Strong Artificial Intelligence and Consciousness

Article *in* Journal of Artificial Intelligence and Consciousness · March 2020

DOI: 10.1142/S2705078520300042

2 authors:

Gee-Wah Ng
DSO National Laboratories
89 PUBLICATIONS **1,298** CITATIONS
SEE PROFILE

Wang Chi Leung
1 PUBLICATION **22** CITATIONS
SEE PROFILE

Journal of Artificial Intelligence and Consciousness
Vol. 7, No. 1 (2020) 63−72

DOI: 10.1142/S2705078520300042

World Scientific
www.worldscientific.com

Strong Artificial Intelligence and Consciousness

Ng Gee Wah*,‡ and Leung Wang Chi†,§

Home Team Science and Technology Agency (HTX)
New Phoenix Park, 28 Irrawaddy Road
Singapore 329560, Singapore
**ng_gee_wah@htx.gov.sg*
†leung_wang_chi@htx.gov.sg

Published 29 April 2020

In the last 10 years, Artificial Intelligence (AI) has seen successes in fields such as natural language processing, computer vision, speech recognition, robotics and autonomous systems. However, these advances are still considered as Narrow AI, i.e. AI built for very specific or constrained applications. These applications have its usefulness in improving the quality of human life; but it is not good enough to do highly general tasks like what the human can do. The holy grail of AI research is to develop Strong AI or Artificial General Intelligence (AGI), which produces human-level intelligence, i.e. the ability to sense, understand, reason, learn and act in dynamic environments. Strong AI is more than just a composition of Narrow AI technologies. We proposed that it has to be a holistic approach towards understanding and reacting to the operating environment and decision-making process. The Strong AI must be able to demonstrate sentience, emotional intelligence, imagination, effective command of other machines or robots, and self-referring and self-reflecting qualities. This paper will give an overview of current Narrow AI capabilities, present the technical gaps, and highlight future research directions for Strong AI. Could Strong AI become conscious? We provide some discussion pointers.

Keywords: Artificial Intelligence; Narrow AI; Strong AI; Artificial General Intelligence; Machine Learning; Reasoning; Active Learning; Emotional Intelligence; Consciousness.

1. What are the Current AI Capabilities?

Artificial Intelligence (AI) is about emulating the human intelligence process by machines. AI has been around for more than 60 years and it has been through several "winter seasons." However, in the last 10 years, AI has showed significant advancements. This is due to three factors: availability of more powerful computing platforms, availability of big data shared from the Internet and electronic gadgets, and improvement in algorithms, particularly in the scalability aspect. Some well-known achievements of AI are as follows:

(1) In 2011, IBM Watson beating the world's greatest Jeopardy Champions.

‡Ng Gee Wah is a staff from DSO National Laboratories who is on secondment to HTX.
§ Leung Wang Chi is a staff from Immigration & Checkpoints Authority (ICA) who is on secondment to HTX

(2) In 2012, a deep neural network showed significant improvement in image classification and won the ImageNet computer vision contest.
(3) In 2016, DeepMind's AlphaGo defeated the best human Go player in the world.
(4) In 2017, Libratus, a Poker AI, defeated a team of human professional players.
(5) In 2017, Waymo's driverless cars had driven four million miles on the road.
(6) In 2018, Deep learning network performed at the same level as radiologists in reading computed tomography (CT) scans [Grewal *et al.*, 2018].

These advances have led to massive investments in AI, from commercial companies to governments in many countries. In the USA, for example, there are AI initiatives by Google, Facebook, Apple, and IBM in diverse areas such as driverless cars, smartphones, social media, cybersecurity, and healthcare; as well as their investments in academic institutions like MIT, CMU, Stanford University, just to name a few, to conduct AI research in cutting-edge topics like deep learning and quantum computing. In China, there is a strong push by the government into AI, with slogans like "match the West in 3 years" and "to lead the world by 2030". In Russia, President Putin proclaimed in 2017 that "whoever leads in Artificial Intelligence will rule the world".

In the defense arena, advances in AI have the potential to change the rules of the game. The US Defense Advanced Research Projects Agency (DARPA) in 2017 characterized AI advancements into three waves, namely Handicraft Knowledge (Wave 1), statistical learning (Wave 2), and contextual adaptation (Wave 3), and emphasized the great advantages of creating the third (future) wave technologies. We envision that AI will significantly contribute to defense capabilities, commercial applications and influence individual life in the 21st century.

2. Narrow AI versus Strong AI

However, current AIs are Narrow AIs. Narrow AIs are good at doing a single task extremely well. An AI that is designed for playing a game of English chess well would not be able to play Chinese Chess, and even less so if the game is changed to Go. Even self-driving car technology is considered Narrow AI, as it is constituted from multiple Narrow AI technologies into the vehicle compute engine. Having said that, Narrow AI remains important and especially useful for humans on laborious and repetitive tasks, as machines are more efficient and accurate, which in turn shortens the decision cycles of the human.

However, we would like to push to build a Strong AI (also known as Artificial General Intelligence (AGI)) that can perform like a human or possess human-level intelligence. Such Strong AI would encompass the ability to reason, understand context, and perform complex tasks that the Narrow AI of today are unable to comprehend or perform. Strong AI system must also exhibit robust and adaptive behavior, and be able to perform a variety of tasks at different situations. Hence, building a Strong AI is the grand challenge in the AI community. One approach to

Strong AI is the cognitive architectures [Chong *et al.*, 2007; Baum, 2017]. Thus far, there are more than 45 such cognitive or AI related architectures that provide some promising directions towards building generic and scalable architectures and many of these mimic human brains. In the next section, I would like to provide some further thought on building Strong AI.

3. Possible Future AI Research in building Strong AI

Strong AI that have human-level intelligence need to have processes that can sense, understand, reason, learn and act in the environment in ways similar to how humans can intelligently do. Strong AIs cannot be achieved by simply combining many Narrow AI technologies into a system. The Strong AI must take into account the broader context of its operating environment, in addition to the task that it is instructed to accomplish. The real operating environment is ever changing and it includes unknown elements that are not present during the training phase. The ideal Strong AI should be able to reason and adapt to the unknown elements, and react appropriately to it without critically failing. This adaptability coincides with DARPA's 3-wave characterization of AI development whereby the third (future) wave is termed "Contextual Adaptation" (see DARPA perspective on AI). According to DARPA's definition, future AI systems capable of contextual adaptation can construct explanatory models for classes of real-world phenomena, and are able to comprehend what and why certain decisions are made, and possibly to use the models to abstract further from data.

While Strong AI remains in the realm of active research, humans are born capable of contextual adaptation. Humans achieve this through making decisions and interacting with the environment. The popular decision-making tool in the military domain is the Observe Orient Decide Act loop or OODA loop (Fig. 1). The OODA loop was developed by military strategist John Boyd to analyze combat operations. It has been extended successfully to other fields such as business strategy and law enforcement. The OODA loop in essence is a model of individual and organization

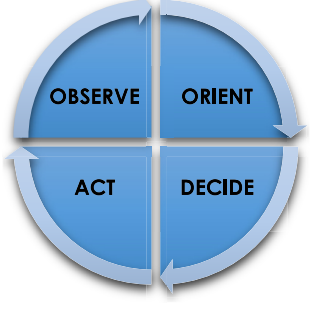

Fig. 1. OODA Loop, a human decision-making model that enables contextual adaptation, a key feature for Strong AIs.

learning and adaption. It serves as one of the useful guides in highlighting the current gaps and future research trends to achieve Strong AI.

3.1. *Observe: Sentient surveillance systems*

To *observe* is more than just to see. It is the active surveillance of the situation, which is the combination of focusing attention at key areas and making connections between disparate sources. In the big data era, the exponential growth in sensory data available has outpaced humans' ability to manually process it. Today, AI systems are deployed to aid human operations by automating the manual processes in a single system. Tomorrow, the mass proliferation of sensors will mean that the AI will be needed to orchestrate the data collection effort on top of processing in multiple systems and domains.

At present, AIs' successes in a specific task prediction performance usually involve deep neural networks that depend on having large amounts of annotated training datasets and high-quality data. However, in certain contexts, it can be difficult, if not impossible, to obtain large training data. Hence, the AIs of tomorrow will have to overcome this hurdle by employing the techniques of transfer learning [Zhang *et al.*, 2017], or domain adaptation and "few-shot" learning. For example, in the prediction of age from images of faces, Rothe *et al.* [2016] starts from a neural network trained on general images (14 million images in ImageNet); adapts it to the task of face and age estimation using data automatically obtained from the Internet (0.5 million age-annotated face images in IMDB-WIKI); and tunes the model for the particular dataset of 2.5 thousand images. The extreme case is training AIs with *no data*, in what is called *zero-shot learning* [Fu *et al.*, 2018]. In such a case, the adaptation is done via constraining the algorithms and extracting features of the algorithm to be used in the new task. This requires mimicking human ingenuity in understanding the relation between different tasks and exploiting the latent features induced by the learning algorithm.

3.2. *Orient: Emotionally intelligent cyborgs*

To orient to a new situation, is to analyze past experiences and synthesize them with the current observations. The real operating environment is not a clinical environment. It is further complicated by human actors, who have different dynamic behaviors. To have the correct orientation in a complex situation requires tacit understanding of human emotions.

The AI needs to be emotionally aware to be an effective team player alongside its human counterparts. It needs to comprehend human idiosyncrasies and biases in complex situations, and exploit human weaknesses when necessary. The aim is not to build a machine that can cry or express its emotions to make it look more human-like, but rather to detect and recognize a particular state of emotion the human user is in and use the additional information to adjust its actions. In addition, human emotions

can lower the quality of his decision-making ability, and at the critical juncture, the AI system can step up to complement or augment the decisions made. Emotional intelligence in AI systems can also have other usages like detecting the emotional states of social groups or even to encourage pro-social behavior in a population by the use of large-scale emotional AI agents [Paiva *et al.*, 2018].

3.3. *Decide: Imaginative machines*

Decide is the generation of the courses of actions or hypotheses after orientating to the situation. Hypotheses generation is not a regurgitation of experiences, for the environment is dynamic and reactive. In the military context, the thinking enemy is part of the equation. Inspired decision-making requires imagination, one of the hallmarks of human intelligence. With the ability to imagine, humans can mentally transcend space and time, and dream of something novel. Current advancements in data analytics address the "what is" and some of the "what" that are embedded in the data. With imagination, it can bring out the "why". In addition, an imaginative AI is able to create new problems to solve, which could be viewed as a form of exploration with no explicit objective functions, and could eventually lead to better methods and solutions.

One area of active research is learning affordance — the ability to perceive an object as in not only its current state, but also what can be done with it. This gives the machine the flexibility it needs to "think out of the data" and manipulate objects or its environment beyond statically defined tasks. While this idea of an imagination machine is not a novel idea, these ideas have been explored in an area known as computational creativity. This creativity has been expressed in areas such as music, novels, paintings, and engineering. We hope future AI could "imagine" and produce strategies and operating concepts, aid in the design of advanced materials and equipment, and aid in cryptanalysis by thinking out of the data.

3.4. *Act: Robots/actuator*

Act is to carry out the decided course of action. For the modern military, "act" includes the use of unmanned systems and robotics. Today, robotic systems that operate autonomously are typically pre-programmed to perform well-defined actions. Future robotic systems have to operate autonomously under varying mission outcomes that deviate from preprogrammed tasks. Mission planning that is dynamically modifiable is a requirement for the autonomous system to cope with changing mission outcomes. Additionally, dynamic re-tasking of sensors and payloads may be needed to support aided target recognition, identification and tracking. Precise navigation and timing is another key enabler for autonomous systems operation that allows understanding of the operational area, sensor cueing and collision avoidance. Cognitive techniques and algorithms are needed to integrate sensing, perceiving,

analyzing, communicating, planning, decision-making, and executing, in order to fully realize the operational benefits of robotic systems.

In the near future, robotic systems have to work autonomously and collaboratively in executing missions. For example, applications such as surveillance and reconnaissance, search and rescue, mapping unknown environments and payload transportation can be performed by having a collaborative system consisting of Unmanned Surface Vehicles (USV) and Unmanned Aerial Vehicles (UAV). By having a USV−UAV collaborative system [Zhu and Wen, 2019] which leverages on their complementary properties, it provides humans at the higher command center with richer information to achieve the mission of a patrol system.

3.5. *Closing the loop: Self-referential and self-reflecting AI system*

After every action by the AI system, the environment would react. The AI system would then observe those phenomena and the decision cycle repeats. This closes the OODA loop. However, without a means of self-inspection, such an AI system would be susceptible to being trapped in a closed loop. A common quote that perfectly sums this up is "insanity is doing the same thing over and expecting a different result". Fortunately, humans have the capacity to exercise introspection, to learn and improve their internal states and purpose. The introspective ability is a defining feature of future Strong AIs.

Introspection can be defined as self-referencing and self-reflecting. Building Strong AI with self-referential ability will allow its function to refer to itself, like recursion in programming, but the ultimate aim is to give it its own ability to modify its function by rewriting its own source code. An AI system with self-referential ability should be able to represent its own state in its own native memory representation. States within the AI system could represent error rates, utility values over time, metadata pertaining to the environment, and an embedding of the system's finite state machine.

4. Man−Machine Collaboration

To achieve Strong AI would require a partnership between Man and Machine. For example, to imbue AIs with emotional intelligence would require factoring of humans' idiosyncrasies and biases. This can be achieved with interactive reinforcement learning where the human acts as a critic of the AI training process. However, this interactive process, while necessary, may introduce "human tendencies" problems such as:

- Humans tend to reward good actions that have diminishing returns over time but consistently punish bad actions.
- Reward hacking (gaming the system).
- Contradictory constraints (yes means no).

Additionally, humans are subjective, and behavioral differences between different people might taint the human-in-the-loop learning process. When building AI with emotional intelligence, the training phase with human-in-the-loop should be designed around human trainers rather than the human trainers simply being a small component in the training process.

For an effective partnership between Man and Machine, the evolving AI agent needs to be trusted by the human. Such trust is more readily established if the agent can explain the manner in which it arrives at its decision or recommendation. While the previous generation of tools based on their human engineered set of rules can generate plausible explanations, the increasing use of deep learning methods due to their superior performance is making this harder to accomplish. To overcome this, DARPA has started a four-year program on *Explainable Artificial Intelligence* [DARPA, 2016]. With explanations, one may envision tighter man−machine interaction to achieve better results. For example, Malasky [2005] has designed a resource allocation and planning strategy where humans are involved in selecting options and variables proposed by an algorithm in an iterative manner. For the task of inferring the relation between two entities in a natural language text, Sameer [2017] has proposed that humans study the chain of reasoning given by the algorithm and correcting only the part of the reasoning that is wrong. Enabling humans to examine the explanations and correct the algorithms is one manner to overcome the challenge of low-data regime. These are examples of Man−Machine collaborative reasoning, which is part of an emerging field known as collaborative and coactive AI systems. The idea is to bring out the best of both worlds by considering the advantages and disadvantages of both human and machine processes when designing AI systems.

5. Reliable and Safe AI systems

As AI sees wide adoption, there will be an increased competition in AI systems, not only between industries and countries, but from malicious actors as well. Adversarial AI is an emerging field that uses AI techniques for the attack and defense of AI models. Examples of attacks can range from evading spam filter, to tricking the AI into seeing something that is not, like misclassifying a turtle as a rifle [Athalye *et al.*, 2018]. Such adversarial attacks could be employed by terrorists to avoid automated detection when publishing online materials or by adversaries to sabotage the perception system of autonomous vehicles during times of conflict. On the other hand, there are advances in adversarial defenses, too. This form of defense includes preprocessing-based methods to remove the adversarial signal, adversarial training to make the AI more robust, and the use of conceptual reasoning to form a secondary check. Future AI systems will likely be complex and incomprehensible to most humans. Thus, other AI systems would be needed to secure the operational AI systems using adversarial defense techniques and be qualified as safe from adversarial attacks.

6. Can Strong AI achieve Consciousness?

So far, my pointers on building Strong AI are largely from an engineering point of view. Building AI systems as an aid to humans and more towards a man−machine collaboration such as augmented intelligence, leads to the amplification of human intelligence. It is more human centric than the AI system centric design. The AI system centric design concept has the ultimate quest of building a Strong AI or AGI where the AI has superintelligence, autonomy and human-like consciousness. Can humans build such an AI system that has consciousness? I would like to provide the following pointers for discussion.

(1) What is consciousness and why are we conscious? Thus far, nobody has any idea exactly what is consciousness in humans. Thus, that brings us back to the basic problem, i.e. we do not have the know-how to design the AI system with the same consciousness as humans, since we do not know it ourselves. However, assuming in the future, humans are able to use material means to design Strong AI-like human intelligence, can the Strong AI intelligence achieve consciousness? That leads to the fundamental question of whether pure material means of intelligence give rise to consciousness; or is consciousness more than just material? This is related to the Mind–Brain (or Mind–Body) problem. Lewis [1996] said, "Consciousness is either inexplicable illusion, or else revelation."

(2) On the Mind–Brain Problem. The Mind–Brain problem breaks down into two possibilities, i.e. the materialism view and the non-materialism view. However, quantum physics debunks the materialism view: Physicist Wigner [1995] pointed out that, "...while a number of philosophical ideas may be logically consistent with present quantum mechanics, materialism is not." Hence, the fundamental nature of reality leads us to the fact that consciousness is not physical [Wheeler and Zurek, 2012]. Likewise, Nagel [2012] and studies from neuroscientists [Dirckx, 2019; Hoffman, 2008] indicated that brain activity itself is not able to explain the conscious experience.

    Hence, we are faced with the question of how our minds have self-awareness and self-dialogue capabilities, a mind that asks questions such as, "What, Who and Why am I?" Furthermore, Planck [1931] pointed out, "All matter originates and exists only by virtue of a force... We must assume behind this force the existence of a conscious and intelligent mind." Hence, there exists something beyond materialism.

(3) Intelligence and Consciousness. In our biological brain, the intelligence and consciousness that have "conscience" are connected. In today's AI system, the designs of intelligence and "conscience" are separated; and we do not yet know how to design it as a tightly connected structure. As intelligence and consciousness are linked and not separated, our purpose and value of doing intelligence work are governed by our conscience. Would AI systems be able to have this capability? This leads us to the point on legal and moral law.

(4) Legal and Moral Law. One of our greatest fears of building Strong AI is the total autonomy of the AI system. Hence, we have debates on legality and morality for autonomous systems such as self-driving cars if an accident were to occur on the road. Can we mirror the human system by building morality rules into the AI system? Picard [2000] said, "The greater the freedom of a machine, the more it will need moral standards." Can AI systems reach full autonomy and have freedom to make its own decisions?
(5) Full AI System Autonomy. Boddington's account [2017] reflects the human situation, "For if we see the Genesis account of the Fall of man as foreshadowing our fears about robots, then Genesis gets the problem exactly right, for exactly the right reasons — it's a worry about autonomy itself: what might robots do if we can't control them fully? Will they adhere to the same value system as us? Will they decide to disobey us? What will our relationship with our creations be? We can thank the Hebrew account of Genesis for pre-warning us thousands and thousands of years ago." Boddington brought up an interesting point on human freedom and turning his back to his creator. However, this will not happen in the AI system because of the gaps mentioned and at best, governing rules will be built into the AI system in its decision process.

The ascending complexity of the above points is for us to ponder on when we build the AI system. The AI system will continue to advance and better aid humans. However, unlike a human brain, it is made of non-living organism and does not possess the non-physical dimension, and hence, we argued that it will never possess human-levels of consciousness. We should not be fearful of AI advancement, and avoid being misled by speculation or misinformation. I would like to end this discussion with the following three quotes:

(1) Rees [2018] — "We can have zero confidence that the dominant intelligences a few centuries hence will have any emotional resonance with us — even though they may have an algorithmic understanding of how we behaved."
(2) Brooks [2017] — "I think people see how well [an algorithm] performs at one task and they think it can do all the things around that, and it can't. For those of us who do work in AI, we understand how hard it is to get anything to actually work through product level."
(3) Prof. Joseph Macrae Mellichamp (University of Alabama) — "The 'artificial' in artificial intelligence is real."

7. Conclusions

Narrow AI has come a long way and it has started to bear fruit in some applications. We will continue to see exciting progress in engineering AI for more applications in the 21st century, and the continual quest to achieve Strong AI. Human−machine collaboration will be a more probable approach as the focus is still on the human.

References

Athalye, A., Engstrom, L., Ilyas, A. and Kwok, K. [2018] "Synthesizing robust adversarial examples," in *Proc. 35th International Conference on Machine Learning (ICLM 2018)* (Stockholm, Sweden).

Baum, S. [2017] A survey of artificial general intelligence projects for ethics, risk, and policy, global catastrophic risk institute Working Paper 17-1, *https://ssrn.com/abstract=3070741*.

Boddington, P. [2017] Myth and the EU study on civil law rules in robotics, *https://www.cs.ox.ac.uk/efai/2017/01/12*.

Brooks, R. [2017] Artificial Intelligence is not as smart as you (or Elon Musk) think, *www.techcrunch.com/2017/07/25/artificial-intelligence-is-not-as-smart-as-you-or-elon-musk-think*.

Chong, H. Q., Tan, A. H. and Ng, G. W. [2007] "Integrated cognitive architectures: A survey", *Artif. Intell. Rev.* **28**(2), 103−130.

DARPA [2016] Broad Agency Announcement — Explainable Artificial Intelligence (XAI); DARPA-BAA-16-53, *https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf*.

Dirckx, S. [2019] *Am I just my brain?* (The Good Book Company).

Fu, Y., Xiang, T., Jiang, Y. G., Xue, X., Sigal, L. and Gong, S. [2018] Recent advances in zero-shot recognition — Towards data-efficient understanding of visual content, in *IEEE Signal Processing Magazine*, pp. 112−125.

Grewal, M., Srivastava, M. M., Kumar, P. and Varadarajan, S. [2018] "RADNET: Radiologist level accuracy using deep learning for hemorrhage detection in CT scans," in *Proc. IEEE 15th Symposium on Biomedical Imaging (ISBI 2018)*, pp. 281−284.

Hoffman, D. D. [2008] Conscious realism and the mind-body problem, *Mind Matter* **6**(1), 87−121, www.cogsci.uci.edu/ ddhoff/ConsciousRealism2.pdf.

Lewis, C. S. [1996] *Miracles A Preliminary Study* (Simon & Schuster Trade).

Malasky, J. S. [2005] Human machine collaborative decision making in a complex optimization system, *https://dspace.mit.edu/handle/1721.1/32514*.

Nagel, T. [2012] *Mind & Cosmos: Why the Materialist Neo-Darwanian Conception of Nature is Almost Certainly False* (Oxford University Press).

Paiva, A., Santos, F. P. and Santos, F. C. [2018] "Engineering Pro-Sociality with autonomous agents," in *32nd AAAI Conference on Artificial Intelligence* (New Orleans).

Picard, R. W. [2000] *Affective Computing* (The MIT Press).

Planck, M. [1931] *Das Wesen der Materie* [Speech on The Nature of Matter], (The Observer, London) p. 17.

Rees, M. [2018] *On The Future: Prospects for Humanity* (Princeton University Press).

Rothe, R., Timofte, R. and Van Gool, L. [2016] Deep expectation of real and apparent age from a single image without facial landmarks, *Int. J. Comput. Vis.* **126**, 144−157, doi: 10.1007/s11263-016-0940-3, https://link.springer.com/article/10.1007.

Wheeler, J. A. and Zurek, W. H. [2012] *Quantum Theory and Measurement* (Princeton University Press).

Wigner, E. P. [1995] *Philosophical Reflections and Syntheses* (Springer-Verlag).

Zhang, W., Fang, Y. and Ma, Z. [2017] "The effect of task similarity on deep transfer learning," in *24th International Conference on Neutral Information Processing (ICONIP 2017)* (Guangzhou, China), pp. 256−265.

Zhu, M. and Wen, Y. Q. [2019] Design and analysis of collaborative unmanned surface-aerial vehicle cruise systems, *J. Adv. Transp.* **2019**, https://doi.org/10.1155/2019/1323105.