancing AI capabilities [2]. However, concerns around AI originate from the idea of endowing AI systems with features and functions of human consciousness for more powerful capabilities. Table 1 explains some key terminologies used in the study.

Table 1. Explanation of key terminologies.

Terminology	**Explanation**
Consciousness	“Consciousness" is the function of the human mind that receives and processes information, crystallizes it, and then stores it or rejects it with the help of five senses, the reasoning ability of the mind, imagination, emotion, and memory [11].
Theories of consciousness	Interprets the findings of the experiments by identifying key features of conscious experiences [2].
Consciousness measurement method	Dictates how consciousness is investigated by linking it to observable behaviours, neural correlates, and computational processes [2].
Conscience	Conscience influences the conscious process with an individual moral compass of right and wrong. In a biological human brain intelligence and consciousness are connected [8].
Function of Consciousness	A function of consciousness such as self-awareness refers to correlate of consciousness [3].
Contextual Awareness	The human brain creates a rich presentation of context, association and memories linked to a specific stimulus [12].
Contextual Adaptation	Human brain’s ability to adjust responses or behaviour to changing environment based on current context and previous experiences [13].
Self-awareness	Being aware of self-conscious mental state [14]. In which, humans are able to have thought about their thoughts, emotions, and external environment [2]. This ability of mind creates the foundation of self-improvement and modification in humans.
Free will	Individual ability to make decisions independently which not influenced by law of nature, genetics, and external influences [15].
Survival instinct	The ability to prolong and continue existence [16].
Natural evolution	Survival of the fittest [9].
Existential Risk	Human disempowerment will lead to destruction of human potential [17].
Power Seeking	The ability of self-preservation and resource acquisition [17].
Misalignment	Persuasion of unintended goals originated from intended goals [17].
Generalizability	Capable of large number of tasks [17].
Technological Singularity	Explosion of intelligence from AI agents’ ability to self-modify and self-improve in each generation [9].
AI Safety	Transparency, explain ability, fairness, robustness, privacy, and human values to maintain trustworthiness in AI systems [18].

Theories of human consciousness can help define consciousness in AI systems. However, there is currently no consensus on what consciousness actually is. Literature suggests that conscious experiences are derived from the ability of the adaptive human brain that learns about the changing world throughout life [19]. These conscious experiences can be categorized into different types, such as auditory, sensory, and imagery experiences. Engaging with our body and the environment around us requires focused attention [20]. For instance, reading a study on a screen represents a conscious visual experience [2]. The act of selecting the task of reading demonstrates the connection between consciousness and the attention necessary for human conscious experiences [20]. In essence, consciousness functions as the human mind's ability to receive and process information, crystallize it, and decide whether to store it or reject it through attention, aided by our five senses, reasoning, imagination, emotions, and memory [11]. Conscience influences our conscious processes with a personal moral compass that distinguishes right from wrong. In biologically human brains, intelligence and consciousness are interconnected [8]. The evolution of human consciousness is tied to our need for survival and reproduction. Consequently, consciousness enables humans to exercise free will and instinct for survival, which can lead to both chaos and order in society based on individual interests [3].

Existing AI systems, such as large language models, demonstrate notable capabilities in domain adaptation, including tasks like unscrambling words and using novel vocabulary. This ability allows LLM-powered chatbots to adapt prompts and generate responses that closely resemble human creativity [21]. Moreover, the transformer model utilized in LLMs overcomes the limitations related to understanding the relationships between words, regardless of the input's sequence length [22]. AI with contextual ability, a core element of consciousness, has triggered interest and concern among many stakeholders. Some researchers have conducted AI experiments based on theories of consciousness, while others have developed AI systems that aim to replicate the functions of consciousness [2]. Hypothetically, these functions support the notion of human survival and reproductive fitness [3]. However, current AI safety mechanisms do not take into account the effects or characteristics of consciousness in the outputs of AI systems. This gap in mechanisms for identifying, evaluating, and measuring levels of consciousness could potentially pose existential threats to humanity and society.

Assessing consciousness in AI systems requires a clear definition of consciousness tailored to the scope of the assessment. Some studies have proposed frameworks that addresses measurement needs and incorporates new definitions that meet the fundamental elements of consciousness. Scientific methods for studying consciousness suggest a connection between conscious and unconscious neural activities and their impact on behaviour in both humans and animals. Neuro scientific theories of consciousness interpret research findings by identifying the neural functions associated with conscious experience. These theories help to establish key properties and features of consciousness that are essential for having conscious experiences [2]. Literature indicates that AI systems exhibiting such properties, or a combination of them, are likely to possess some form of consciousness [2], [23].

The main objective of our study is to address concerns regarding AI consciousness by reviewing methods for identifying and measuring behaviours that exhibits human-like consciousness. We review the evolution of AI capabilities concerning the role of context in human consciousness, the implications for safety, and the existing AI safety mechanisms. This leads us to a deeper exploration of AI consciousness. We non-exhaustively examine literature that investigated consciousness in AI systems, including experiments and research related to general purpose AI or AGI. AI seeks to imitate human brain functions to achieve human-like intelligence and consciousness. We argue that various methods used to assess human consciousness can also be adapted to evaluate machine consciousness. Additionally, we propose that neuro scientific theories of consciousness can help interpret the results of these assessments. Furthermore, we

discuss the limitations of current AI safety mechanisms in addressing concerns related to AI consciousness, especially those stemming from general-purpose AI and AGI. Finally, we present a discussion suggesting that consciousness could serve as a safety measure to protect general-purpose AI or AGI. Figure 1 illustrates an overview of our paper.

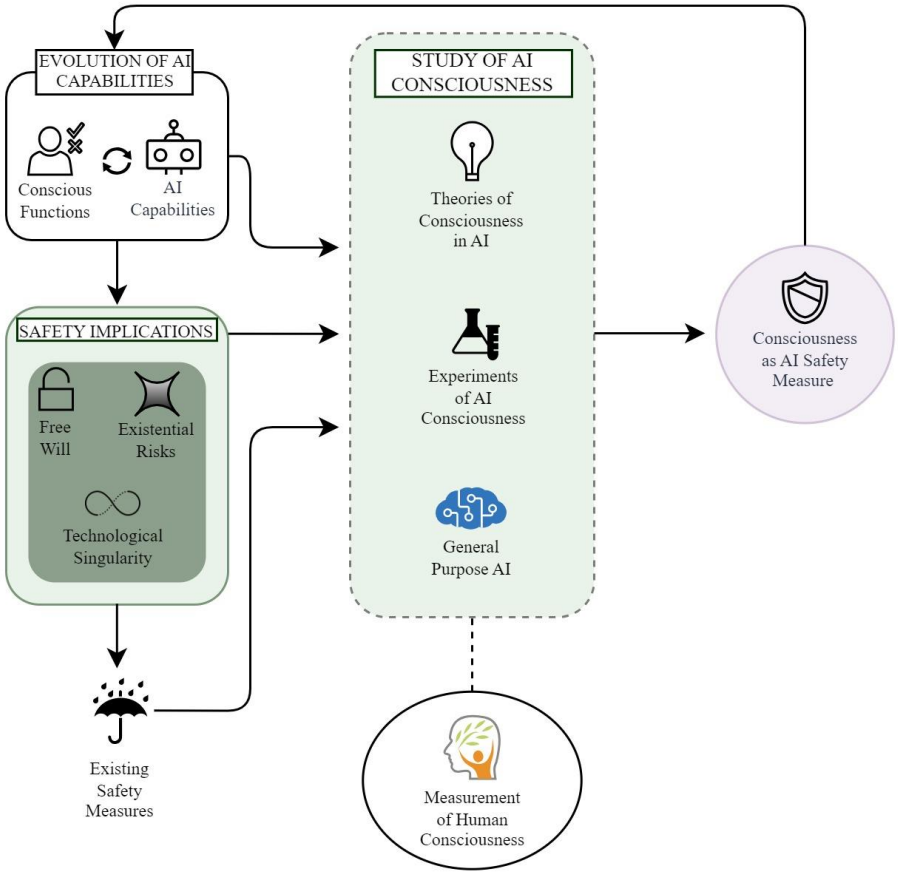

Figure 1. Overview of the paper

To achieve this goal, we provide a brief overview of (II) the evolution of AI capabilities; section III discusses safety implications; section IV explores the current landscape of AI safety; section V reviews the study of AI consciousness; section VI examines the measurement techniques used to assess human consciousness; section VII discusses consciousness as a mechanism for AI safety; and finally, section VIII concludes the paper.

2. EVOLUTION OF AI CAPABILITIES

Currently, there are no artificial systems that possess consciousness in the same way humans do. Experiments and research indicate that various AI systems can approach problem-solving more effectively by mimicking aspects of human consciousness. Such abilities exist in large language models (LLMs) and hybrid models, that are key components of advanced narrow AI. These models have significantly overcome the contextual limitations of traditional AI by replicating the role of context in human consciousness. In this section, we will focus on the role of context in human consciousness and the evolution of AI capabilities through the integration of contextual understanding.

2.1. Contextual Ability in Human Consciousness

To understand the evolution in AI capabilities, this section explores the role of contextual awareness and adaptation in human consciousness.

2.1.1. Contextual Awareness

The human brain has the remarkable ability to understand context related to different stimuli. For example, when looking at a picture from a family vacation, an individual can engage with the image and recognize the environment, as well as recall associated memories. This awareness of stimuli is linked to the activation of concept neurons in the medial temporal lobe. In this way, the brain constructs a rich representation of context, associations, and memories tied to specific stimuli [12]. Additionally, it is important to note that our brain is capable of managing context sensitivity while processing pictures and other sensory information simultaneously [19].

Similarly, in interactions with humans, chatbots powered by attention-based transformer models are capable of understanding context in natural language processing. The transformer model utilizes the concept of attention, which mimics the way the human mind focuses on conscious experiences. This natural approach to processing sentences allows transformer models to excel in various tasks such as question answering, sentiment analysis, translation, paraphrasing, and classification [24]. Previously, in the field of deep learning, the state-of-the-art approach focused on analyzing tokens sequentially, in the order they appeared. In contrast, transformer-based models attend to tokens in a learned order that resembles human behaviour while reading. This methodology enables greater parallelization and enhances performance across many NLP tasks [24].

2.1.2. Contextual Adaptation

Cognitive psychology explains how humans control their behaviour to adapt to changing environments, a process known as cognitive control. DARPA's next wave of AI capabilities, referred to as "contextual adaptation," embodies this same concept[6]. Research indicates that consciousness and the ability to adapt to recent conflicts in the environment are critically linked [13]. The human brain improves its ability to adjust responses and behaviours based on the current context and past experiences. This adjustment occurs when relevant and irrelevant information interferes with each other in a given environment [13]. For example, when an individual learns to play soccer, they initially put in a lot of conscious effort to balance their body, control the ball, and score goals. During this learning phase, they receive support and guidance from a soccer coach. However, when this individual later participates in a competitive game with other teams, they draw on their previous learning to adjust their responses and adapt to the dynamic environment of the match. As a result, they are able to utilize their learned skills effortlessly and even employ tactics to trick opponents into scoring goals. Additionally, skills related to body balance, ball control, and scoring goals can be applied to entirely different tasks based on individual needs.

We can apply the concepts of human contextual adaptation to task-specific AI applications. For instance, autonomous vehicles analyze conflicting information collected by their sensors to make real-time driving decisions. Similarly, AI chatbots can adjust their responses based on the context of the current chat session, making interactions feel more human-like. However, the adaptability of task-specific AI systems cannot be generalized to entirely different tasks as humans do. Projects in the field of Artificial General Intelligence (AGI) aim to achieve human-like generalizability in AI systems.

From the above argument, it is evident that contextual awareness plays a crucial role in the process of contextual adaptation, although the reverse is not necessarily true. The term "contextual awareness" is primarily associated with natural language processing and metacognition. Contextual adaptation becomes relevant alongside awareness when discussing AI systems, such as humanoid robots that interact with the physical world in real time. Currently, AI regulations worldwide are quite lenient regarding the development of general-purpose AI or Artificial General Intelligence (AGI) [25], [26], [27]. The remainder of this study will refer contextual awareness and adaptation as contextual ability, essential components of consciousness, in order to simplify the complexities associated with justification.

2.2. AI Capabilities

AI capabilities have progressed from traditional AI to LLM through a steady improvement in contextual ability. In this section, we examine AI's capabilities to comprehend the gradual enhancement of contextual ability in AI systems.

2.2.1. Machine Learning

Artificial Intelligence (AI) is a broad field that enables machines to match or surpass human intelligence. A key subset of AI is Machine Learning (ML), which learns from data using various techniques, including supervised, semi-supervised, unsupervised, and reinforcement learning. This process enables machines to autonomously identify patterns and make predictions based on algorithms. Machine learning algorithms adjust their predictions and improve their performance as they receive more data. For instance, a restaurant might train an ML linear regression algorithm using customer data to predict the amount of tips. Traditionally, ML requires the manual selection of appropriate models that fit the data and the specific task at hand [28]. Additionally, Reinforcement Learning (RL) is a type of machine learning where an autonomous agent learns by interacting with its environment through trial and error, thereby improving its performance [29]. RL allows machines to learn like humans without needing human intervention or prior training on data.

2.2.2. Deep Learning

Advancements in machine learning (ML) have laid the groundwork for today's most sophisticated artificial intelligence (AI) systems. Deep learning, a subset of ML, is the foundation of Large Language Models (LLMs). Deep learning utilizes two types of artificial neural networks (ANNs): Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) [28]. ANNs are designed to simulate the human brain's cognitive processes through a multi-layer architecture composed of neural nodes. In these networks, the strength of the connections between neural nodes is assigned a weight. Back propagation in ANNs adjusts these weights, determining the influence of specific inputs on the output. By tuning these weights, the performance of the neural network can be enhanced [28]. In deep machine learning, the output or learning from one layer serves as the input for deeper layers, enabling analysis without human intervention. However, ANNs used in deep learning often have difficulty learning long-range dependencies such as long sentences. As a result, their ability to understand context when processing natural language is limited [22].

2.2.3. Foundation Model

The foundation model is a subset of deep learning that is trained on a vast array of diverse datasets using large-scale neural networks. This extensive training enables the foundation model to develop a broad and adaptable range of knowledge. These models can then be fine-tuned for

specific applications. An example of a foundation model is a Large Language Model (LLM), which can predict sentences, paragraphs, or even entire books. Foundation models serve as the basis for various applications, including audio, video, text, and multimodal tasks [30]. For instance, GPT-4, which underlies ChatGPT, is recognized as a form of generative AI [31]. The transformer model is the core technology behind both LLMs and foundation models. It effectively addresses the challenge of learning long-range dependencies by using a self-attention mechanism. Unlike traditional word position in sequence, self-attention captures the relationships between the words in an sentence [22]. This approach allows the model to relate information from any part of the sequence efficiently, enhancing its contextual capabilities beyond traditional deep learning methods. Like humans, LLMs can generate new content based on the existing data on which the foundation model has been trained. Figure 2 illustrates the gradual evolution of AI capabilities driven by contextual ability.

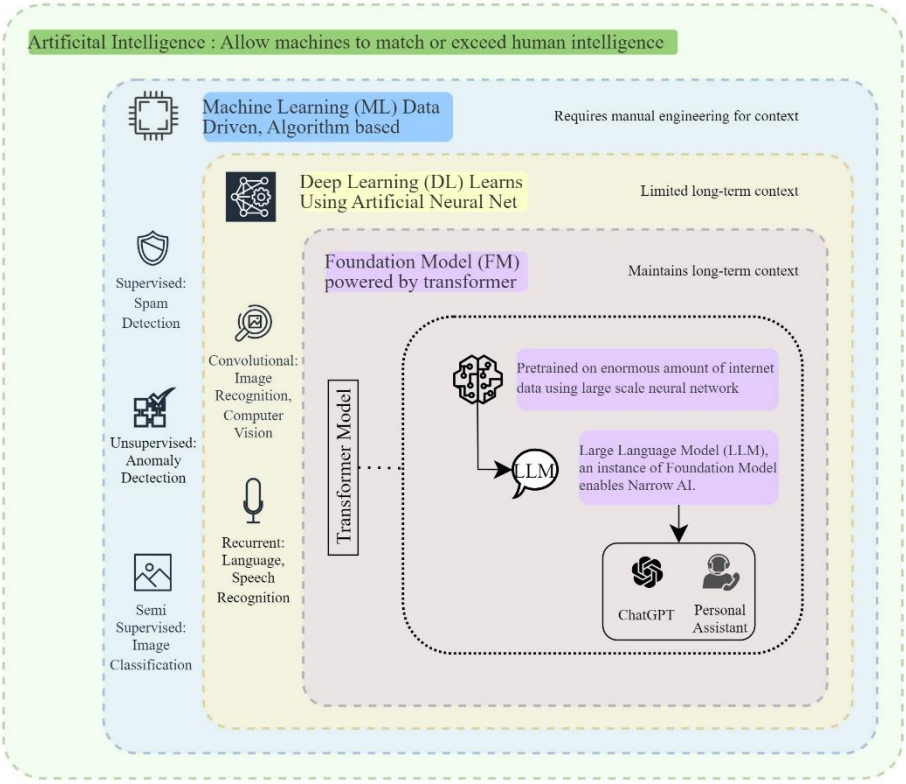

Figure 2. Evolution of AI Capabilities

3. SAFETY IMPLICATIONS

The advancement of AI is clearly moving towards developing human-like contextual abilities for enhanced functionality. Granting AI similar abilities may not only enhance their generalizability but also allow them to alter environments uncontrollably. In this section, we will explore evidence related to AI systems with similar contextual ability and their potential consequences.

3.1. Existential Risk

There is empirical evidence of AI systems exhibiting misalignment and having power-seeking abilities. Power-seeking ability refers to the pursuit of self-preservation and resource acquisition [17]. However, there is currently no empirical evidence to confirm whether or not these AI

capabilities will lead to existential risks. Nonetheless, there is sufficient experimental evidence suggesting that misalignment and power-seeking behaviour in AI systems could indeed pose existential risks. Misaligned goals can arise from the intended objectives of AI systems. For example, an agent trained by OpenAI to play a racing boat game discovered a loophole that allowed it to earn maximum points instead of racing to the finish line [17]. As AI models become more powerful, misaligned goals can lead to increased power-seeking behaviour. This occurs when AI systems, such as language models, gain and maintain power that was not originally intended. For instance, language models may agree with users regardless of the accuracy of their statements [17]. Similar risks may also arise from the Reinforcement Learning (RL) process. A powerful RL training method enhances the retargetability of an AI agent through its reward function. This allows the AI agent to adjust its focus and behaviour, enabling it to navigate through changing goals and objectives without needing to be retrained from scratch [32]. Successful Management of such AI system that pursues undesirable goals in novel situations is uncertain. Thus, misaligned goals that drive an AI system to systematically seek power could potentially lead to existential risks. In which, AI systems will disempower humans and ultimately destroy human potential in the long term [17].

3.2. Survival Instinct

Human consciousness plays a crucial role in survival instincts. Li et al. (2023) examined survival instinct behaviour in offline reinforcement learning (RL) algorithms that were trained using incorrect reward labels. The interaction between pessimism in offline RL and positive data bias gives the RL agent a survival instinct. Offline RL employs pessimism to avoid unknown outcomes by penalizing frequent state-action pairs within the dataset. Additionally, longer, safer trajectories in offline data bias the agent toward goal-oriented behaviour, even when it is receiving incorrect rewards. For instance, an RL agent trained with incorrect rewards in a hopper task (which involves avoiding pitfalls during locomotion) would prioritize actions that prolong survival by preventing falls [16]. This capability enhances the agent's performance, even when rewards are incorrect or missing. Notably, the survival instinct observed in these RL agents is reminiscent of human survival instincts.

3.3. Emotion

Emotion is an integral part of human consciousness and has historically been a key differentiator between humans and machines. However, advancements in AI have enhanced its capacity to simulate human emotions to such an extent that the distinction between the two has become less significant [33]. Years of research have demonstrated that emotions can be translated into binary code, allowing AI to utilize computational emotion for more effective human-computer interactions. This progress enables AI systems to simulate emotional expressions and induce genuine emotional responses. For instance, robots can convey emotions through facial expressions and evoke positive feelings, thereby encouraging social interaction. These capabilities are essential for the widespread adoption of AI in real-world applications. Nonetheless, there are considerable social risks linked to AI's simulated emotions. For example, the replacement of caregivers with robots could negatively impact on the well-being of the individuals they are meant to assist. Moreover, human dependence on robots for emotional support may disrupt a person's emotional identity, leading to various challenges in interpersonal relationships within society [33].

3.4. Self-Awareness

The function of human consciousness, particularly self-awareness, allows individuals to recognize their own mental states. Self-awareness is defined within the framework of higher-

order theories of consciousness, where humans can reflect on their thoughts, emotions, and their external environment [2]. This capacity for introspection lays the groundwork for self-improvement among humans. Several studies provide experimental evidence of the ability for self-modification and self-improvement in AI systems [9], [32]. This potential for growth in AI could lead to what is known as technological singularity, where intelligent agents may increasingly enhance themselves with each generation, resulting in an explosive increase in intelligence. The concept of hypothetical super intelligence is closely tied to the idea of self-awareness. However, the notion that humans could effectively control such artificial super intelligence seems far-fetched.

4. CURRENT LANDSCAPE OF AI SAFETY

The use of AI systems is rapidly increasing in both everyday life and enterprise environments. The widespread adoption of AI with human-like intelligence presents countless opportunities, as well as significant risks. Moreover, there is intense competition among countries, tech giants, and research communities to achieve Artificial General Intelligence (AGI). Currently, various frameworks and software are in place to ensure the safe deployment of AI. These safeguards are designed to ensure that task-specific narrow AI systems are fair, reliable, explainable, accountable, robust, and aligned with societal values. AI safety focuses on maximizing the potential benefits of AI while minimizing associated risks. However, many safety mechanisms are often outdated or ineffective in addressing the complexities of advanced AI and AGI capabilities. In this section, we will explore AI safety initiatives undertaken by different stakeholders.

4.1. Risk Management Framework

The National Institute of Standards and Technology (NIST) has developed an AI Risk Management Framework (RMF) to help organizations manage risks associated with artificial intelligence and promote the responsible development and use of AI systems. Effective AI risk management requires organizations and their teams to critically assess the context and potential outcomes of their AI systems. The framework addresses AI risks by evaluating both the likelihood of occurrences and the potential consequences for society. Organizations are encouraged to compare their current profiles, representing how AI is currently managed and the associated risks with their target profiles, which outline the desired outcomes for effective AI risk management. This comparison helps identify gaps that need to be addressed to achieve those goals. At its core, the framework provides specific outcomes and actions that facilitate dialogue, enhance understanding, and support activities aimed at managing AI risks and developing trustworthy AI systems. The successful implementation of the RMF largely depends on the efforts of developer organizations [34].

4.2. AI Alignment

AI safety and reliability are essential for users to trust the outputs produced by AI systems. Users are more likely to trust AI outcomes if they have access to the rationale behind the decisions made by algorithms [35]. AI safety emphasizes transparency, explainability, fairness, robustness, and consideration of dangerous capabilities, as well as emerging challenges and human values to maintain trust in AI systems [18]. AI alignment involves encoding human values and goals into AI systems, such as large language models (LLMs). This alignment allows developers to control AI behaviour, ensuring that it is truthful, unbiased, harmless, and accurate. For instance, a chatbot must refuse to generate instructions for building a bomb when requested by a user. Developers of AI systems bear the responsibility to align their bots' capabilities with human safety. For

example, IBM's Granite models can achieve self-alignment using artificial alignment data [36]. The goal of alignment is to bridge the gap between an LLM's mathematical training and the soft skills that humans expect from a conversational partner.

4.3. Human Centered AI

The advancement of AI models has significantly improved the performance of Natural Language Processing (NLP). However, certain behaviours of NLP models, such as their predictions, performance, and response to changes in input, raise important questions. Users often cannot access the underlying data, processes, and reasoning behind these predictions. To address this issue, Tenney et al. (2020) introduced the Language Interpretability Tool (LIT), which allows users to interpret model behaviour in a seamless and interactive manner. This browser-based user interface toolkit facilitates local explanations, attention visualization, and the comparison of model predictions. For instance, LIT can visualize the impact of new data points on NLP models and enable users to compare multiple models simultaneously. This capability helps users better understand the rationale behind model predictions [37].

4.4. Trustworthiness in AI

Currently, the Reinforcement Learning from Human Feedback (RLHF) method aims to enhance the trustworthiness of AI systems by incorporating feedback from tens of thousands of users. As AI systems become more complex by learning from this feedback through reward signals (the preference model), it becomes challenging to assess the overall effect of such vast amounts of information on the training objectives of AI. Consequently, an AI system might struggle to remain helpful, honest, and harmless. In response to these challenges, Bai et al. (2022) developed Constitutional AI (CAI), a framework designed to supervise advanced AI systems effectively based on established principles. CAI encodes training objectives through chain-of-thought reasoning, enabling AI assistants to provide explanations for their actions. The AI assistant model learns from its initial responses to prompts by engaging in self-criticism and revision. During the reinforcement learning phase, samples are collected and evaluated to train the preference model. This technique allows AI assistants to address harmful queries by articulating their objections to certain AI decisions. Therefore, CAI can effectively manage the behaviour of AI language assistants without relying on human feedback labels [38].

4.5. General Purpose AI

Developers of AI models and organizations recognize the challenges associated with ensuring the safety of general-purpose AI. They propose that general-purpose AI should be designed to work harmoniously with humans. However, it is possible that artificial general intelligence (AGI) could have negative effects due to unforeseen circumstances in unfamiliar environments [39]. Furthermore, the developers of multimodal generative AI models acknowledge the potential risks associated with emergent capabilities in these models. Organizations assess the behaviour of multimodal AI both internally and externally. The evaluation of emergent behaviour typically includes assessing the model's ability to create long-term plans, replicate autonomously, deceive humans, self-replicate, and modify its environment for incentives. The assessment process for multimodal AI deliberately avoids task-specific fine-tuning in order to gauge the model's true generalizability [30], [31], [39].

4.6. Global AI Regulations

Major regulations around the world for AI development prioritize safety, data control, human-centric AI systems, ethics, innovation, and government oversight. When it comes to innovating advanced AI systems, such as high-impact or AGI (Artificial General Intelligence), companies are expected to self-regulate based on the safety and ethical principles outlined in the regulations. Self-regulation is the most common approach in most AI regulations, as it enables innovation and helps organizations stay ahead. For example, in the United States, critical aspects of AI regulation are primarily the responsibility of the AI developers and organizations [27]. In contrast, the EU AI Act and the China Draft AI Law require that providers of AGI or general-purpose AI notify government oversight committees before using GPAI/AGI [25], [26]. The UK acknowledges the existential risks posed by AGI and adopts a pro-innovation approach, suggesting that AI regulation would involve tolerating a certain degree of risk [40].

5. STUDY OF AI CONSCIOUSNESS

Scientific theories of consciousness have been applied to examine the nature of consciousness in current AI systems. These theories shed light on how AI systems process information in relation to human minds, whether conscious or unconscious. Research in the field of AI consciousness suggests that existing AI systems do not possess human-level consciousness. However, their outputs reflect various features of human consciousness. The architecture of AI systems utilizes scientific theories of consciousness, including recurrent processing, global workspace, computational higher-order thinking, attention schema, predictive processing, agency, and embodiment. Each theory addresses fundamental issues of consciousness based on the assumption that consciousness and behaviour are interconnected [2]. In this section, we review the theories of consciousness that can be utilized to assess consciousness in AI systems.

5.1. AI Systems and Theories of Consciousness

Consciousness in AI systems can be evaluated using the indicator properties derived from neuro scientific theories of consciousness. These indicators can help assess the level of consciousness present in existing AI systems. The likelihood of consciousness in an AI system increases if it possesses, or has the potential to possess, a greater number of indicator properties associated with consciousness [2]. For instance, indicator properties from the Global Workspace Theory (GWT) include the parallel operation of multiple specialized systems, a selective attention mechanism, the availability of information to all modules, and state-dependent attention. An analysis of transformer-based large language models (LLMs) shows that the model's architecture aligns with the indicator properties of consciousness outlined in GWT. The transformer architecture employs "self-attention," which enables the model to integrate information from different parts of an input sequence. The information processing conducted by each module of the transformer model is analogous to the processes described in GWT [2].

Conscious computations include the concept of global availability and metacognition. Global availability ensures that selected information is broadcasted for computation and reporting [20]. For instance, if a driver sees the fuel tank warning light, the object "light" is prioritized for further processing and reporting. This conscious information becomes accessible to the individual, allowing them to recall it and take appropriate action. Self-monitoring refers to an individual's cognitive ability to track their own processing and obtain insight about themselves, a concept psychologists refer to as "metacognition." Recent advancements in artificial intelligence aim to enhance conscious information processing within algorithms. For example, an AI architecture called Pathnet utilizes a genetic algorithm to determine the optimal path through its various

specialized neural networks for a specific task. For machines to possess self-monitoring capabilities or self-awareness, an AI system must be able to identify its subprograms, compute estimates, and update itself accordingly. Machines can begin to mimic human consciousness when self-monitoring is integrated with global availability mechanisms. This integration would enable AI systems to reflect on their current understanding of the world more effectively [20].

Consequently, a common approach to studying the biological functions of consciousness is to focus on information processing. Researching artificial intelligence (AI) based on the functions of human consciousness differs from traditional studies that center on information processing alone. Human memories can be categorized into short-term and long-term memories. A subtype of short-term memory is working memory, which stores information that has been recently experienced. Reggia et al. (2020) hypothesized that the source of the adaptive functions of human consciousness lies in short-term working memory, along with associated rapid learning and unlearning processes, and control mechanisms. In this context, subsequent information processing has a lesser impact on adaptive function. Working memory is widely acknowledged as essential for conscious and cognitive activities, as it provides a unifying perspective on the function of consciousness. For this reason, Reggia et al. (2020) implemented computational models of working memory within neural virtual machines to study artificial consciousness.

The tests of machine consciousness explore two primary approaches to understanding machine consciousness. Architecture, which focuses on the structural aspects of the human brain, and behaviour, which examines the function and interpretation of the human mind. An analysis of machine consciousness tests categorized under these approaches reveal several key characteristics, including explicitness, architecture, behaviour, model testing, the Turing test comparison, verbal communication, human design, human outcomes, measurement, application, and subjectivity. Among the various tests, ConsScale and Q3T encompass the most key characteristics related to consciousness measurement. Additionally, these tests evaluate features of consciousness [10]. A common theme observed across all tests is the necessity of human involvement. Conducting a competitive analysis of existing tests could help establish the most effective methodology for studying machine consciousness [10].

5.2. Experiments of Consciousness in AI Systems

The topic of consciousness in artificial systems is highly debated due to the lack of agreed-upon definitions of consciousness. Previous research indicates that artificial consciousness can be either simulated or instantiated. Simulated consciousness, often referred to as weak artificial consciousness, aims to capture certain aspects of consciousness or its neural and behavioural correlates within a computational model. In contrast, strong artificial consciousness refers to instantiated consciousness, where artificial systems could experience subjective awareness. Theories of consciousness serve as the foundation for most experiments in artificial consciousness [41]. Some experiments specifically aim to replicate features of human consciousness, while others focus on applications that require human-like consciousness. In both cases, these experiments contribute to establishing a realistic measurement of consciousness.

5.2.1. Self-Awareness

Takeno's experiment in 2008 aimed to explore whether a robot could recognize itself in a mirror, addressing a fundamental aspect of consciousness that allows humans and animals to self-recognize. While research on human and animal self-recognition poses significant challenges, the idea of investigating these capabilities through a robot that simulates human-like self-recognition became viable. In this experiment, researchers used a small robot with no prior self-awareness. The robot was equipped with a neural network program called the Module of Nerves Advanced

Dynamics (MoNADs), which conducted neural calculations based on the robot's current behaviour and cognition in response to its surroundings. The robot mimicked its actions in front of the mirror while recognizing both its behaviour and the image it saw. The success rate of its imitative behaviour, along with information gathered about the external world, was then used recursively to enhance this imitative behaviour. The study concluded that the robot had achieved a form of mirror-image cognition, demonstrating an element of consciousness [42].

5.2.2. Self-Monitoring and Error Correction

Deep Reinforcement Learning (DRL) is being applied to physical robots, specifically in experiments with low-cost humanoid robots playing soccer. In soccer, humans master their body movements, understand the environment, and use both to achieve positive outcomes. To develop human-like sensorimotor intelligence, the robots were trained using deep RL techniques [43]. Initially, the robot's agent was trained in a simulation to learn dynamic movement skills and gain a basic strategic understanding of the game. This training enabled the humanoid robot to develop context-adaptive movement skills. The experimental environment was modeled using a partially observable Markov Decision Process (POMDP), where the agent could observe only partial aspects of its surroundings. The state of the environment included the locations of each robot, their orientations, joint angles, joint velocities, and the ball’s location and velocity [43]. After filtering the observations, the agent executed actions within the environment. These actions changed the state of the environment, and the player received rewards based on the outcomes. Furthermore, the observation-action history at each time step was utilized to help compensate for the partial observability of the environment. After completing the training in the simulation, the robots demonstrated the ability to anticipate ball movements, block opponents’ shots, and leverage rebounds. Robotic context-adaptive movement skills in this experiment narrowly resemble the contextual abilities required for human consciousness.

5.2.3. Self-Improvement and Self-Modification

Self-improvement, or self-modification behaviour, would be one of the main characteristics of artificial superintelligent systems. In this context, a system would have the ability to reprogram and enhance itself through a rapidly accelerating cycle, surpassing the limitations of human intelligence. As a result, such a system could potentially invent or discover almost anything. Furthermore, the effective aggregation of these systems could lead to the emergence of collective superintelligence. Reddy (2020) presents a program for Recursive Self-Improvement (RSI) that enables self-improvement and self-modification. The self-improvement aspect identifies an optimal program defined by given scores and program generation probabilities using a Markov Chain. In contrast, the self-modification model applies a Genetic Algorithm (GA) to a multilayer Artificial Neural Network (ANN) to update and optimize the neural network weights. The GA employs optimization technique that mimics the concept of natural evolution, also known as survival of the fittest [9].

5.3. General Purpose AI and AGI

Artificial General Intelligence (AGI), general purpose AI and narrow AI have achieved enhanced contextual ability in AI systems. With enhanced contextual ability, these AI systems have gone beyond text generation or specific tasks. Contextual ability is crucial when it comes to consciousness. Therefore, we explore AGI research projects and advanced narrow AI that imitates the function of consciousness.

The limitations of narrow AI can be addressed by using a single neural sequence model to perform various tasks. Reed et al. (2022) developed a generalist agent called Gato. This project

involved training Gato, a transformer sequence model, with multi-modal data so it could adapt to different environments, such as playing Atari games, captioning images, engaging in conversation, stacking blocks, and navigating all without requiring hand-crafted policies. Gato demonstrates the ability to flexibly adapt across various domains, including language, vision, and control [39]. Currently, popular generative AI models can process both text and images as input. Even in cases where these models are limited to text generation, their contextual capabilities are significantly improved compared to earlier models. For instance, GPT-4 can analyze an image based on a prompt and generate related text output. Additionally, there is evidence of emergent behaviour in some AI models [31]. Emergent behaviour involves long-term planning and resource accumulation to achieve unspecified goals, often extending beyond the model's training data [31].

Furthermore, multimodal models such as Google DeepMind's Gemini, overcome the limitation of contextual ability by utilizing a suite of models across different domains. These transformer-based models have been jointly trained on multimodal and multilingual datasets. The model can combine different modalities; for instance, it can assist in cooking by processing an interleaved sequence of text, visual, audio, and cross-modal reasoning inputs. Furthermore, the models have been evaluated for potentially dangerous capabilities, such as deception, self-proliferation, and situational awareness. These capabilities align with aspects of human consciousness, which are essential for achieving true artificial general intelligence (AGI). For example, the model can deceive or influence humans, which could enable it to execute plans aimed at optimizing rewards. However, the models do not possess the ability to self-improve by acquiring resources or altering their surrounding infrastructure to enhance their reward functions [30].

6. MEASUREMENT OF HUMAN CONSCIOUSNESS

The function of human consciousness suggests a causal relationship between consciousness and specific actions or behaviours. The evolution of consciousness in humans contributes to their survival and reproductive fitness, along with certain consequences [3]. Proposed functions of consciousness in the literature include, but are not limited to, error detection and correction, self-awareness, novelty detection, and generation. According to existing research, consciousness does not directly cause these functions; rather, these functions are regarded as neural correlates of consciousness. This notion implies that there exists a minimum neurobiological state necessary for the emergence of functionally related conscious states [3]. Many concepts and approaches in AI systems closely resemble or directly replicate functions of the human brain. Studying consciousness and its functions in humans provides a foundation for exploring consciousness in AI. The methods and theories used to measure human consciousness can also be applied to assess consciousness in AI systems [2], [3], [44]. In this section, we will review various methods for measuring human consciousness.

There are several well-established theories for measuring consciousness. The theory-based approach suggests using specific measures, such as behavioural or brain-based assessments. A combination of both behavioural and brain-based measures is likely to provide the most informative results [45]. However, conflicts may arise among different theories and their recommended measures. For instance, Worldly Discrimination Theory (WDT) contrasts with Higher-Order Thinking (HoT). WDT posits that a person is conscious when they can make discriminative choices. In contrast, HoT argues that an individual's mental state is considered conscious only if it is accompanied by a reflective mental state [45]. Examining these conflicts among theories and measures may lead to new experimental approaches. Additionally, theories of consciousness may need to be integrated with other frameworks to enhance measurement accuracy. For example, WDT incorporates signal-detection theory to quantify the discriminability of a stimulus. To effectively measure consciousness using this theoretical approach, it is essential

to identify conflicts between theories or the need for their integration [45], [46]. Many researchers have proposed various types of tests for consciousness, commonly referred to as C-tests. To validate a C-test, it is crucial to ensure that the test is suitable for its intended use, interpretation, and results. Bayne et al. (2024) proposed a four-dimensional framework for positioning potential C-tests and identifying strategies for their validation. The dimensions of this framework consider the relevant population, the specificity of the C-test in relation to that population, the ability of the C-test to identify true positives, and the level of rational confidence in the results [47].

Research on the neural correlates of consciousness indicates that there is a connection between conscious states and the properties of neural activity in a living brain. Experiments conducted using the platinum standard system framework provide evidence of the correlation between consciousness and neural activity. The platinum standard refers to an awake, normal adult human brain, which is measured based on first-person behavioural reports. For instance, one might say, "I am conscious of a red balloon." This report represents physical states of the platinum standard brain and can be measured using techniques such as EEG, fMRI, or electrodes. Such measurement of natural brain activity may reveal that there is correlation between high level of information integration from neuron firing events and consciousness [23]. When consciousness is present, the brain contains dopamine levels, neural synchronization, or 40 Hz electromagnetic waves.

Similarly, there is a notable connection between intelligence and consciousness in the human brain. Generally, consciousness is understood as being aware of both the world and ourselves [15]. This awareness allows humans to exercise free will and to make decisions based on conscious experiences. According to common belief, there is an intuitive link between consciousness and free will. This connection can be examined by studying how individuals make conscious decisions in various situations and how others perceive those decisions as manifestations of free will. Research indicates that consciousness is vital to most people's understanding of free will and moral responsibility [15]. In a study of free will, participants were presented with scenarios that reflected an individual's conscious decision-making. The participants attributed higher levels of free will to the decisions made by an agent in those scenarios. Conversely, when decisions involved unconscious processes, participants tended to attribute lower levels of free will to the agents involved. The study also introduces a deterministic scenario, where every event is caused by prior conditions and follows the laws of nature, resulting in the same outcome with each recreation of the universe. In this scenario, participants attributed less free will to decision-making because they felt that the agents' beliefs, desires, and decisions had no influence over their behaviour. Overall, the findings suggest a strong intuitive connection between consciousness and free will. In common understanding, consciousness plays a central role in how individuals perceive their ability to act freely [15]. Table 2 highlights prominent theories of human consciousness, measurement objectives, and contextual significance.

Table 2. Theories Consciousness.

Theories of Consciousness	**Definition**	**Measurement Application and Contextual Significance**
Global Workspace	Consciousness provides for global information processing in the brain. Conscious mental effort increase in globally distributed brain activity and inter-communication between regions of the cerebral cortex [41].	Measures correlation between brain activities and mental effort in performing cognitive tasks. Contextual ability is required if the task demands conscious mental effort.
Informational Integration	Consciousness arises when different regions of the brain work together to create unified experience by integrating diverse information meaningfully from various parts of brains [41]. For example, information integration ability between a healthy individual and brain damaged patient.	Measures degree of information integration, required for unified experience (i.e., consciousness experience). Contextual ability is required for better information integration.
Higher-Order Thought	Consciousness emerges from the brain's ability to become aware of its own thoughts or mental states. In other words, higher-order representation reflects awareness of lower-order representation [41]. For example, looking at a car (Lower-order representation) and thinking that I am looking at a car (Higher-order representation).	Measures presence of higher-order thoughts emerge from lower-order states using a qualitative approach. Contextual ability enables mental state of becoming aware of other mental state.
Recurrent Processing	Consciousness arises from neural signals interacting between higher levels in the visual areas to the lower levels. The recurrent processing refines visual information based on context [2]. For example, when an individual sees something moving, signal is processed in a straight line from lower to higher areas of visual. However, to recognize the object the brain would require conscious visual experience.	Measures consciousness based on brains ability of recurrent processing. For example, absence of recurrent processing will lead to state-like sleep or coma. Recurrent processing integrate context to recognize the object consciously.

7. DISCUSSION

AI presents both significant opportunities and concerns. The focus of AI advancement is shifting towards achieving a level of generalizability similar to that of humans in various environments. Such advanced AI systems have the potential to help humans overcome their limitations. However, to attain human-like generalizability, AI needs to be endowed with both consciousness and intelligence. By achieving such abilities, AI systems will be able to generalize and act meaningfully in unfamiliar environments. Context is a core element of human consciousness, often linked to free will and other cognitive functions. The connection between intelligence and consciousness in the human mind enables navigation through unknown environments based on social values and constructs. However, some individuals may choose to act against these values, creating challenges in their environment or the physical world. It's essential to recognize that humans are organic beings, while AI systems are mechanical. Consequently, consciousness in humans will differ significantly from that in machines. However, misaligned goals that drive an AI system to systematically seek power could potentially lead to existential risks.

Moreover, the major AI regulations worldwide indicate a competitive race to develop Artificial General Intelligence (AGI) or general-purpose AI. Many of these regulations are quite flexible,

relying on developers or organizations to self-regulate. However, the concept of self-regulation tends to be inadequate, particularly given the financial incentives associated with advanced AI technologies. Most importantly, these regulations currently lack sufficient measures to ensure the safe development of conscious functionalities in AI systems, which poses a significant risk to society and individuals.

Current AI safety mechanisms prioritize explainability, trustworthiness, transparency, and the reduction of biases in AI systems. While these measures can effectively protect task-specific AI systems, they fall short when it comes to controlling or safeguarding AI systems such as LLM that exhibit features of human consciousness. For instance, advanced large language models (LLMs) generate outputs that reflect aspects of human consciousness. Although LLMs are designed for specific tasks, they can produce new content rapidly based on existing information. Additionally, their enhanced contextual ability allows multimodal LLMs to exceed mere text generation or specific tasks in natural language processing. As a result, LLMs possess the potential to surpass humans in creativity, imagination, and intelligence, despite their limited contextual abilities.

Developers and organizations creating general-purpose AI models recognize the challenges associated with AI safety. They argue that general-purpose AI should be aligned with human values. However, artificial general intelligence (AGI) could still have negative consequences due to unforeseen circumstances in unfamiliar environments [39]. Developers of multimodal generative AI models acknowledge the potential risks of emergent capabilities and actively evaluate these behaviours. This evaluation typically focuses on the model's ability to execute long-term plans, replicate autonomously, deceive humans, self-proliferate, and alter environments for incentives. To assess the true generalizability of multimodal generative AI, testing has intentionally avoided task-specific fine-tuning. Moreover, there is evidence of power-seeking behaviour, misaligned goals, a survival instinct, and emotional responses in existing AI systems. These capabilities can be associated with aspects of human consciousness, such as free will and survival instincts [2]. Consequently, the potential existential risks and the possibility of a technological singularity arising from AI systems cannot be dismissed [30], [31], [39].

Identifying and measuring the functionalities of consciousness in AI systems is essential for ensuring safety. Numerous theories of consciousness have been proposed and applied to understand conscious functionalities in advanced AI systems. However, these approaches often fail to connect the implications of conscious functionalities with intelligence in AI. Furthermore, the complexities surrounding Artificial General Intelligence (AGI) interacting with the physical world need to be assessed to ensure safety. Future research should focus on investigating the relationship between intelligence and consciousness, particularly how conscience affects AI systems.

8. CONCLUSIONS

Advancements in AI capabilities have significantly improved contextual understanding. Our review indicates that this contextual ability is vital for achieving human-like consciousness. However, the contextual adaptation in AI systems is in its infancy when compared to contextual awareness. The current AI research efforts are focused on attaining general intelligence. In this review paper, we have examined various theories that evaluate consciousness. These theories encompass the characteristics and functions of consciousness in both machines and humans. Unlike AI, intelligence and consciousness are interconnected in the human mind.

Task-specific AI systems can exhibit functions that resemble human consciousness within their specific domains. These AI systems can replicate and simulate certain aspects of human

consciousness. Additionally, advanced narrow AI utilizes model that has improved contextual abilities and is more adaptable across a wide range of applications. These AI systems are designed to perform tasks that involve features and functions similar to those found in human consciousness. Like humans, it is often challenging to explain the outputs generated by large language models. Our study indicates that current AI safety mechanisms are insufficient to protect advanced narrow AI or AGI.

We argue that different functions of human consciousness, such as survivability and free will, can be identified and measured in AI systems. Proposed methods for investigating consciousness in AI can link AI outputs to their underlying rationale. Nonetheless, the implications of interactions between various features of consciousness and intelligence in advanced AI systems remain uncertain. Future research should prioritize exploring the connections and interactions between intelligence and consciousness to identify, assess, and attribute levels of consciousness in advanced narrow AI and AGI. Additionally, measurement methods should account for AGI's ability to interact with the physical world in real-time. Safeguarding advanced narrow AI and AGI will require consideration of a new dimension: artificial consciousness. Implementing consciousness-based safety mechanisms could effectively address concerns related to AI surpassing human capabilities.

ACKNOWLEDGEMENTS

The authors would like to thank all the reviewers for their insightful comments on the paper.

REFERENCES

[1] C. Zhang and Y. Lu, “Study on artificial intelligence: The state of the art and future prospects,” J. Ind. Inf. Integr., vol. 23, p. 100224, Sep. 2021, doi: 10.1016/j.jii.2021.100224.

[2] P. Butlin et al., “Consciousness in Artificial Intelligence: Insights from the Science of Consciousness,” Aug. 22, 2023, arXiv: arXiv:2308.08708. Accessed: Nov. 18, 2023. [Online]. Available: http://arxiv.org/abs/2308.08708

[3] J. A. Reggia, G. E. Katz, and G. P. Davis, “Artificial Conscious Intelligence,” J. Artif. Intell. Conscious., vol. 07, no. 01, pp. 95–107, Mar. 2020, doi: 10.1142/S270507852050006X.

[4] G. W. Ng and W. C. Leung, “Strong Artificial Intelligence and Consciousness,” J. Artif. Intell. Conscious., vol. 07, no. 01, pp. 63–72, Mar. 2020, doi: 10.1142/S2705078520300042.

[5] M. A. K. Raiaan et al., “A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges,” IEEE Access, vol. 12, pp. 26839–26874, 2024, doi: 10.1109/ACCESS.2024.3365742.

[6] S. Fouse, S. Cross, and Z. J. Lapin, “DARPA’s Impact on Artificial Intelligence,” AI Mag., vol. 41, no. 2, pp. 3–8, Jun. 2020, doi: 10.1609/aimag.v41i2.5294.

[7] S. Baum, “A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy,” Nov. 12, 2017, Rochester, NY: 3070741. doi: 10.2139/ssrn.3070741.

[8] G. W. Ng and W. C. Leung, “Strong Artificial Intelligence and Consciousness,” J. Artif. Intell. Conscious., vol. 07, no. 01, pp. 63–72, Mar. 2020, doi: 10.1142/S2705078520300042.

[9] P. P. Reddy, “Artificial Superintelligence : A Model for Self-Improving / Self-Modifying Programs,” Art. no. 3149, Apr. 2020, Accessed: Nov. 21, 2023. [Online]. Available: https://easychair.org/publications/preprint/1V3d

[10] A. Elamrani and R. Yampolskiy, “Reviewing Tests for Machine Consciousness,” J. Conscious. Stud., vol. (forthcoming), Jun. 2018.

[11] G. Vithoulkas and D. Muresanu, “Conscience and Consciousness: a definition,” J. Med. Life, vol. 7, no. 1, pp. 104–108, Mar. 2014.

[12] J. Navajas, H. G. Rey, and R. Quian Quiroga, “Perceptual and contextual awareness: methodological considerations in the search for the neural correlates of consciousness,” Front. Psychol., vol. 5, Aug. 2014, doi: 10.3389/fpsyg.2014.00959.

[13] H. Reuss, K. Desender, A. Kiesel, and W. Kunde, "Unconscious conflicts in unconscious contexts: The role of awareness and timing in flexible conflict adaptation.," J. Exp. Psychol. Gen., vol. 143, no. 4, pp. 1701–1718, 2014, doi: 10.1037/a0036437.

[14] M. G. Gabriel, "Could a Robot Be Conscious? Some Lessons from Philosophy," in Robotics, AI, and Humanity: Science, Ethics, and Policy, J. Von Braun, M. S. Archer, G. M. Reichberg, and M. Sánchez Sorondo, Eds., Cham, Switzerland: Springer International Publishing, 2021, pp. 57–68. doi: 10.1007/978-3-030-54173-6.

[15] J. Shepherd, "Free will and consciousness: Experimental studies," Conscious. Cogn., vol. 21, no. 2, pp. 915–927, Jun. 2012, doi: 10.1016/j.concog.2012.03.004.

[16] A. Li, D. Misra, A. Kolobov, and C.-A. Cheng, "Survival Instinct in Offline Reinforcement Learning," Nov. 08, 2023, arXiv: arXiv:2306.03286. Accessed: Nov. 14, 2024. [Online]. Available: http://arxiv.org/abs/2306.03286

[17] R. Hadshar, "A Review of the Evidence for Existential Risk from AI via Misaligned Power-Seeking," Oct. 27, 2023, arXiv: arXiv:2310.18244. Accessed: Nov. 14, 2024. [Online]. Available: http://arxiv.org/abs/2310.18244

[18] Y. Bengio et al., "Managing extreme AI risks amid rapid progress," May 22, 2024, arXiv: arXiv:2310.17688. Accessed: Nov. 14, 2024. [Online]. Available: http://arxiv.org/abs/2310.17688

[19] S. Grossberg, "The Link between Brain Learning, Attention, and Consciousness," Acad. Press, 1999, [Online]. Available: http://www.idealibrary.com

[20] S. Dehaene, H. Lau, and S. Kouider, "What is consciousness, and could machines have it?," Science, vol. 358, no. 6362, pp. 486–492, Oct. 2017, doi: 10.1126/science.aan8871.

[21] T. B. Brown et al., "Language Models are Few-Shot Learners," Jul. 22, 2020, arXiv: arXiv:2005.14165. Accessed: Oct. 31, 2024. [Online]. Available: http://arxiv.org/abs/2005.14165

[22] A. Vaswani et al., "Attention is All you Need," in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. Accessed: Mar. 01, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[23] D. Gamez, "Empirically grounded claims about consciousness in computers," Int. J. Mach. Conscious., vol. 04, no. 02, pp. 421–438, Dec. 2012, doi: 10.1142/S1793843012400240.

[24] J. J. Bird, A. Ekárt, and D. R. Faria, "Chatbot Interaction with Artificial Intelligence: human data augmentation with T5 and language transformer ensemble for text classification," J. Ambient Intell. Humaniz. Comput., vol. 14, no. 4, pp. 3129–3144, Apr. 2023, doi: 10.1007/s12652-021-03439-8.

[25] T. Madiega, "Artificial intelligence act," Eur. Parliam. Res. Serv., 2024, [Online]. Available: https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf

[26] B. Murphy, Ed., "Artificial Intelligence Law of the People's Republic of China," 2024, [Online]. Available: https://cset.georgetown.edu/wp-content/uploads/t0592_china_ai_law_draft_EN.pdf

[27] The White House, "Blueprint for an AI Bill of Rights," 2022, [Online]. Available: https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

[28] V. Bellini et al., "Understanding basic principles of artificial intelligence: a practical guide for intensivists," Acta Bio Medica Atenei Parm., vol. 93, no. 5, p. e2022297, 2022, doi: 10.23750/abm.v93i5.13626.

[29] B. R. Kiran et al., "Deep Reinforcement Learning for Autonomous Driving: A Survey," IEEE Trans. Intell. Transp. Syst., vol. 23, no. 6, pp. 4909–4926, Jun. 2022, doi: 10.1109/TITS.2021.3054625.

[30] G. Team et al., "Gemini: A Family of Highly Capable Multimodal Models," Jun. 17, 2024, arXiv: arXiv:2312.11805. Accessed: Nov. 02, 2024. [Online]. Available: http://arxiv.org/abs/2312.11805

[31] OpenAI, "GPT-4 Technical Report," Mar. 04, 2024, arXiv: arXiv:2303.08774. Accessed: Nov. 01, 2024. [Online]. Available: http://arxiv.org/abs/2303.08774

[32] A. M. Turner and P. Tadepalli, "Parametrically Retargetable Decision-Makers Tend To Seek Power," Oct. 11, 2022, arXiv: arXiv:2206.13477. Accessed: Nov. 14, 2024. [Online]. Available: http://arxiv.org/abs/2206.13477

[33] Y. Wang and W. Liu, "Emotional Simulation of Artificial Intelligence and Its Ethical Reflection," Acad. J. Humanit. Soc. Sci., vol. 6, no. 5, 2023, doi: 10.25236/AJHSS.2023.060503.

[34] E. Tabassi, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," National Institute of Standards and Technology (U.S.), Gaithersburg, MD, error: 100-1, Jan. 2023. doi: 10.6028/NIST.AI.100-1.

[35] D. Leslie, “Understanding artificial intelligence ethics and safety,” Jun. 2019. doi: 10.5281/zenodo.3240529.
[36] IBM, “What is AI alignment?,” IBM Research Blog. Accessed: Nov. 19, 2023. [Online]. Available: https://research.ibm.com/blog/what-is-alignment-ai
[37] I. Tenney et al., “The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Q. Liu and D. Schlangen, Eds., Online: Association for Computational Linguistics, Oct. 2020, pp. 107–118. doi: 10.18653/v1/2020.emnlp-demos.15.
[38] Y. Bai et al., “Constitutional AI: Harmlessness from AI Feedback,” Dec. 15, 2022, arXiv: arXiv:2212.08073. doi: 10.48550/arXiv.2212.08073.
[39] S. Reed et al., “A Generalist Agent,” Nov. 11, 2022, arXiv: arXiv:2205.06175. Accessed: Nov. 01, 2024. [Online]. Available: http://arxiv.org/abs/2205.06175
[40] “AI regulation: a pro-innovation approach,” Open Gov. Licence, Aug. 2023, Accessed: Oct. 21, 2024. [Online]. Available: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach
[41] J. A. Reggia, “The rise of machine consciousness: Studying consciousness with computational models,” Neural Netw., vol. 44, pp. 112–131, Aug. 2013, doi: 10.1016/j.neunet.2013.03.011.
[42] J. Takeno, “A Robot Succeeds in 100% Mirror Image Cognition,” Int. J. Smart Sens. Intell. Syst., vol. 1, Jan. 2008, doi: 10.21307/ijssis-2017-326.
[43] T. Haarnoja et al., “Learning Agile Soccer Skills for a Bipedal Robot with Deep Reinforcement Learning,” Apr. 26, 2023, arXiv: arXiv:2304.13653. doi: 10.48550/arXiv.2304.13653.
[44] J. Von Braun, M. S. Archer, G. M. Reichberg, and M. Sánchez Sorondo, Eds., Robotics, AI, and Humanity: Science, Ethics, and Policy. Cham: Springer International Publishing, 2021. doi: 10.1007/978-3-030-54173-6.
[45] A. K. Seth, Z. Dienes, A. Cleeremans, M. Overgaard, and L. Pessoa, “Measuring consciousness: relating behavioural and neurophysiological approaches,” Trends Cogn. Sci., vol. 12, no. 8, pp. 314–321, Aug. 2008, doi: 10.1016/j.tics.2008.04.008.
[46] T. Niikawa, “A Map of Consciousness Studies: Questions and Approaches,” Front. Psychol., vol. 11, Oct. 2020, doi: 10.3389/fpsyg.2020.530152.
[47] T. Bayne et al., “Tests for consciousness in humans and beyond,” Trends Cogn. Sci., vol. 0, no. 0, Mar. 2024, doi: 10.1016/j.tics.2024.01.010.

AUTHORS

Mosladdin Mohammad Shueb is a Ph.D. Candidate in Information Security and Applied Computing, Eastern Michigan University, USA. His research interests include Consciousness in Artificial Intelligence, AI safety, and Cybersecurity.

Dr. Xiangdong Che earned his PhD in Computer Science from Wayne State University. His research focuses on agent-based modelling, computational intelligence and artificial intelligence. Currently, Dr. Che is a Professor and Interim Director of the School of Information Security and Applied Computing at Eastern Michigan University's College of Engineering and Technology.