https://doi.org/10.26599/AIR.2024.9150018 CAAI Artificial Intelligence Research

Interactive Embodied Intelligence of Machines

Deyi Li[1] ✉

ABSTRACT

This paper discusses how intelligent machines have replaced humans in tasks requiring, heavy, and repetitive labor, whilst being better suited to the requirements of these jobs. The increased capacity for brute force computation has facilitated increased collaborative innovation between man and machines. For example, the intelligent farming machines have overcome the confines of computational power, algorithms, and data, and the next generation of intelligent farming machines is expected to interact, learn, and grow autonomously. In the future, in addition to self enhancement, humans are expected to teach machines to learn and work. Scientists and engineers will collaborate with machines to accomplish invention, discovery, and creation. For "embodied intelligence" in the farming machine context, we propose (1) deep learning should be performed iteratively via real-time interactions with the external world; (2) embodied control and self-regulation can ensure coordination between behaviors of machines and their environment; (3) intelligent farming machines are characterized by the ability to interact, learn, and grow autonomously.

KEYWORDS

embodied intelligence; human-machine interaction; machine intelligence

Embodied intelligence is a term first proposed in 1948 by Alan Mathison Turing, the father of artificial intelligence. As suggested by the word "embodied", it is a form of intelligence in which the body and intelligence are inseparable. In contrast, disembodied intelligence, is a form of intelligence in which the body and intelligence are separated. At the Dartmouth Conference in 1956, Claude Elwood Shannon and his colleagues defined disembodied intelligence as "artificial intelligence" (AI). We add the word "interactive" to the title of this article to emphasize interactions in machines.

The potential of ChatGPT[1] released by OpenAI to replace Google Search is under an intense debate in academia. ChatGPT can interact intuitively with humans via conversations. According to Albert Mehrabian's 7-38-55 rule of personal communication[2,3], 55% of information exchange during human interactions is achieved through nonverbal cues such as facial expression, posture, and body language; another 38% is realized via audio cues, such the tone, emotion, intonation, and speed of the speech; the remaining 7% is conveyed through words. ChatGPT is solely reliant on words.

As can be seen, interactions play a crucial role in the embodied cognition. Embodied intelligence could be both the origin and the destination of human cognition, which may have originated from action and mimicry. Initially, body language results in abstract thoughts, and action is the external manifestation of intelligence. This is called embodied intelligence. Body movements are silent language. Similar to dancers using actions for artistic expression, machines use actions as a manifestation of interaction. For example, the anti-lock braking systems stabilize cars and increase movement precision, while intelligent and dexterous hand movements gently serve tea or rice to the elderly. If driverless wheeled robots used for urban traffic cannot recognize the sign language of the traffic police and ride-hailing gestures of the pedestrian, they should not be allowed to drive. Autonomous cars should have an excellent sense of positional, directional, and geographical awareness, as well as qualified driving performance. In this case, it is necessary to have embodied intelligence comparable to that of a human-driven car. This requirement also applies to those intelligent farming machines.

1 Deep Learning Should Be Performed Iteratively via Real-Time Interactions with the External World

Learning results in the creation of memories, and memory-based intelligence generally supersedes computational intelligence. Advancements in deep learning have facilitated the development of highly capable AI that extends beyond the conventional "programming-limited" paradigm in which pre-written programs are interpreted to create intelligence. In deep learning, labels, rather than memories, are used to directly extract classification-related knowledge from large datasets. The parameters of the algorithm are then amended with data, indicating the arrival of a new machine-learning era.

However, deep learning has certain limitations due to its intrinsic inadequate explainability. These limitations are as follows:

- Training samples are generally provided in the nonsequential third- or fourth-party perspective instead of the active collection with a consistent perspective from the machine's "self".
- Multichannel cross-modal perception is seldom used in deep training, particularly in tasks pertaining to vision, language, and body action.
- Although labeling is crucial for deep learning, it is expensive, as stated in the adage "the amount of intelligence you get depends on how much labor you put in".
- Deep learning has no selective attention mechanism, preventing effective interpretation of new observations from current job and long-term memories.

1 Deparment of Computer Science and Technology, Tsinghua University, Beijing 100084, China.
Address correspondence to Deyi Li, leedeyi@tsinghua.edu.cn

© The author(s) 2024. The articles published in this open access journal are distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

- The universality and robustness of the system are poor, with biases in data. Algorithms become vulnerable to attack with the use of adversarial samples.
- Training large neural networks using large models and numerous parameters is computationally taxing.
- Deep learning models is unable to accomplish online real-time learning. For example, if a machine is commanded to recognize new image objects, the model should be first amended and then retrained with new data. This hinders the autonomous growth of intelligent machines.

Yann LeCun, a Turing award winner, proposed a scenario for future deep learning[4], which consists of six modules, i.e., the configurator, perception, world model, cost, short-term memory, and actor. The configurator provides executive control for the configuration of other modules. The perception module receives sensory signals from the physical world to estimate the current status of the system. The cost module evaluates the actions of the machine based on energy minimization. Short-term memory remembers the world model and can enhance or slightly modify the world model. The actor module computes action commands based on the current system state and executes these commands. This model provides satisfactory results; however, there is a lack of human control and interaction with the machine. Therefore, deep learning should be performed iteratively via interactions with the external world.

2 Embodied Control and Self-Regulation Can Ensure Coordination between Behavior of Machines and Their Environment

A farming machine needs to be coordinated with the soils, footpaths, and plants of the field prior to the use. Thus, it is critical to ensure the coordination between the behaviors of these machines and their environment via embodied control and self-regulation.

"Turing computability" laid the foundations for brute force computation. In 200 BC, Archimedes improved the precision of the numerical estimation of Pi (π) from 3.1 (which stood for 1 700 years) to 3.14; in 500 AD, Chongzhi Zu computed π to 3.141592. It takes 2400 years for the π value to move from 2 decimal places to 6 decimal places. But a Turing machine used only 70 years to improve the precision of π to 1012 decimal places. This exponential improvement in computational power exemplifies the precision of the Turing machine and brute force computation. However, even Turing machines have limitations.

Multichannel cross-modal interaction is an integral part of embodied machine intelligence. As behavioral interactions are the manifestations of the exploration and feedback processes of machine cognition, a cognitive machine should be able to learn and grow through its interactions with the environment. The von Neumann computer architecture only has inputs and outputs, which generally lags behind the cognitive machine. The lack of multichannel cross-modal perception and interaction has become a major flaw for computerized intelligent machines, thus necessitating the development of cognitive machines.

Cognition is an "upwards spiral" of perception, cognition, and action; meanwhile, cognition is inseparable from perception and action. To create an intelligent machine, we need to overcome the the limitations of the Turing machine. These limitations include (1) the total focus on cognition, without considering interactions between the machine and the environment, and (2) the focus on computation without considering memories.

Learning is an interactive process characterized by guided learning or self-learning. Natural evolution has equipped mankind with excellent short-term, working, and long-term memories, allowing us to conceptualize time. Time is the cornerstone of human cognition as our memories enable the maintenance of cognitive continuity and the accumulation of knowledge, thus human civilization and history came into being. Humans rely on memory to form boundaries that constrain and shape their thoughts, and memories always supersede calculation. As intelligence normally exists in many forms and modalities, it is inappropriate to confine the definition of intelligence to "the ability to compute".

Turing is regarded as "the father of artificial intelligence" with his 18 years of dedication to studying artificial intelligence. In a life span of 42 years, Turing published a paper on Turing machines at 24 , followed by an 18-year research on AI. His seminal paper in 1950, Computing Machinery and Intelligence, openly broached the subject of whether machines were capable of thought[5]. He analyzed and refuted nine common objections against thinking machines and proposed the concept of teaching machines for improved learning. Furthermore, he concluded that if a machine cannot be differentiated from a human from their verbal behavior (i.e., in conversation), then the machine has the ability to think and is intelligent. This was subsequently used as the "Turing test". In his opinion, a "child program" can then be instructed until it reaches adult intelligence.

However, Turing's proposal has yet to undergo a comprehensive evaluation globally. For example, limited research has been conducted on the "child program', i.e., imparting the "cognitive core of a child" into a program and subsequently teaching the machine, guiding its learning, and making it self-sufficient. Analysis of Turing's refutations against the nine common hindrances to thinking machines reveals that these refutations essentially critique the current fear of machines.

Norbert Wiener, the father of control theory, published the book Cybernetics: Or Control and Communication in the Animal and the Machine[6] in 1948. He stated that if we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere once we have started it, because the action is so fast and irrevocable that we have not the data to intervene before the action is complete, then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it." John McCarthy postulated that "AI should have been called control theory, as it is the automation of intelligence." Wiener claimed that "control is the pursuit of negative entropy, and negative feedback can be used to ensure the stability of a machine's embodied behavioral intelligence." In this vein, self-control is the origin of reinforcement learning, and any reward or punishment function can be equated to a negative feedback control system with a deviation of zero.

For over a decade, our team has been actively researching a machine-driving brain. Although different from that of Yann LeCun in brain architecture (presented in Fig.1), it is equally effective.

In short-term memory, positioning sensors, particularly devices using BeiDou and GPS services, are required to provide centimeter-level navigation precision. Position sensors track the acceleration and the speed of a car; visual sensors capture images; and radar sensors measure distance and direction. The data from these sensors can be fused to create a "driving situation map", which is then sent to the job memory. Similar to the human brain that has long-term memories to store and retrive driving maps and traffic rules, it is necessary for the driving brain to have

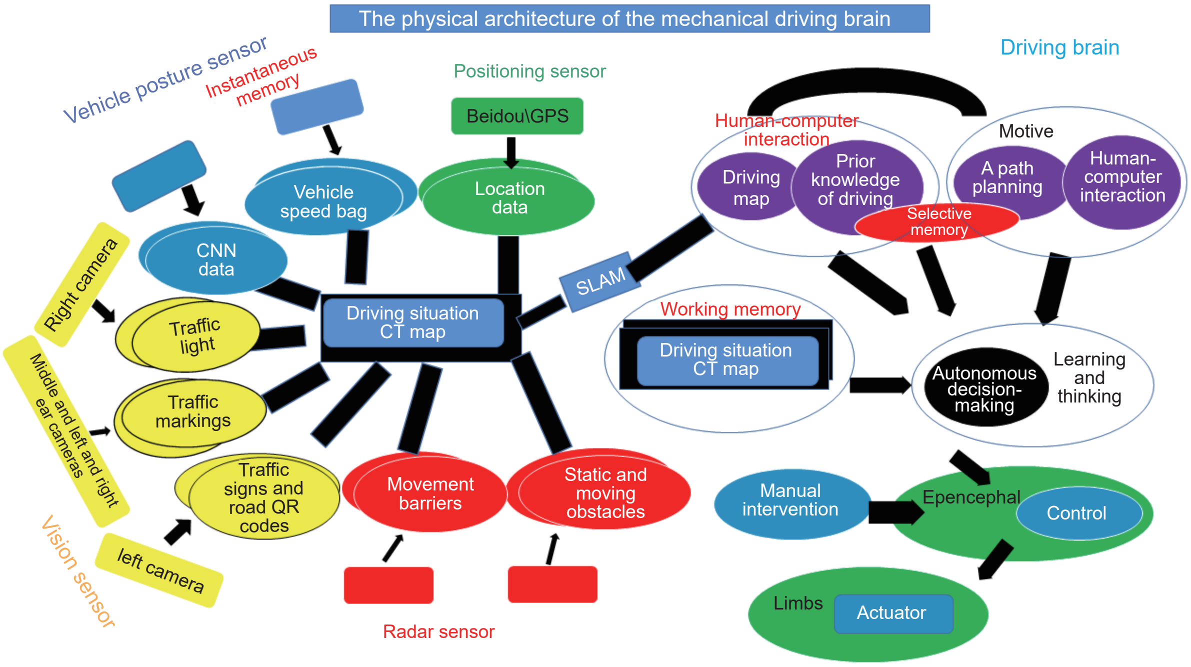

Fig. 1 Driving brain.

memory sticks for such functions. The human–machine interaction is also necessary for performing route planning. The driving brain should be able to learn to make independent decisions, and the movement of the car can be controlled through the control platform and three controller area network buses. In our opinion, in the future, deep learning can be perfected iteratively to learn via real-time interactions in addition to pretraining and preprogramming.

Highly paid "prompt engineers" were hired to ensure effective training of ChatGPT. Similarly, "guide engineers" should train self-driving farming machines; thus, experts in farming techniques can be asked to teach farming machines to work autonomously. In Fig. 2, the physical space is displayed in blue, the cognitive space in light blue. All learning and deductive processes are performed in physical and cognitive spaces.

As shown in Fig. 2, in the cognitive space, short-term memories are formed from situational awareness and cross-modal sensor fusion. In the job memory, a "decision-making blackboard" for current driving situation can be used to compute driving decisions (such as waiting at a junction, overtaking a car, or switching lanes). Furthermore, memories are extracted from long-term memory to change the current driving status based on selective attention and right of way. Meanwhile, the physical space is used for controlling the car. Here, the data from the pose and motion sensors are used as feedback for the operational behaviors of the care to keep it moving in accordance with the driving decisions made by the driving brain. Environmental data from the surroundings of the car can be simultaneously collected, and the inputs of the cognitive space are dynamically adjusted following changes in the environment.

The physical architecture is thus a "perception–cognition–action" architecture with an embedded control circuit used for preprogramming via human–machine communication, thereby fulfilling the responsibility of a "guide engineer." The processes of human teaching, self-learning, and machine understanding the objectives of human-defined tasks can be referred to as "mission alignment," which allows the machine to precisely complete its tasks and manifest its embodied intelligence.

3 Intelligent Farming Machines Determine the Ability to Interact, Learn, and Grow Autonomously

Intelligent farming machines are not constrained by

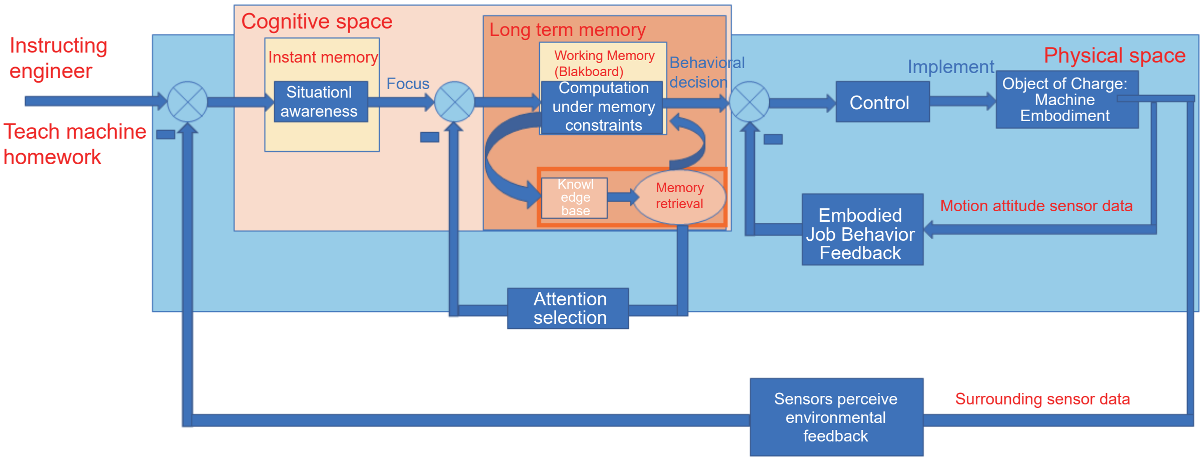

Fig. 2 Perception-cognition-behavior model architecture.

computational power, algorithms, and data and should be capable of interaction, learning, and self-growth. It can be stated that the abilities to interact, learn, and grow autonomously are the defining characteristics of next-generation intelligent machines. First, let us consider the past: why those tools from the agricultural age and machines from the industrial era are unable to think?

The tools of the agricultural era comprised two elements: a solid matter and a virtual structure, with the structure directly parasitizing the matter to form a hard structure. Why does the structure parasitizes matter? Let us consider wheels from the agricultural era. Sap from natural tree trunks was heat bent, hardened into a circular shape, and subsequently affixed to a vehicle. The raw material was then transformed into a tool called wheels. In human history, the significance of the wheel is comparable to the invention of fire.

Machines from the industrial era comprised three elements, namely matter, energy, and structure, where the structure directly parasitizes matter and energy to form a hard structure. For example, in a clock, the swinging pendulum is a structure that directly parasitizes matter and energy to move in a precise manner. Similarly, the steam engine and generator are examples in which the structure directly parasitizes matter and energy. However, time was not considered in industrial-era machines; even clocks just displayed time as a readable number. Therefore, the definition of time by Einstein was critical: as time is just a readable number on a clock, it makes no contribution to the precise movements of the clock itself.

Machines from the AI age have four elements: matter, energy, structure, and time. Due to major changes during the AI age, the "view of life" machines in this era can be explained from perspectives of cognitive and behavioral layers. Matter and energy are the physical representations of the physical layer. Structure and time are the abstract concepts in the cognitive layer. Structure is utilized to describe the topological (geometric) relations of structure, while time is utilized to decribe the motions and changes of matter in space, and the flow and exchange of energy. Structure and time parasitize matter and energy to form a hard structure, and the information contained within the machine are "soft" structures, or "spirit", which can parasitize a hard structure or other existing soft structures, act independently, and reuse itself appropriately. Therefore, the organization of these machines allows the maintenance of their self and generation-ordered events, manifesting as thought and behavior. For example, a self-driving car has a hard structure that consists of the chassis, integrated circuits, and driving brain of the car. The soft structures (software) include the programming of the driving brain, maps, and traffic rules. Because the machine can conceptualize time, it can maintain order and function independently and autonomously, which results in the ability to think.

Figure 3 reveals the relationship between matter, energy, structure, and time. The top half is the cognitive space, representing the thoughts of the machine. The lower half represents the physical space, i.e., the actions of the machine. The part located between these two reveals that structure and time parasitize matter and energy to form a hard structure. The wheel is an example of such a (hard) structure. By filling the gap between matter and energy, these hard structures increase the difficulty of separating information and matter. Currently, integrated circuits are the most prominent example of these hard structures, and they represent the "bottleneck" of AI. Soft structures are highly diverse, and they can be categorized into low- and high-level structures. Soft structures are the elements of thought that support abstract, logical (language), and intuitive thoughts; they are manifestations of the imagination and creativity of mankind as well as the spiritual world. Furthermore, these soft structures exhibit a sense of space, time, and hierarchy. Low-level soft structures include symbols, letters, strokes, numbers, front and back, left and right, up and down, ordering, and fast and slow. Soft structures are not natural languages because a child can think even without learning language; linguists call this phenomenon "the language of the heart". Concepts, news, information, and knowledge are all upper-level soft structures because they are mirror and abstracted images of the physical world in the cognitive space, which form a virtual reality. A language is an upper-level tool that conveys human thought. Currently, the reality imagined in the cognitive space is called virtual reality, and the cognitive space the metaverse. Hard and soft structures have been used to fill the gaps between matter, energy, and information, creating an entangled state between matter, energy, structure, and time. This interconnection is analogous to Schrodinger's cat.

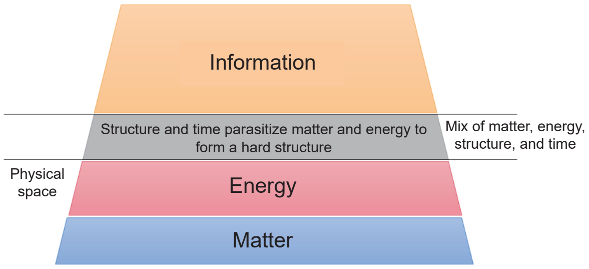
Information
Structure and time parasitize matter and energy to form a hard structure
Mix of matter, energy, structure, and time
Physical space
Energy
Matter

Fig. 3 Relation between matter, energy, structure, and time.

Is intelligence corporeal or spiritual? Let us use music as an example to explore this phenomenon. Sheet music produced by a composer is a soft structure that expresses information, emotion, technique, art, style, and culture. The same sheet music can be expressed using various types of hard structures (instruments), such as violin, piano, or drum, with only the sheet music being constant. Sheet music is spiritual, virtual, and noncorporeal; however, the music heard by human beings in the physical realm is an acoustic art that is material, embodied, and objective being. This physical embodiment is embedded in matter, energy, structure, and time. The combination of these four elements is a manifestation of beauty and the fusion of cognition and action. Furthermore, hard structures can be locally transformed into soft structures, such as virtual robots and hosts. Similarly, soft structures can be locally transformed into hard structures. For example, the robot in the slide and the wheeled robots developed by AIForceTech can interact, learn, and grow autonomously. Therefore, corporeal and spiritual are interconnected, but software cannot define everything.

Schrodinger, who used the physical perspectives of living cells to explain life, speculated that life is a machine. His ideas allow us to understand why machines can be considered alive–this is what we call the "view of the life of machines". The physical layer of life corresponds to the matter layer of the machines; the biological layer of life corresponds to the electronic circuitry and machine instructions in the machines; the psychological layer of life corresponds to the control system and middleware in the machines and the cognitive layer of lifen corresponds to high-level software and data of machines. This reveals the importance of four elements, namely matter, energy, structure, and time. A clock relies on energy; time relies on clocks, order on time, and soft structures parasitize hard structures. Meanwhile, the machines independently achieves automation of thought and autonomously grow their cognitive capabilities via self-reuse. The operation of a machine relies on programming, which relyies on the sequencing of codes, and software on interaction. As sequencing and interactions lead to the generation of negative entropy, machines rely on negative entropy to “live”. As the clock never stops, the machine never stops interacting with its external world, and its thought and cognitive processes will continue endlessly.

An analysis of over 200 definitions of intelligence shows that the term has been “loosely” defined. Intelligence, cognition, or thinking can be broadly defined as the cultivation and transmission of the ability to learn how to solve pre-set problems, interpret, and solve real problems.

In the cognitive space, cognitive and thinking capabilities are attained via computational and memory intelligence, whereas in the physical space, embodied interaction capabilities are obtained through perceptual and behavioral intelligence. Therefore, perception and cognition form an endless loop. In perceptual intelligence, spatiotemporal recognition is the synchronization of positioning, navigation, and time, integrating target and facial recognition. The cultivation and transmission of the ability to solve pre-set problems is learning, considered as a subset of real problems. Once a problem is solved, knowledge will remain. Machines can accept guided teaching and self-learning. To improve the ability to explain and solve real problems, the pre-set problems of where, how, why, and what should be solved.

The learning and operational processes of an intelligent farming machines in a field consists of guided learning tasks such as pre-configuration, assigning tasks, providing guidance, answering questions, cognition through interaction, and supervision. Self-learning is a critical process that involves converting the results of guided learning into long-term memories, undergoing processes such as revision and digesting knowledge. If we refer to guided learning as supervised learning, such a labeling would be an oversimplification of categorizing self-learning as unsupervised learning.

Therefore, the learning processes of an intelligent farming machine should cover three processes: (1) manual operation of the farming machine to teach the robot; (2) robot operates the farming machine with manual intervention from a human; (3) robot operates the machine autonomously and conducts self-learning. The continuous iteration of these processes results in guided learning, semi/weakly supervised learning, and self-learning. In practice, all machine learning processes follow an analogous process. However, previous studies have emphasized the automation of L0 to L5 while neglecting learning, interaction, and growth.

Let us consider the “SenseRobot”, a Chinese chess-playing robot from SenseTime, which beat three generations of grandmasters Hu Ronghua, Xie Jing, and Gu Bowen. The robot has 26 levels of difficulty and more than 100 endgame settings. Furthermore, the robot can autonomously monitor changes on the board and calculate moves accordingly. The dexterous robot can pick up and place chess pieces precisely to the millimeter level. Furthermore, it can respond within seconds, exhibiting perfect “hand-eye” coordination, clean movement, and a tight rhythm. Li Xiaolong, a famour Chines chess player, stated that the SenseRobot is an excellent opponent and training partner. It is not just another AlphaGo program; it is a robot with arms that can pick up and place chess pieces and is equipped with mechanical eyes to see the chessboard. This robot has beaten many Chinese chess champions, and it successfully passed one after another Turing test. But it remains to be seen why SenseRobot was not equipped with voice interaction ability, despite its being equipped with four elements of perception, cognition, behavior, and interaction. Is it able to learn? If the model is placed in a chess institute or made to learn from high-level Chinese chess players, then could it grow and lead to some innovation?

An embodied intelligence grows iteratively from each conversational Turing test it undergoes. Conversational Turing tests are highly diverse and casual, with all coding languages based on natural languages. Thus, Turing’s proposal to use conversations as Turing tests is highly insightful. The use of language marks an excellent achievement for machine intelligence. However, sound, words, and symbols used in coding languages have certain limitations as they are constrained by the axioms of natural-language expressions. Therefore, these models could be somewhat formalized following a “math first and physics second” principle that obeys Gödel’s incompleteness theorems. To impart thinking ability to a machine, the latter’s working language should be formalized; when formalization is conducted, mechanization and the subsequent automation should be performed. Regarding automation, machines can surpass mankind in the profundity of thought.

The Turing test can be performed in many fields and diverse areas. In social dialogues, the Turing test can be conducted in the form of conversations. In literary linguistics, it can be performed by a machine mimicking an actor. In gaming languages, it is using games such as Go. In mathematics, the machine can be tested to define proofs. In arts, the machine could be modified to create pieces of artwork. When it comes to the Tang and the Song poetry, the machine could be requested to create new poems. In law, the Turing test could be performed with the machine providing legal advice. In the physical language, the Turing test can be performed by creating intelligent farming machines. This shows a need for intelligent farming machines to be able to communicate via vocal sound. In the future, conversational Turing tests are expected to be more diverse and casual.

4 Conclusion

The essence of thought is abstraction and association, analogous to creating and linking soft structures. Having intelligent machines learn from “casualized” Turing tests can result in embodied intelligence. With increasing replacement of humans by machines in intellectual and technical tasks, it remains a challenging work to train machines for highly specific positions in various fields and industries. In the future, in addition to learning and working alongside machines, humans can teach machines to learn and work. The result of learning is minute adjustments to the long-term memories of a machine, that is, the network topology of AI neurons and self-learning——a critical process for converting job memories into long-term memories. Machines can easily replicate this process on a large scale and continue learning by themselves. By interacting with machines in such a manner, humans can also learn from machines and collaborate with machines to create and innovate, e.g., robot engineers will create formulations for new materials and robot scientists propose new scientific hypotheses, thus driving scientific and technological discoveries.

The interactive embodied intelligence of machines refers to the intelligence to learn and eventually create. The significance of machine intelligence to human intelligence is like telescope to astronomers and microscope to biologists. By expanding the memory and computational intelligence of humans, machine intelligence can relieve mankind of heavy and repetitive work. Crucially, the capacity of machine intelligence for brute force computation will increase man–machine collaborations to facilitate engineers and scientists for further invention, discoveries, and creation. By then, mankind will not consider whether an innovative technology or paradigm is created by machines or humans.

Article History

Received: 1 June 2023; Accepted: 4 August 2023

References

[1] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg et al. , Sparks of artificial general intelligence: Early experiments with GPT-4, arXiv preprint arXiv: 2303.12712, 2023.

[2] A. Mehrabian and M. Wiener, Decoding of inconsistent communications, *J. Pers. Soc. Psychol.* , vol. 6, no. 1, pp. 109–114, 1967.

[3] A. Mehrabian and S. R. Ferris, Inference of attitudes from nonverbal communication in two channels, *J. Consult. Psychol.* , vol. 31, no. 3, pp. 248–252, 1967.

[4] Y. LeCun, A path towards autonomous machine intelligence version 0.9. 2, https://openreview.net/forum?id=BZ5a1r-kVsf, 2022.

[5] A. M. Turing, Computing machinery and intelligence, in *Parsing the Turing Test*, R. Epstein, G. Roberts, and G. Beber eds. Dordrecht, the Netherlands: Springer, 2009, pp. 23–65.

[6] N. Wiener, *Cybernetics, or, Control and Communication in the Animal and the Machine*. Cambridge, MA, USA: MIT Press, 2019.