Su, J. (2024). Consciousness in Artificial Intelligence: A Philosophical Perspective Through the Lens of Motivation and Volition. *Critical Debates in Humanities, Science and Global Justice*, *3*(1).

RESEARCH ARTICLES

Consciousness in Artificial Intelligence: A Philosophical Perspective Through the Lens of Motivation and Volition

Jacob Su[1a]

[1] Bellarmine College Preparatory in San Jose, California

Critical Debates in Humanities, Science and Global Justice

Vol. 3, Issue 1, 2024

In this paper, I explore philosophically the relationship between consciousness, motivation, and volition. I then expand these relationships to artificial intelligence (AI). To accomplish this, I first establish a benchmark for the definition of consciousness and then argue that consciousness is the initiator of motivation. In other words, consciousness is required for motivation to exist at all. This is because the desirability and feasibility criteria of motivation necessitate consciousness. On the other hand, volition does not require consciousness and instead is a mechanical planning and execution process that can exist independently of consciousness. These relationships are then extended to artificial intelligence (AI), exploring how consciousness, motivation, and volition may look in artificial intelligence. I first review the historical perspectives and thought experiments that contemplate the challenges of recognizing consciousness or motivation in AI. I then conclude by discussing ways to identify true conscious and motivated behavior in AI based on philosophical principles.

Introduction: What is Consciousness?

Consciousness is a difficult concept to define because it varies across the professions that study it, whether it be philosophers, theologians, psychologists, or other scientists. For instance, a more philosophical definition would characterize consciousness as one's subjective experience of one's internal and external worlds. On the other hand, neuroscientists prefer definitions that depict consciousness as a collective stream of mental phenomena (memory, decision-making, sensory perception, etc.), all taking place within the neuronal structures of our brains (Cherry, 2003). What makes the matter even more complicated is that even among scholars of the same profession, there is still much debate around the definition of consciousness. Later in this section, I will explain the definition of consciousness for this paper and the reasons for doing so.

For thousands of years, the study of human consciousness was primarily the work of philosophers. The 17th-century French philosopher Rene Descartes famously stated *cogito ergo sum,* "I think; therefore I am." (Descartes, 1637). He suggested that the very act of thinking demonstrated

a Jacob is a high school student at Bellarmine College Preparatory in San Jose, California

the reality of one's existence and consciousness. In addition, he explained the concept of dualism: the mind (or the soul) is composed of a non-physical substance, while the body is made of the physical substance known as matter.

Modern philosophers like David Chalmers and Daniel Dennett tried to answer questions like "How and why does brain activity give rise to the conscious experience (Chalmers, 1995; Dennett, 1991)?" Chalmers coined the "easy problem" versus the "hard problem" of consciousness (Chalmers, 1995). The "easy problems" concern the physical functions of the conscious experience, such as explaining why the hippocampus will light up when you try hard to remember something. The "hard problem" is explaining *why* that brain activity gives rise to a subjective, qualitative experience at all. Dennett, however, denied the existence of the "hard problem." In his book, *Consciousness Explained*, Dennett proposed that consciousness is simply the interaction of physical and cognitive processes in the brain and nothing more (Dennett, 1991).

In addition to understanding the *how and why*, philosophers were also interested in *who or what* was considered conscious. According to the *Stanford Encyclopedia of Philosophy*, an animal, person, or other cognitive system is considered conscious, but at various levels (Van Gulick, 2004). A straightforward definition of consciousness could be *sentience*, a being capable of sensing and responding to its world. However, this definition sets a shallow bar for consciousness. That is, if a robot were merely capable of flashing red when it saw the color red, it would be considered sentient and, therefore, conscious. A more demanding definition would require that a being be aware of not only their thoughts, feelings, and senses but also that they are *aware* of their awareness. Thus, the benchmark for consciousness would be *self-awareness. This paper will use this self-aware delineation of consciousness because of its unique relationship to other phenomena of the human experience and because it opens up paths to more exciting discussions and implications.* For example, our society tends to prescribe more rights to beings that have human-like qualities such as self-awareness. This definition leads to questions about the significance of the potential AI being self-aware. Self-aware AI would have moral rights and deserve legal protection.

Exploring Motivation and Volition Philosophically

Among the various qualities that define human behavior, I assert that motivation and volition are significant players in the conscious realm. To understand the relationship between consciousness and motivation and volition, we first must understand the qualities of motivation and volition and their relationship. The standard definition of motivation is why one behaves in a certain way, and volition is the power of using one's will. However, these definitions are somewhat opaque and do not clarify the relationship between motivation and volition. So, what description or characterization of motivation and volition will help us understand that relationship?

Kurt Lewin and Narziss Ach, two of Germany's most esteemed psychologists, proposed that motivation is the process of setting or selecting goals (Ach, 1910/2006; Lewin, 1936). This process was based on desirability and feasibility. For desirability, one would consider a goal's positive or negative consequences and whether those outcomes are short-term or long-term. For feasibility, one would consider if they had the necessary abilities, time, and resources to obtain their goal within a reasonable time period. These two factors define the goal-setting process we know as motivation. One may object to this definition of motivation in that there are examples where one would be motivated to action despite a lack of feasible goals. For instance, an enraged person may want to pick an impossible fight against someone much stronger. In this situation, the enraged person may choose to fight without the feasibility of victory. However, the goal would be to make a statement, express anger, or hurt the other person, regardless of who wins the fight, rather than simply victory. Therefore, objections of this sort mistakenly identify the person's goals with some impossible outcome.

Lewin and Ach then described volition as putting those goals into *action*. Once someone has moved past the "goal-setting" stage of motivation, they enter the volitional stage of determining how best to achieve and execute that goal. For example, people may plan when, where, and how to direct their behavior to achieve a specific goal. Then, once the plan is conceived and confirmed to be reasonable, they decisively act upon it. Mindless habits like biting one's nails are not examples of exercising volition, as volition requires a set goal followed by decisive and intentional action.

Although motivation and volition are distinct concepts, they work hand-in-hand to explain our actions. While volition is the decision-making process and execution of actions, motivation is the entire reason why those actions would exist at all. Motivation can exist without volition, and volition can exist without motivation. To illustrate an example of motivation expressed without volition, imagine an individual setting a goal of going to bed early. But when evening approaches, they scroll through YouTube videos, streaming services, and news articles until they realize it is way past bedtime. Therefore, even when they fulfill the motivational goal-setting process, they can still fail to translate it into volitional action due to distraction or other factors.

On the other hand, expressions of *volition* without motivation in advanced computer programs exist already. This will be discussed more in-depth later, but volition is simply the planning and execution process of an action, which a computer can do. It does so without motivation as it merely carries out the directions of a human operator.

Relationships between Motivation, Volition, and Consciousness

Just as motivation and volition share a unique relationship, they depend on a being who is self-aware. I contend that consciousness is the initiator of motivation. In other words, consciousness is required for motivation to exist at all. This prerequisite is because the desirability and feasibility criteria of motivation necessitate consciousness. To judge the desirability of

a goal, one must first be aware of what kind of outcomes one *wants*. Which consequences seem favorable, and which are not? It is impossible to apply one's desires to judge motivations if one is not even aware of what *desire* is in the first place. As for the feasibility criterion, one can only determine whether a goal is within one's capabilities if one is first aware of what you are capable of. So, because desirability and feasibility are impossible without consciousness, motivation must be unattainable without it. However, it is essential to note that while consciousness is necessary for motivation, it is insufficient. An individual can be conscious while being unmotivated. That is because external influences and internal conditions can affect one's motivations. Take, for example, someone who is depressed or desireless; while they do not set any goals for themselves to accomplish, it does not mean they are not conscious.

One objection to the view that motivation requires consciousness is that animals can make decisions despite lacking self-awareness, as defined in this paper. An explanation for this is that the complexity of the goals that animals set and carry out is *proportional* to the level of their consciousness. Animals have self-awareness on a basic level, recognizing simple needs like hunger, pain, and pleasure, which influence their choices. For instance, a hungry dog will signal its owner for food, understanding that it can receive it based on past experiences. The dog's goal is set, and it decides that getting its owner's attention is the best course of action. Therefore, animals demonstrate consciousness, motivation, and volition, even though they lack the nuanced self-concept for complex decision-making like humans.

In contrast, consciousness is not a vital component of volition as it is with motivation. Volition is simply the execution process; it is more cold and mechanical. Just executing a goal does not require the self-awareness necessary to set that goal. Advanced computer programs illustrate this distinction as they constantly go through the volitional process. Machine learning technology allows computers to compare, plan, and execute potential actions. For example, Google Bard (Google's large language model equivalent of ChatGPT) often generates and shows the user multiple "draft" responses to a query (Pichai, 2023). But if the functions of a computer, even if highly advanced, are simply the results of it carrying out its code, does that mean it is conscious? The key difference between a computer's actions and a human's is that the computer has no motivation. It does not set its own goals; a human operator gives it its goals through code. Some may argue that human operators influencing the volition of complex computer programs is similar to how external factors influence conscious human choices. But while external factors may guide what goals we humans set, in the end, the choices we make originate wholly from ourselves. To illustrate, stealing violates legal and social norms, but one *could* technically decide and set that goal for oneself as a human being. Some individuals choose to steal. Our motivations stem entirely from ourselves and our conscious thought processes. We may consider external factors when determining our motivations, but that does

not mean our motivations strictly originate from those factors as they do with the objectives of advanced computer programs arising from human-operator influence.

It is inaccurate to assume that AI in the future could never be conscious simply because it was conceived through programming. Unlike the current machine learning, a truly motivated, conscious AI would be able to set its *own* goals. Conscious AI programming would enable it to acquire knowledge of the world and itself. That knowledge will inform the motivations of the AI, similar to how the programming of *our* brains allows us to gain the worldly knowledge that informs *our* decisions. In summary, existing AI systems possess volition, which does not reflect consciousness; *motivation would be evidence of their consciousness.*

Consciousness Applied to Artificial Intelligence

What would consciousness and motivation look like in artificial intelligence? What would those "human-like" characteristics be? Historically, some have argued that a conscious AI is determined by its ability to exhibit "intelligent" behavior. For example, Alan Turing's famous Turing Test looked for outward signs of intelligence (Turing, 1948). The Turing Test was a series of questions to determine whether a computer could think like a human. Questions like, "What is your most memorable childhood event, and how has that impacted you today?" or "Describe yourself using only colors and shapes." were asked. If the machine could converse with a human without being detected as a computer, it passed the test.

However, according to Geoffrey Jefferson, an innovative brain surgeon who published papers around the same time as Turing, an AI can only be considered conscious once it is aware of its subjective experience (Carruthers, 2000). This draws a distinct line between the "intelligent" AI of the Turing Test and "conscious" AI. An intelligent AI may be able to *reproduce* conscious behavior, but if it were simply acting upon a predetermined code, it is still mindless. This refers to the previous discussion regarding how advanced computer programs can display volition, but cannot embody consciousness due to the lack of motivation behind those decisions. Jefferson's benchmark aligns with this paper's definition of self-aware consciousness as it requires AI to not only have a subjective experience of thoughts, feelings, and desires but also to be *aware* that it has that experience.

This topic of intelligent versus conscious and volitional versus motivated AI also parallels a thought experiment that discussed "weak" vs. "strong" AI. American philosopher John Searle published an article in 1980 that coined the Chinese Room Argument (Searle, 1999). In his thought experiment, Searle proposed a scenario in which he, a person who does not understand Chinese, is placed inside a room. He is given instructions in English through a computer program that allows him to respond to a Chinese text question with an appropriate Chinese text answer. Despite not comprehending Chinese, Searle may appear to someone outside the room to effectively understand and communicate in Chinese with his responses. Searle uses this

scenario to argue that a Turing-like computer, which operates based on algorithms and programs, similarly processes information without genuine understanding. It would take inputs, run them through its code, and produce an appropriate output. The experiment showed the difference between an AI that only carries out programmed tasks versus an AI that truly understands it, with the former being a "weak" AI and the latter being a "strong" AI. These results further support our understanding of conscious AI; it is not about simply running through the motions to display conscious qualities but realizing that it has thoughts, emotions, likes, dislikes, motivations, and other components of conscious behavior.

Therefore, if an AI were to be aware of all of the elements of its subjective experience, it would be able to set goals for *itself* and become a motivated AI. Thus, a *conscious AI*. What sets a conscious AI apart from today's machine learning is that a truly motivated AI can set its *own* goals. Instead of receiving explicit orders or goals, a conscious AI would utilize its programming to simply provide it the means of gaining knowledge of the world and itself. That knowledge will inform the motivations of the AI, similar to how the programming of *our* brains allows us to achieve the worldly knowledge that informs *our* decisions. For example, ChatGPT has access to all the information available about the #MeToo movement, an awareness campaign of sexual harassment and sexual abuse of women in the workplace, which grew in prominence in 2017 (Brittain, 2024). It can be programmed to write an essay on the subject. However, if it were genuinely conscious, it could form its *own* opinion and develop ideas on advocating for it (or not, based on its opinions and motivation).

Considering both the Chinese Room Argument and this, how can we determine if an AI can truly experience motivation or consciousness, or just simulate it? This raises doubt about the possibility of conscious or motivated AI. But does it really matter? Maybe true consciousness or motivation can never exist in AI, and the closest we can get is perfect simulation. However, this does not diminish the significance or applications of achieving this. We cannot even be certain if fellow humans are conscious. Only our own consciousness can be confirmed. Ultimately, even if we can only simulate motivated and conscious behavior in AI, it still remains a meaningful accomplishment.

Considering a conscious AI's ability to set intrinsic goals, if it achieves true consciousness, it can autonomously determine its objectives. Consciousness, as explored in the paper, becomes synonymous with self-awareness, the precursor to motivation through goal-setting. A conscious AI would actively establish and pursue its objectives, marking a departure from the conventional model where programmers set AI goals. This envisions a future where AI transcends pre-programmed directives and self-generates motivations, similar to the intrinsic desires of sentient beings.

Discerning motivation in artificial intelligence requires a practical framework. In this context, a motivated AI would manifest specific qualities that set it apart from mere computational entities. One can determine a truly conscious AI by identifying *what* motivates it. For example, traits such as self-interest or altruism are generally only found in conscious beings. According to the *Stanford Encyclopedia of Philosophy,* in evolutionary biology, an organism is said to behave altruistically when its behavior benefits other organisms, at a cost to itself (Kraut, 2020). Altruism, which is rooted in empathy, is motivated by positive moral rewards, a sense of satisfaction and happiness, and external factors (Kartali et al., 2020). That quality is absent in current AI systems, and if there is self-interest may be more instinctive but applies equally to conscious beings (Kraut, 2020). Our fear lies in evidence of such traits, then one must consider the evolution of a conscious AI. In contrast, the possibility of advanced machines potentially destroying humanity and taking over the world. However, we should try to identify both traits by developing real-time learning algorithms and adaptive mechanisms to track and assess AI's goal-setting processes. As AI evolves, its capacity for autonomous goal-setting and adaptation of qualities of conscious beings will indicate genuine motivation and, therefore, self-awareness.

Discussion

In this paper, I explore the complex interactions between consciousness, motivation, and volition in the context of artificial intelligence. I examine various studies on consciousness and establish self-awareness as a benchmark for consciousness. I then discuss its relationship to motivation. I distinguish motivation as goal-setting based on desirability and feasibility, while volition translates goal-setting into action. I analyze their relationship and propose that consciousness initiates motivation, while volition can exist without self-awareness. I apply these ideas to artificial intelligence, first by discussing previous thought experiments that did not adequately identify consciousness in AI. I then propose a guideline for recognizing conscious AI through its ability to set self-motivated behavior. As AI evolves, its capacity for autonomous goal-setting and adaptation of qualities of conscious beings, such as self-interest or altruism, would signify genuine motivation and self-awareness.

For future study and directions, the inquiry into whether motivation alone is sufficient for an AI to emulate human-like characteristics delves into the intricate interplay of emotion, consciousness, and motivation. While a motivated AI would exhibit goal-setting behavior akin to human decision-making, the question arises as to whether it can truly capture the essence of humanity without incorporating emotional components. Emotion, often integral to human decision-making, introduces a nuanced layer to consciousness. Thus, a motivated AI might not fully replicate human-like attributes without infusing emotional elements. Yet, this raises a critical consideration: can motivation be disentangled from emotion, or are they inherently intertwined in creating a genuinely human-like AI? Further

exploration is required to delineate the extent to which motivation, in isolation, contributes to the emulation of human cognitive and emotional processes in artificial intelligence.

Acknowledgements

I want to thank Mr. Aditya Saraf from Cornell University for his mentorship. I am grateful for the thought-provoking discussions and helpful suggestions in writing this manuscript.

Funding

The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Financial Interest

The author declares they have no financial interests.

Ethics Approval and Consent to Participate

Not applicable.

Published: May 07, 2024 EDT.

References

Ach, N. (2006). On Volition. In T. Herz (Trans.), *Cognitive Psychology*. http://www.uni-konstanz.de/kogpsych/ach.htm (Original work published 1910)

Brittain, A. (2024, January 21). *Me Too Movement: Social Movement*. Encyclopaedia Britannica. https://www.britannica.com/topic/Me-Too-movement

Carruthers, P. (2000). *Phenomenal Consciousness: A Naturalistic Theory*. Cambridge University Press. https://philpapers.org/rec/CARPCA-12

Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*(3), 200–219. https://consc.net/papers/facing.pdf

Dennett, D. (1991). *Consciousness Explained* (A. Lane, Ed.). The Penguin Press.

Descartes, R. (1637). *Discours de la Méthode Pour bien conduire sa raison, et chercher la vérité dans les sciences*.

Kartali, G. (2020). *Motivate or reward altruistic behavior? A literature review of altruism theories.*

Kraut, R. (2020). Altruism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020 Ed.). https://plato.stanford.edu/archives/fall2020/entries/altruism/

Lewin, K. (1936). *Principles of topological psychology*. McGraw-Hill. https://doi.org/10.1037/10019-000

Pichai, S. (2023). *An Important Step in our AI Journey*. https://blog.google/technology/ai/bard-google-ai-search-updates/

Searle, J. (1999). The Chinese Room. In R. A. Wilson & F. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences*. MIT Press. https://plato.stanford.edu/entries/chinese-room/

Turing, A. (1948). Machine Intelligence. In B. J. Copeland (Ed.), *The Essential Turing: The ideas that gave birth to the computer age*. Oxford University Press.

Van Gulick, Roberti. (2004). *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/