Title:

The criteria for Artificial Intelligence (AI) to possess consciousness and self-consciousness as conditions of personhood

Research question:

How should we set the criteria for artificial intelligence (AI) to possess consciousness and self-consciousness as conditions of personhood?

Subject: Philosophy

Word Count: 4000

Table of Contents

Introduction

Recently, artificial intelligence (AI) has become one of the growing fields of interest in not only science and technology, but also in its philosophical and social implications. There are generally two types of AI: weak artificial intelligence (weak AI) and artificial general intelligence (AGI). Weak AI denotes AI that only serves certain given purposes well, but lacks the so-called general intelligence across various tasks, which is possessed by most human being; AGI denotes AI possessing human-level intelligence capable of performing "general intelligent action" (Newell and Simon 116). Today, although its archetypes are still resource-intensive, AGI is not too far away from civil use (Goertzel 27-28). At the threshold of its advancement, it is of significance and practical value to discuss the condition for AI to possess personhood. Future laws, regulations, and social morality system will all need reference from AI's criteria to be considered as a person. Consciousness and self-consciousness have always been considered as fundamental and essential properties of personhood because of the close relationship between being sensible and being able to think as a person (Locke 44) and the characterization of a person as "a thinking intelligent being that … can consider it self as it self …" (Locke 183), respectively. Therefore, it is important to address these two qualities for AI before addressing other properties such as agency, morality, etc. In this work, I will discuss the criteria for AI to possess these two properties of personhood thus shedding light on future researches on this crucial topic.

Consciousness and self-consciousness are two crucial qualities of philosophical concern that has constantly been debated since the dawn of humanity. From *Oedipus Rex* to Descartes, from Aristotle to Plato, from Hume to Kant, countless influential and thoughtful philosophers have touched upon the topic of consciousness and self-consciousness, and developed numerous theories and perspectives. However, since the prevalence of science after 18th centuries, many

fields, previously of purely philosophical concerns, fell into the intersection of philosophy and science, so does the issue of consciousness and self-consciousness. In this paper, I will discuss these two concepts mainly with philosophical thoughts in combination with supportive scientific facts as evidence.

The nature of consciousness has been hotly debated. In this part, I will examine different criteria for a cognitive system to possess consciousness from various perspectives, and then arrive at a conclusion from the combination of phenomenology and scientific psychology, providing a criterion for AI to possess this quality.

The view on self-consciousness, another similar yet more contentious topic, has been divided into two main parts: an intuitive, introspective consciousness of the self and a functional, empirical consciousness of the self. I would argue for the latter from the combination of Hume's "bundle of perception" view with my own improvements and Kantian account of being aware of a transcendent ego, as well as show the exact criteria for being considered as an entity with self-consciousness.

Apart from the nature of consciousness and self-consciousness, whether these two qualities are necessary and/or sufficient conditions of personhood has also received considerable attention among various philosophers. I will argue for consciousness as necessary condition of personhood and self-consciousness as sufficient condition.

Consciousness

Since ancient times, numerous scholars have given their positions to the problem of consciousness. Although there has not been a definitive conclusion upon the nature of consciousness, at least we have reached some consensus ("Consciousness (Stanford Encyclopedia Of Philosophy)"). Consciousness is generally defined as the capability of sensing and responding to its immediate environment (Armstrong 55-67). Consciousness is clearly a spectrum with infinite degrees and consciousness and self-consciousness are two different concepts (Leibniz 307). Beyond these consensus is the battlefield of three major schools of thoughts: associationist psychology, phenomenology, and scientific psychology. In the following sections, I will argue for a combination of phenomenology and scientific psychology for the nature of consciousness.

Phenomenological view on consciousness

Phenomenological view on consciousness rose in response to associationist psychology, which claims that consciousness appears as mental processes interacted with each other in succession states (Hume 25-28). Phenomenologists like Kant argued that consciousness could not possibly be produced by succession of thoughts; instead, there exists a mental structure with respect to many important qualities that a mind perceives when experiencing the objective world consciously (qtd. in "Consciousness (Stanford Encyclopedia Of Philosophy)").

The first and foremost feature of a phenomenal consciousness is a mental structure for perception. We could find evidence of this claim in our daily experience. When our eyes met a person, we are not perceiving just his/her eyes, then face, then clothes, then decorations, etc. in succession. We are perceiving them through a *priori* (independent of empirical experience) knowledge with various lens, including not only sensory data of shapes, colors, tones, volumes,

and smells, but also cognitive properties such as space, time, and causation. Although the exact framework properties used in a perception process depends on the nature of the perceived object, the fact that there exists such a mental framework for our perception can be extrapolated by our own experiences and introspection (qtd. in “Consciousness (Stanford Encyclopedia Of Philosophy)”). Some "thin" view holder of the phenomenal properties may argue that these properties only contain sensory properties, but nothing beyond that level. I would argue against this perspective, because some framework of consciousness cannot possibly be reflected by pure sensation. Think of a scene of a balloon punctured by a needle and the next scene of the balloon’s explosion. Instead of just perceiving sensory data of the extreme shrinkage of shape and disappearance of colors, with a blasting sound, we could intuitively sense that there is something more than we have perceived just within the level of sensations. The sequence of time as well as the causal relationship between the needle punctured on the balloon and the explosion of the balloon are also perceived through our consciousness. These parts that are not defined by simple sensory properties belong to the so-called cognitive properties, which include *priori* knowledge of body, person, space and time, causation, meaning, reference, and truth (qtd. in Snowdon). They are also an indispensable part of the mental framework of consciousness. Therefore, we could reasonably conclude that the first feature of a phenomenological consciousness is the possession of some forms of *priori* structure of perception of sensory and cognitive properties through which the cognitive system exerts consciousness in our experience. The second essential structure of phenomenal consciousness is the subjectification of the perceived object. Starting from his famous work *Ideen*, as an influential phenomenologist after Kant, Husserl (qtd. in De Boer 121) elaborates the process in which these framework properties emerge from an object. He argues that once the mind started to consciously perceive an object,

that object ceases to be perceived an objective entity. Instead, it becomes a set of perceptional and functional properties, which are subjectively defined by the nature of that cognitive system, under the name of that object itself. This definition clearly develops Nagel's (435-450) argument that consciousness for a being is the ability to tell "what it is like" to be that being. In his example, bats possess consciousness because something is there for bats to be bats by their unique echo-locatory way of perception. However, he does not elaborate on the specific aspects of a being's subjective what-it-is-like experience. Husserl successfully demonstrated the origin of this what-it-is-like experience. He clearly showed that the physiological basis for the subjective "what-it-is-like" state is the objective difference of the consciousness framework across various cognitive systems. So, in conclusion, if the being has of its own framework of subjectification, we could deem that this being possess the second feature of a phenomenological consciousness. Therefore, for a cognitive system to have phenomenological consciousness, it must possess certain established framework of perception with sensory and cognitive properties that are unique to that being's species. One problem with this argument is the way to determine the cutting point for a cognitive system to possess consciousness in its spectrum. We need to consider the category and strength of the sensory and cognitive properties that a being possesses, and then evaluate whether this being should be regarded as a conscious one. However, artificial intelligence's framework might be extremely different from us, because its sensory component is defined by the properties and performances of the camera, microphone, electronic nose, artificial hands, etc., and its cognitive component is defined by the codes pre-programmed into its central system, which may result in a common-sense system totally different from human being. In short, the great divergence between the origin and cognitive system of artificial machine and biological creatures have prevented us from giving them the name of consciousness. However,

the sole fact that they are different from us should not prevent them having consciousness. Though our cognitive system is shaped by numerous genes within us through evolution and natural selection, it is necessary to recognize another path of developing cognitive system - purely human-made artifact. Despite of the differences these cognitive entities have from us, their non-carbon-based physical composition, and their origin apart from mother nature, as long as they satisfy the basic definition of making responses to its world, and are deemed to possess a cognitive system unique to its species, we should grant them phenomenal consciousness. However, one problem is left unsolved in this theory. There exist some properties that we perceive unconsciously into our brains. We all have the experience that we are dreaming in the middle of a class: every word the teacher said is received by our mind, however we could not comprehend them, because our center of consciousness isn't on these words. This example draws a significant different between possessing the ability of consciousness and being in a conscious state. If a being is only occasionally being in a conscious state, while most of the time staying unconscious about its world, or are only conscious of a limited range of stimulus, while most of the external stimulus is ignored, then we could not deem it as a conscious cognitive system. This distinction has not been fully explored by phenomenologists, but was instead solved by a cognitive theory called *global workspace* developed by Baars (70-72) which I will discuss in the following section.

Scientific psychological view on consciousness

Scientific psychology's research on consciousness starts in early twentieth century with the rise of behaviorism. Though behaviorists such as Skinner tries to grasp the whole nature of human mind from truly objective perspective (27-31), due to limitation of advancement of equipment, it pays much attention to the observable features of human mind - explicit behaviors, while

completely ignoring the internal activities happened within the brain (e.g. consciousness). While reaching the 1960s, cognitive psychology has gradually replaced behaviorism because of their attention of inner mental processes (Neisser 5; Gardiner 32-33), thus paying much attention to consciousness.

According to Baars' theory of *global workspace* (70-72), consciousness is a limited resource competed by different mental processes, on the physical basis that the working memory of human cognition is limited and competed by various mental activities. Once a mental process has entered the conscious state, its information will be broadcasted inside our mind, so that other systems of our brain - behavior control, rational thinking, attention, etc. - could assist and alter its process in a positive way. One thing corresponds with our phenomenal model is the claim that only with global network could the conscious perception becomes possible. If the mental activity only stays at the level of primary senses, it could not be deemed as conscious (Dehaene and Naccache 1-37). The global network that involves cooperation among various cognitive systems is where our own feeling of consciousness lies. In phenomenological consciousness, in conclusion, if cognitive properties of the framework are involved in the perception process, this perception could be deemed as a conscious one. Here scientific psychology corroborates this claim because it explains the cognitive mechanism of working memory behind the reason of the necessity of cognitive properties for the perception framework. So, if a cognitive system exercises its ability of synthesizing different cognitive processes through a global cooperative working memory, then this entity could be deemed as performing.

Combining the two perspectives into the criteria for AI to possess consciousness

Thus, combing the phenomenological and scientific psychological theories, we could give AI a standard to possess the quality of consciousness. We could set the criteria that if AI is able to

structure a perception framework of not only primary sensory systems but also higher cognitive processes, moreover synthesizing these different systems into a global working memory that each system will work together to strengthen the processing result of other systems, then we could satisfyingly grant them the quality of consciousness.

Currently, AI has achieved this level of consciousness only in specific fields, such as AlphaGo for the game of *weiqi*, and Siri for human voice recognition. Although most of today's AI could fulfill the framework part to a great extent, the development of global workspace as well as the responsiveness to various stimulus has been extremely limited. Thus, it is fair to conclude that the current level of AI is only able to satisfy a limited degree of consciousness in the spectrum, thus falling below of the cutting point of possessing consciousness.

Self-consciousness

Self-consciousness is one of the most important topics in the philosophy of mind. It is fundamentally defined as a conscious perception of oneself ("Self-Consciousness (Stanford Encyclopedia Of Philosophy)"). Ancient time discussion of self-consciousness mainly focuses the origin of self-consciousness. Some argue that being aware of oneself is roughly the same experience as being aware of external stimulus (Aristotle 430), so that self-consciousness depends on the conscious perception of external stimulus (Cory 26; Owens 707-722). However, Augustine argues that knowledge of mind could be obtained just by going through itself (qtd. in Matthews 130; qtd. in Cary 63), so that self-consciousness solely depends on introspection. The problem with these claims is that they depend solely on intuition or inner perception, so that there is no way to examine or duplicate these states of self-consciousness as well as to transfer and test its criteria across different cognitive systems. As doubts gradually rose upon these perspectives, two major schools of thoughts rise successively in response to these concerns: Hume's "bundle of perceptions" view and Kantian transcendental view. In combination of these two theories and my own perspective we could achieve a currently preferable theory of self-consciousness.

Hume's view on self-consciousness

Hume's philosophical view on this issue is that it totally rejects the existence of self-consciousness as an inner perception. He only considers ourselves as merely a "heap or collection of different perceptions" (Hume 148; Strawson 61), the subjective perception of oneself totally rejected. He and his successors made two main arguments about the nature of self-consciousness.

Introspection is traditionally suggested to be an appropriate way of perception. However, the claim that "because through introspection I am aware of my own mental properties, I am self-conscious" has been repudiated by Shoemaker (114-118) by arguing that the concept of perception is an understanding of the object through senses, whereas the process of introspection does not involve any account of sensory data. Therefore, the premise for the claim that we could introspectively perceive of something is fundamentally untenable. Thus, introspection could not be deemed as a way of approving self-consciousness (Martin "Self–observation." 119; Rosenthal 37-38). This claim confirms with our common sense in the way that introspection could not give us substantial evidence to the existence of anything. We were at best modeling the real situation in our mind, not to say just imagination. There must exist something beyond mere introspection to prove the existence of that thing (e.g. physiological data for mental states).

Second, they rejected the awareness of a bodily self as a form of self-consciousness. Though some philosophers argue that one could perceive one's body as oneself through sensations from one's body perceived by one's consciousness (Brewer 291-309), it is obvious that our self-consciousness implies something more than just bodily awareness (Martin "Bodily awareness: A sense of ownership." 119-140; "Self–observation." 267-289).

Therefore, it is concluded from Hume's perspective that self-consciousness only involves this bundle of perception, because no other things could provide evidence for something other than this.

However, two problems are left unsolved. The first is about the nature of mental system. We have discussed in the part of consciousness that one's mental activities are not just bundle of perceptions. It involves the perception framework of sensory and cognitive properties and the synergy of various cognitive processing units into a working space that coordinates globally.

Therefore, built upon Hume's argument of self-consciousness being aware of the "collection of different perceptions", I further his argument by claiming that self-consciousness is one's awareness of the existence of the framework through which one interacts with the outer world, and the whole global workspace system that dominates the mental process within oneself. Only if one perceived, whether by inference or experience, such framework and workspace system, can one truly be deemed as possessing self-consciousness. Despite the difficulty of observable evidence for such awareness, we can infer from the being's behaviors and reflections. Once one is aware of the existence of the perceptive framework, one could further learn the advantages and limitations of that, thus avoiding bias in perceiving the world and utilizing advantages in the cognition process (for example, humans could utilize their exceptionally fast facial recognition system once they become self-conscious).

The other problem is that of the self-identity. Imagine that self-consciousness only denotes the awareness of my consciousness at this moment, the next moment I cannot associate that consciousness state with this moment's state. At the same time, we all have the experience of being conscious of ourselves by an identity across the flow of time from birth to death. These all suggest that beyond the immediate awareness of one's mental state at any moment, there should be a subjective state of continuum secured by the existence of self-consciousness. This issue is solved by Kantian argument.

Kantian view on self-consciousness

Kant's argument has been heavily built upon Hume's denial of inner perception of the elusive self as an owner of all the experiences. He argues that there does not exist an intuitive understanding of the self as a given object (Kant 446; Brook 71; Ameriks 279). However, he did not completely agree with Hume's argument of "bundle of perceptions." Instead, he argues that there exists a

necessity of self-consciousness for the continuity of conscious experience over time (Kant 246; Keller 66-67). Although we do not possess an owner of our experiences in ourselves, we do need an awareness of the collection of all the experience across time into one concept of "I", and this recognition of continuous experience in the framework of time is at the essence of Kant's view on self-consciousness.

The problem of this view is that it is also hard to determine the degree to which one associates current experience with past ones. For an elder person, it might be very hard to recall what has exactly happened forty years ago, when he/she was still a student. However, we should still deem him/her as self-consciousness because 1) he/she is still consciousness of this subjective continuum in a considerable long framework (he/she could associate most of his/her experience in the recent two years); and 2) he/she could still recall the essence of those experiences happening in childhood time, so that these experience in part still maintains within this continuous self.

Thus, in conclusion, if a being possesses, to a certain degree, this capacity of relating oneself to one's past and future, then one should be granted to possess self-consciousness.

Combining the two perspectives into the criteria for AI to possess self-consciousness

Following the arguments, if a cognitive system could be aware of the perceptive framework and inner workspace system that operates inside itself, as well as the recognition of the continuous subject over considerable scale of time, then we could confidently grant the existence of self-consciousness.

Therefore, we could set the criteria for AI to possess self-consciousness that it is aware of the framework given by its electronic sense organs and human-programmed consciousness system, and aware of the experiential data of itself across time.

Although the second part is rather simple for AI to possess due to the nature of its memory system that stores all information across time into a single place, the first part seems rather hard and elusive. Granted, currently few, if any, AI could achieve this rather strict standard, but in the future, we have that possibility of strong AI. Therefore, this criterion is set for the coming of that day to determine a truly self-conscious AI.

Consciousness and self-consciousness as conditions of personhood

Since consciousness and self-consciousness are both spectrums, we set each up a cutting point beyond which the name "consciousness" or "self-consciousness" is granted to the cognitive system. When discussing these two conditions of personhood, I will look for cognitive systems that are at or beyond this critical point.

For consciousness, I would deem it as necessary but not sufficient condition of personhood. It is necessary because it confirms to our commonsense that those creatures without framework of cognitive properties (such as plants, invertebrates) is excluded while most vertebrates including human beings are included to be conscious. Beyond that, it is not a sufficient condition because otherwise some important qualities such as intelligence and morality that greatly affect the quality and manner of an entity's response to outer stimulus will not be taken into consideration.

For self-consciousness, I would argue that it is not a necessary but rather a sufficient condition of personhood. First, not all human beings are self-conscious, according to my criterion. It requires considerably high level of rationality, objectivity, and thinking ability that only certain educated or reflective people could meet this standard. Setting this as necessary condition will exclude many human beings in our world, especially human babies and children who has not fully developed their brains to understand self-consciousness, thus exempting them from the moral rights and responsibilities that have been essential to the advancement of civilization. However, it should be deemed as a sufficient condition, since this level of thinking required by being self-conscious implies the possession of other important qualities (e.g., intelligence and morality). The thinking process involved in recognizing the inner system of consciousness implies some level of intelligence; the ability of synthesizing experience into the current self indicated the preservation of acquired morality in childhood. With this level of intelligence, morality, although

we still need to look for other necessary conditions, it is of high possibility that this AI will possess most of them and is thus granted personhood. Therefore, self-consciousness should be considered as sufficient condition of personhood.

Thus, for current level of AI, which is mostly weak AI that focuses on specific task, we could say that they might fulfill a few parts of the criteria for consciousness, but there is still a long way to go before they could be granted self-consciousness, because being aware of inner system of consciousness requires thinking outside of the operating system, which is apparently beyond current level of AI. Thus, for these two fundamental qualities of personhood, AI currently fall behind and are around the lower quartile of the spectrum of personhood.

Conclusion

Throughout the essay, I discussed phenomenological and scientific psychological consciousness, and suggested that the combination of those two could best yield what we called consciousness, which is the existence of a perception framework of primary sensory systems and higher cognitive processes, and a global working memory that synthesizes each mental process into a synergistic system in a cognitive system. Further I evaluate Humean and Kantian view and developed the combining idea which could best depict the nature of self-consciousness, *i.e.*, the awareness of the perceptive framework and inner workspace system that operates inside oneself, and the awareness of the subjective continuum into the name of "I".

To draw a conclusion, consciousness is a necessary condition of personhood, whereas self-consciousness is a sufficient condition for personhood. With these two premises, I developed the criteria for AI to possess such qualities and examined its implications of whether AI possesses personhood or not in its current state. Future research may develop more properties of personhood (such as agency, morality, authenticity, etc.), so that the criteria for AI to be a person is more refined and specified. One day human can confidently decide whether to include a newborn AI as a member of human's society. Not until then could we be fully prepared for the arrival of AGI in terms of not only science and technology, but also morality and legislation.

Works Cited

" Consciousness (Stanford Encyclopedia Of Philosophy) ." *Plato.stanford.edu*. N. p., 2017. Web. 6 Oct. 2017.

" Self-Consciousness (Stanford Encyclopedia Of Philosophy) ." *Plato.stanford.edu*. N. p., 2017. Web. 25 Aug. 2017.

Ameriks, Karl. *Kant's Theory of Mind: An Analysis of the Paralogisms of Pure Reason*. Oxford University Press, 2000.

Aristotle. *De anima (On the soul)*. Vol. 23. Penguin UK, 1986.

Armstrong, David M. "What is consciousness." *The nature of mind* (1981): 55-67.

Baars, Bernard J. *A cognitive theory of consciousness*. Cambridge University Press, 1993.

Brewer, Bill. "Bodily awareness and the self." *The body and the self* (1995): 291-309.

Brook, Andrew. *Kant and the Mind*. Cambridge University Press, 1997.

Cary, Phillip. *Augustine's invention of the inner self: the legacy of a Christian Platonist*. Oxford University Press, USA, 2000.

Cory, Therese Scarpelli. *Aquinas on human self-knowledge*. Cambridge University Press, 2014.

De Boer, Th. *The development of Husserl's thought*. Vol. 76. Springer Science & Business Media, 2012.

Dehaene, Stanislas, and Lionel Naccache. "Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework." *Cognition* 79.1 (2001): 1-37.

Goertzel, Ben. *Artificial general intelligence*. Ed. Cassio Pennachin. Vol. 2. New York: Springer, 2007.

Hume, David. *A treatise of human nature*. Courier Corporation, 2003.

Kant, Immanuel. "Critique of Pure Reason (translated and edited by Paul Guyer & Allen W. Wood)." (1998).

Keller, Pierre. *Kant and the Demands of Self-consciousness*. Cambridge University Press, 2001.

Leibniz, Gottfried Wilhelm. "Discourse on metaphysics." *Philosophical papers and letters*. Springer Netherlands, 1989. 303-330.

Locke, John. *An essay concerning human understanding*. Awnsham and John Churchil, at the Black-Swan in Pater-Noster-Row, and Samuel Manship, at the Ship in Cornhill, near the Royal-Exchange, 1700.

Martin, M. G. F. "Self–observation." *European Journal of Philosophy* 5.2 (1997): 119-140.

Martin, Michael G. F. "Bodily awareness: A sense of ownership." *The body and the self* (1995): 267-289.

Matthews, Gareth B. *Thought's ego in Augustine and Descartes*. Cornell University Press, 1992.

Nagel, Thomas. "What is it like to be a bat?." *The philosophical review* 83.4 (1974): 435-450.

Neisser, Ulric. *Cognitive psychology: Classic edition*. Psychology Press, 2014.

Newell, Allen, and Herbert A. Simon. "Computer science as empirical inquiry: Symbols and search." *Communications of the ACM* 19.3 (1976): 113-126.

Owens, Joseph. "The self in Aristotle." *The Review of metaphysics* 41.4 (1988): 707-722.

Rosenthal, David. "Awareness and identification of self." (2011).

Shoemaker, Sydney. "Introspection and the Self." *Midwest Studies in Philosophy* 10.1 (1987): 101-120.

Skinner, Burrhus Frederic. Science and human behavior. Simon and Schuster, 1953.

Snowdon, Paul. "Peter Frederick Strawson." Plato.stanford.edu. N. p., 2009. Web. 25 Aug. 2017.

Strawson, Galen. The evident connexion: Hume on personal identity. Oxford University Press, 2011.