The Application of Artificial Emotions in Artificial Intelligence

Yunze Gu

[1]*Department of Philosophy, King's College London, London*
**Corresponding author. Email: yunze.gu@kcl.ac.uk*

ABSTRACT
The current research effort on AI is mainly focused on how to more efficiently completing specific tasks. Some incredible achievements can already be seen in navigation, data prediction, personalized advertisement and more. However, researchers seem less motivated to study how AI behaviour can be made more like an actual human. The academic engagement on this specific issue is commonly seen in Philosophy and Cognitive Science. This paper shall attempt to establish a relationship between the human memory system and Artificial Intelligence, which helps us understand how human cognition can be reduced to a computational process. The paper introduces some historical psychologic studies of memory and learning, including Pavlovian Conditioning, H.M., Lashley's Law of Mass Action. Some basic ideas in the philosophy of mind and neurophilosophy are also discussed to build the argument. The conclusion of this paper suggests that there is far more computational process hidden behind our daily behaviour than what we tend to perceive. From a behavioral standpoint, artificial intelligence is perfectly capable of simulating human actions. It is also worth noting that studying AI as a bionic technology allows a certain level of ignorance. The paper is built on the assumption that a human-like AI is possible without a complete and thorough understanding of brain science.

Keywords: *Artificial Intelligence, Neurophilosophy, Memory, Artificial Emotions*

1. INTRODUCTION

Humanity is currently living through a boom of computer science. Ever since the establishment of the first MIT cognitive science department in 1952, researchers have been improving our computing power using the study of cognition/brain science. The current stage of artificial intelligence has already made substantial changes to our daily life. From the finical market predictions to personalized advertising, computers seem to be taking over human works at an increasingly faster rate. However, the motivation for making AI behaviors more human-like appears to below. Indeed, the market demand for this type of study can hardly compare with task-specific research. This paper shall cover historical studies in psychology, including the Pavlovian Conditioning, Lashley's Law of Mass Action and the famous H.M. case. Throughout the introduction to the psychological study of memory and learning, issues that relate to AI's working memory and categorization of memory types shall be philosophically analyzed. The main topic of the discussion is focused on how human cognition is fundamentally similar to a pure computational process. The paper shall use the case studies mentioned above to analyze the similarity between the animal memory system and liner computer programs. Following that, the paper shall also touch on the prescriptive feature of human emotions and how it can be the causal link between scripted cognitive functions and unpredictable human behaviors. The combination of knowledge in behavioral psychology, neuroscience and philosophy can hopefully assist in establishing the argument. The aim is to analyse our present knowledge on learning and memory and then find out how they can be potentially transformed into computer science. As for the meaning of this type of research, although the current society does not require, or even sometimes resists, the idea of a human-like AI, the effort in forging consciousness with artificial means is essential for a possible utopian future.

2.THE HISTORICAL STUDY OF MEMORY AND ANIMAL LEARNING

The psychological study of memory can be traced back to William James' conception of brain-paths. As one of the earliest scientific attempts on understanding memory formation, James suggested that memory can result from physical paths formed in the brain, which is strengthened by the persistency of these paths[1]. Although James' proposal can only serve asintuition as it lacks anatomic evidence, the idea of engram remains influential until today. One important aspect of James' intuition is the proposal of memory enhancement; it draws a line between real-time experiences and stored experiences. However, memory can hardly be unified with a single line of definition. Many distinct mantel types that have come to be called memory are fundamentally different from one and another. At the early stage of the research, Ivan Pavlov designed a set of experiments targeting the dog's learning ability. Learning obviously can be seen as a form of memory formation. Pavlov's experiments are intended to trigger dogs' salivation with an artificial signal. Pavlov concluded that learning occurs as the result of associations forming between the signal for food and the actual food. The study of this particular type of memory formation is latter to be called Pavlovian Conditioning.

The consensus is that engram is located somewhere in the brain. However, the early effort in pinpointing the area where memory is stored was largely unsuccessful. Following the research in conditional learning, Karl Lashley attempted to find out the correlation between animal's conditioned responses with knife cuts that are strategically applied to areas of their brain. This simple method of elimination didn't help Lashley to find any specific part of the brain that is responsible for learning. At the end of this effort, he had to conclude that "in reviewing the evidence on the localization of the memory traces, that the necessary conclusion is that learning just is not possible[2]." Obviously, learning is indeed possible. His experiments have shown to the world of psychology that the localized view of engram has little potential for understanding memory. He then famously suggested the law of mass action, where he points out that the function of memory is being distributed rather than located in a specific area.

3. THE STUDY OF HUMAN MEMORY THAT INDUCES HUMAN-LIKE BEHAVIORS IN ARTIFICIAL INTELLIGENCE

3.1. The Application of Memory in Artificial Intelligence

In a 2016 paper published by Google DeepMind UK, the researchers identified a new type of computer by giving it a kind of working memory. The authors said that "In contrast to computers, the computational and memory resources of artificial neural networks are mixed in the network weights and neuron activity. This is a major liability: as the memory demands of a task increase, these networks cannot allocate new storage dynamically, nor easily learn algorithms that act independently of the values realized by the task variables[3]." This quotation gave a good overview of how neural networks are superior to a traditional computer in many ways. More importantly, it shows that neural networks have an architectural advantage of being more similar to the principle of a biological neural network seen in animal brains. The authors also noted that "neuro-scientists argue that neural networks are limited in their ability to represent variables and data structures." To challenge this view, Google researchers started providing AI with read-write access to an external memory. The results in the paper are promising; their AI with external memory system learned how to mathematically use the stored memory to significantly improve the efficiency of certain computational tasks. This successful implementation of memory in AI has led to much more comprehensive research on similar topics. The popularity of memory-related research remains high today. With the help of this approach, it is now possible for AI to resolve much more complicated workflows, such as learning how to beat a video game. More practical applications of this approach are still rare as classical machine learning remains efficient enough for the tasks given. However, the classical approach focuses on a more mechanical part of cognition, such as sensory processing, sequence learning and reinforcement learning. The potential for the memory application may be better seen in the future effort of forging a true, human-like consciousness.

3.2. The H.M. Case and Subdivisions of Human Memory System

In the 1950s, the famous H.M. case presented a challenge to the distributed model of memory. To treat his severe epilepsy, Henry Malmaison had a bilateral medial temporal lobectomy to remove a large portion of his hippocampus. Although the surgery was successful in curing epilepsy, H.M. had become unable to form new memories. Hippocampus is located in the limbic system, which is traditionally thought to be an area in charge of processing emotions. After the H.M. case, people started to suspect that the hippocampus or limbic system might be related to memory formation [4]. As previously mentioned, memory is better seen as a collection of functions. This notion is first proven by the mirror drawing task of the H.M. case. This task trains the participant to draw a simple object without looking directly at the paper but through a mirrored reflection. The task is usually poorly performed at first, but participants can learn the skill fairly quickly after some

practice. In this experiment, H.M. is asked to perform the task regularly, although he couldn't remember ever doing such a task, the skill seems to persist as he cancomplete the task at an increasing speed each day[5]. This result suggests that H.M. could still form new implicit memory after the hippocampus damage, which led to a later discovery that shows the basal ganglia are handling procedural memory.

The sophistication of the human memory system makes it very hard to categorize different types of memory accurately. The early studies tend to focus on animal learning, which is of course a representation of memory. But, as the H.M. case finely sorted the subdivisions such as episodic and semantic memory, it has become more challenging to describe human memory as a linear computational process. From the H.M case, it has been discovered that memories are likely to be processed separately as the hippocampal lesions suffered by H.M. only resulted in the inability of forming new episodic memory. Therefore, it is undoubtedly a challenge for AI research efforts to reconstruct each subdivision of memory and make it a coherently integrated body.

3.3 The Implementation of Emotion in Specific Subdivisions of Memory

When one single type of memory is examined separately, such as the procedural memory, the function of this type of memory suddenly seems more computational than when it is well-integrated as a part of human cognition. For example, a healthy, functional human being is capable of learning various tasks such as playing the piano. As a result, the act of playing the piano is commonly recognized as humane behaviour. However, when a piece of music is played on a computer or being physically played by a pianola, the extreme accuracy and the lack of randomness make the same action seem less humane. Procedural memory is arguably the most accessible mental type to be mechanized as it carries a mere functional role. Following the same rationale, other memory types, when being separately examined, may also show their computational nature. I believe that the key to making a human-like AI is to simulate the randomness of human behaviour. When examined closely, it is not hard to realize that the unpredictive nature of our daily actions may partially be the result of emotional changes. With a deterministic assumption of the mind, we can then conclude that there is no true randomness in human behaviour. If emotional instability is the main factor that causes such randomness. With the help of machine learning to generate the causal relationship between the external world and the appearance of painful/pleasurable reactions, AI has the potential to perfectly simulate human behaviour.

4. SIMULATION OF CHARACTERISTICS OF ANIMAL LEARNING AND BEHAVIOUR

4.1. A Brief Introduction to Reinforcement Learning

Memory can be better described as a collection of brain functions rather than a single process of storing information. For example, the Pavlovian conditioning experiment can trigger a conditioned response of salivation by using a conditioned stimulus. Although this learning process reflects some aspect of memory, it only seems to be showing a single, natural and subconscious response. Following that, learning of non-innate behaviors was carried out and was called operant conditioning. Thorndike's puzzle box showed that animals could be trained to solve complex puzzles which are not their natural behaviors. It is shown from Thorndike's experiment that through reinforcement learning, the time in which the test subject takes to solve a puzzle decreases gradually. This means that non-innate behaviors can be learned, and the memory of such behaviour can be reinforced through reward or punishment.

4.2. How the Hedonistic Element of Reinforcement Learning can be Applied to Computer Science

AI research that focuses on simulating human consciousness, just like other bionic technologies, usually does not require a complete understanding of the principle of operation of the human brain. Therefore, the behavioral data observed during these studies can significantly assist in building a memory system for AI. The methods of reinforcement learning usually consist of the element of reward and punishment. According to hedonism, all living things tend to have an innate desire for pleasure and fear for pain. This view is widely discussed in the field of philosophy of mind. Although the hedonistic assumption can be debated, criticisms of this idea mainly land on its moral justifications. When it comes to describing individual actions, however, the hedonistic argument holds incredibly well. Therefore, the animal behaviour when facing pain and pleasure can be recorded then programmed into an AI, which may work as a universal rulebook to limit or allow certain actions. The benefit of this method is that there is no need for programmers to painstakingly line out every possible activity then decide whether it is acceptable for an AI to perform such action. Alongside machine learning, the underlying pattern of what actions cause pain or pleasure can be perpetually generated, allowing AI to encounter new environments in the same way as humans.

4.3. Advantages of Simulated Emotion System in Artificial Intelligence

The method of punishment and reward in the traditional animal learning experiments has recently been introduced to develop machine learning. The principle of training a neural network is fundamentally similar to animal reinforcement learning. The programmers can first recognize correct and faulty results and then allow the neural network to procedurally repeat the calculation until most of the generated results are correct. Commonly, this process is understood as helping the program mathematically identify higher-level patterns of a valid result. Moving back to the animal learning experiments, more precise effects of punishment or reward may have been ignored. Multiple types of punishment or reward only carry a single instructional function. For example, an electrical shock can communicate to the test subject when an option is incorrect. Similarly, other punishments or rewards serve the same purpose. However, each specific type of punishment or reward does a lot more than simply giving instructions to the test subject. When an external interaction is applied to an animal, human included, it triggers multiple emotional changes that all have unique functions that affect the cognition process. The instability of animal cognition may also be applied to AI. Instability is not suited for existing AI tasks such as sequence learning. Still, it may be vital for AI to gain a similar behaviour pattern to humans, which can be applied to practical usage such as companion robots.

4.4. Additional Aspects of AI Development that Require An Emotion System

Other than the points mentioned above, emotions may play a much broader role in artificial intelligence. In her work "affective computing," Ross Picard pointed out the importance of human effect and how computers should be made "effect-aware"[6]. Matthias Scheutz briefly summarised the possible usage of emotions in AI in a serval way. It may be essential for animated characters to be more believable. The ability to correctly recognizing human emotions is crucial for a system to be adapted to the users' needs [7]. Applications as such are focused on the human-AI interactions. From a user's perspective, an emotionally relatable AI may be far superior to what is available today. However, it may be hard to achieve an automated emotional simulation if the research only focuses on how the user perceives the behaviour of AI. This is similar to the weakness of the Turing Test. When the goal of the test subject is to mimic human behaviour and fool the examiners superficially, the direction of research may lead to a single functional approach of emotion. Scheutz also mentioned that "others take emotions to be an integral part of the control of complex agents, and thus focus on architectural mechanisms that are required for emotion processes." [7] The functions of emotion are highly versatile; some may argue that it plays the role of central control for complex behaviour. This can include action selection, motivation, goal management and memory control. From those functions that can be observed from human behaviour, the potential of emotions in AI can hardly be ignored.

5. CONCLUSION

The paper started by analyzing William James' intuition on how memory is being processed in animals. He suggested the concept of the brain-paths, which was one of the first physical representations of memory. His vision provided an important starting place for later scholars such as Ivan Pavlov to scientifically study the formation and storage of memory. Pavlovian Conditioning achieved scientific proof of learning. By using conditional stimulus, conditioned response was successfully observed in animals. Despite its limitations, a simple act of learning represents the ability to store external information that can affect one's behaviour. The methods of reinforcement learning are used to train neural networks. This successful methodology transformation may show a high possibility that animal consciousness can be represented mathematically. As of the current research stage, some human-like cognitive abilities can already be realized by well-trained neural networks, some of which greatly exceed human capability. However, as previously discussed in the paper, the current research efforts tend to focus on improving the efficiency of a specific function like sensory processing. However, the day-to-day behaviour of any animal is the result of various functions working together harmoniously. As more and more individual functions of an animal are successfully simulated, the need for central control that unifies those functions becomes urgent. An important point noted in the paper is that the instability of natural organisms is partially what distinguishes animals with programs, and emotion is in some way both unstable and prescriptive. Suppose a dynamic system of complex emotions were to be simulated and implemented in the AI. In that case, the versatility and the unpredictive nature of human behaviour may also be possible in a carbon-based life form.

REFERENCES

[1] James, W. (1918). The principles of psychology. New York: H. Holt. p695.

[2] Lashley, K. S. (1950). In search of the engram. In Society for Experimental Biology, Physiological mechanisms in animal behavior. (Society's Symposium IV.) p454–482. Academic Press.

[3] Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A.,

Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., Badia, A. P., Hermann, K. M., Zwols, Y., Ostrovski, G., Cain, A., King, H., Summerfield, C., Blunsom, P., Kavukcuoglu, K., & Hassabis, D. (2016). Hybrid computing using a neural network with dynamic external memory. Nature, 538(7626), 471–476. https://doi.org/10.1038/nature20101.

[4] Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. Journal of Neurology, Neurosurgery & Psychiatry, 20, 11–21. https://doi.org/10.1136/jnnp.20.1.11.

[5] Squire, L. R. (2009). The legacy of PATIENT H.M. for neuroscience. Neuron, 61(1), 6–9. https://doi.org/10.1016/j.neuron.2008.12.023.

[6] Picard, R. (1997). Affective Computing. Cambridge, MA: MIT Press.

[7] Scheutz, M. (2013). Artificial emotions and machine consciousness. The Cambridge Handbook of Artificial Intelligence, 247–266. https://doi.org/10.1017/cbo9781139046855.016.