Živan Lazović
Faculty of Philosophy
at the University of Belgrade
zlazovic@f.bg.ac.rs

Mirjana Sokić
The Institute for Philosophy
at the University of Belgrade
mirjanasokic19@gmail.com

Original Scientific Paper
UDC 165.1
004.81:165.1

ARTIFICIAL THINKERS AND COGNITIVE ARCHITECTURE

Abstract: *This paper aims to propose and justify a framework for understanding the concept of personhood in both biological and artificial entities. The framework is based on a set of requirements that make up a suitable cognitive architecture for an entity to be considered a person, including the ability to have propositionally structured intentional states, having a form of sensory capabilities, and having a means of interacting with the environment. The case of individuals in a persistent vegetative state, as studied by Owen, serves as an example to show the importance of each of these requirements and the possibility of a "hybridization" of personhood. The proposed set of requirements provide a complete framework for understanding the concept of personhood and highlight the significance of cognitive architecture in determining personhood.*

Keywords: *Personhood, cognitive architecture, artificial intelligence, artificial thinkers, hardware, program.*

1. Introduction

Isaac Asimov's seminal work, *Bicentennial Man* (1990), presents a thought-provoking exploration of the concept of artificial personhood through the characterization of its protagonist, Andrew, an advanced robot engineered with cutting-edge technologies and designed to resemble and emulate human behavior. The novel's narrative progression, which is marked by Andrew's interactions with human beings and his subsequent questioning of his own identity, raises important queries about the definition and determining factors of personhood. Asimov's novel posits that the concept of personhood may not be reducible to physical characteristics or capabilities alone, but rather encompasses a complex interplay of consciousness, emotions, and intellect.

 DOI: 10.5937/BPA2336049L

The novel's treatment of this question leaves Andrew's personhood open to interpretation and invites readers to reflect on their own perceptions of what it means to be a person.[1] Central to the narrative is the question of whether it is the hardware or advanced software that imbues Andrew with his sense of self and humanity. Additionally, the novel raises ethical implications of this question and its potential impact on the understanding of personhood in the context of artificial intelligence. This study aims to proffer a response to the inquiry pertaining to the attributes that endow some artificial entities or systems with the capability of being deemed as artificial persons. Furthermore, it serves as the underlying foundation for a comprehensive examination of the issues pertaining to synchronic and diachronic personal identity, as well as various other relevant philosophical concerns.[2] Yet, prior to engaging in these contemplations, it is imperative to furnish a more exhaustive explanation of the philosophical framework in which this discourse is situated.

As technology advances at a rapid pace, the possibility of creating authentic thoughts and consciousness in artificial systems becomes increasingly plausible. This has led to significant attention being paid to the field of artificial intelligence within both scientific and philosophical communities, with much of the discourse centered on determining whether programming a computer in a specific way can result in the production of authentic, conscious thought. However, as Eric Olson (2019: 69) points out, these debates often overlook the fundamental aspect that a thought can only exist in the presence of a "thinker" – i.e. an entity that serves as the embodiment or manifestation of that thought. This raises the question of the nature of the artificial thinker, and prompts the inquiry into what the *subject* of these artificial thoughts would be. The most common answers to this question are that (a) the *computer* itself would be the intelligent subject or (b) that it would be the *program* running on the computer.

These answers are often accepted without further argumentation or critical evaluation, presenting a significant challenge in the field of artificial intelligence and calling for further exploration and analysis of the nature of the artificial thinker. The philosophical problem that directly arises from this uncritical response to the posed question is highlighted by Olson in the following passage:

1 It is important to note that the physical resemblance of Andrew to a human is not a crucial determinant in assessing his personhood. The purpose of this research is to establish the necessary and sufficient conditions that any biological or artificial entity or system must fulfill in order to be considered a person.

2 The distinction between diachronic and synchronic personal identity can be described as follows: diachronic personal identity refers to the continuity of an individual's identity across different stages of their life, including their memories, personality, and physical characteristics (see Maslin 2001: 242; Noonan 2019: 14). In contrast, synchronic identity refers to the characteristics and attributes that an individual possesses at a particular point in time. It can be conceptualized as a snapshot of the individual's identity at that specific moment, including all the elements that make up their identity at that time.

> For there to be thought or consciousness is for there to be *something* that thinks or is conscious – just as for there to be life is for there to be living things, and for there to be movement is for something to move. For there to be artificial intelligence, then, there must be *an artificially intelligent being: a thing that is intelligent* because of what a computer does. So there are two different questions concerning the possibility of artificial intelligence. One is whether anything in the nature of thought itself prevents it from occurring in computers. We might call this the *question of artificial thought*. The other is whether anything could be an artificial thinker. We might call this the *question of artificial thinkers*. The second question has to do with the *sort of entity an artificial thinker would be. What properties would it have, in addition to its mental properties? Would it be a material thing?* If so, what matter would make it up? If not, what sort of immaterial thing could it be? What might it be made of if not matter? (2019: 68, emphasis added)

It is important to note that Olson does not provide a comprehensive justification for his assertion that every thought necessitates a bearer, instead treating it as an axiom in his examination of the issue of artificial intelligence (2019: 69–70). In this paper, we accept the thesis that thought, whether in the context of biological (natural) or artificial systems, must have a carrier. In other words, we maintain that the presence of a thinking entity is necessary for the existence of thought. This claim is rooted in the understanding that thought emerges as a property of complex systems and therefore requires a substrate or carrier to exist and manifest. This thesis holds substantial implications in both cognitive science and philosophy, and has significant implications for our understanding of the nature of thought and consciousness in both biological and artificial systems. Additionally, we acknowledge that Olson's analysis highlights important challenges with the concept of artificial intelligence.[3] The most significant of these problems can be summarized as follows:

1. The term "artificial intelligence" typically refers to the possibility of creating streams of conscious thoughts, but these thoughts cannot exist without an artificial thinker.
2. The ontological question of artificial persons (or thinkers) is largely neglected in scientific and philosophical discussions.

---

3 Olson's use of the term "intelligence" is somewhat idiosyncratic. In his understanding, this term refers to mental phenomena such as beliefs, desires, emotions, consciousness, etc. – that is, thoughts and consciousness in general. He also notes that the term "artificial intelligence" in its common sense is often used to refer to forms of intelligent *behavior* in computers and machines (e.g. sorting different types of shapes or materials, playing chess, automated cars, etc.). In contrast to this common use, Olson refers to this term as the conceivable possibility of producing *authentic thoughts* and consciousness in artificial systems such as computers (see Olson 2019: 67).

3. Instead of a precisely specified concept of an artificial thinker, discussions about artificial intelligence often use vague terms such as "system", "substrate", or "medium", the reference of which is unclear. (Olson 2019: 69)

We concur with Olson's assessment that the field of artificial intelligence is beset by misunderstandings and ambiguities in terminology. In this research, we propose a framework for determining personhood in entities, based on the fulfillment of certain conditions such as propositional intentional states, sensory apparatus, and an apparatus for interaction with the environment. This framework, which we term "the architectural view", provides a comprehensive and unified understanding of the concept for both artificial and biological entities.

The structure of the remaining course of this paper is as follows: Initially, we will undertake a more comprehensive examination of the perspectives known as the "hardware view" and the "program view", as well as the criticisms they face. Subsequently, we will conduct a critical analysis of Olson's conceptualization of the notion of artificial persons. This critical analysis will demonstrate the need for a hybrid understanding of the notion of persons, both in biological and artificial contexts, one that circumvents the drawbacks of both aforementioned perspectives while retaining their advantages and drawing upon the concept of cognitive architecture. In the final section, we will evaluate the cogency of our hybrid proposal. Our aim is to conclude that our focus on cognitive architecture allows for a more coherent examination of the implications of personhood in various contexts and sheds light on the nature of both biological and artificial thinkers.

2. The hardware view

As advancements in technology pave the way for the possibility of simulating streams of conscious thoughts within computers, a fundamental question arises regarding the agency responsible for such thought processes. According to Olson (2019: 70), it is commonly assumed that the computer, as the physical embodiment of the system, would be the entity engaged in the act of thinking (see e.g. Turing 1950; Putnam 1964; Searle 1980: 417; Haugeland 1985; Russell & Norvig 2010).

Despite its prevalence, Olson notes that this answer is rarely supported by further arguments and is often accepted uncritically in debates about artificial intelligence (Olson 2019: 70). In addition, discussions of the nature of artificial intelligence often do not specify the type of objects that these thinking computers are: it is simply assumed that the intelligent entity is a *physical object* made of metal, plastic, and silicon chips. Olson refers to this assumption as "the hardware view" and asserts that it is the best answer to

the question of the nature of artificial thinkers. However, it is important to note that this view is not without its significant criticisms. The main obstacle to arriving at a satisfactory answer to the question of the nature of artificial thinkers in the spirit of the hardware view is the presence of the following two assumptions, which align closely with our common-sense intuitions and appear in discussions of the nature and conditions of the diachronic identity of biological persons.

One assumption is that programming a computer for intelligence does not simply *bestow* intelligence upon a previously non-intelligent being, but rather *creates* an intelligent being. This would imply that installing and uninstalling a program results in the creation and destruction of an intelligent thinker. However, it is clear that such actions do not affect the physical hardware of the computer. This leads to the conclusion that artificial thinkers and computers, which are identical in terms of hardware, have distinct histories or conditions of persistence. For example, the computer hardware would exist before and after the existence of the intelligent thinker (Olson 2019: 71).[4] In other words, the hardware view is confronted with the problem of diachronic identity in relation to artificial persons.

In the context of biological individuals, Olson presents a solution to the problem of diachronic identity that is based on the preservation of numerical identity. Specifically, he argues that an individual, *A*, at time t_1, is numerically identical to another individual, *B*, at a future time t_2, if and only if *B* at t_2 possesses the same biological organism as *A* at t_1. To elaborate, the individual who defeated Persian King Darius III at the Battle of Guagamela is numerically identical to the individual whose teacher was Aristotle years prior to that event, as the individual is the *same biological organism* belonging to the species Homo sapiens. Olson posits that psychological characteristics do not play a role in determining diachronic identity; for instance, an individual in a coma, lacking all psychological activity and content, is still considered the same person as the individual who was once a renowned Formula 1 driver, as they represent the *same living organism*. According to Olson, this illustrates that the preservation of numerical identity is dependent solely on the biological organism and not on any psychological characteristics or traits (see Olson 2000: 16–18).

It is important to note, however, that this solution is not applicable in the context of artificial persons. In the context of biological individuals belonging to the species Homo sapiens, a fetus and newborn, while lacking psychological characteristics that would classify them as individuals, still represent living organisms that will develop these characteristics through natural development. Thus, the emergence of psychological characteristics

4 It also raises the well-known problem of "too-many-thinkers" in the context of artificial intelligence. For further information regarding this problem, see Snowdon 1990; Olson 2000; Sutton 2014.

that define an individual is an inherent aspect of biological organisms belonging to the species Homo sapiens. In contrast, a collection of processors and silicon chips can exist and function without ever becoming an artificial thinker; this capability is only achieved through the pairing of appropriate software with the hardware. This shows that the reduction of artificial persons solely to hardware does not offer a convincing solution to the problem of diachronic personal identity. In simpler terms, Olson posits that the hardware view fails to furnish a convincing resolution to the issue of the diachronic identity of artificial thinkers or persons, in a manner similar to how animalism addresses the issue in the case of biological persons.

Another assumption that is often made in the context of artificial intelligence is that an intelligent entity can be transferred from one piece of hardware to another through the transfer of *data*, such as a *program file*. This would imply that the first piece of hardware would lose all of its mental properties, including memories, beliefs, preferences, and cognitive abilities, while the second piece of hardware would acquire them after the program is installed. However, this assumption is incompatible with the hardware view, for it implies that an intelligent entity possesses a property that hardware does not have, namely the ability to be transferred through data transfer; i.e. allowing for such a transfer would entail that the artificial thinker or person is identified with the program, rather than the hardware (Olson 2019: 71).

One potential solution proposed by Olson for addressing the issues associated with the hardware view is to adopt the perspective that programming a computer does not *create* an intelligent entity, but rather, it *imbues* a previously unintelligent entity with intelligence. Analogously, the deletion of data or software does not eliminate an intelligent entity, but rather renders it non-intelligent. Additionally, proponents of the hardware view may reject the idea that an artificial thinker can be transferred from one piece of computer hardware to another, despite the transfer preserving the psychological continuity of the artificial thinker. In other words, they may argue that when the hard disk of one computer (C_1) is transferred to another computer (C_2), it should be seen as C_2 receiving a new hard disk with new programs, rather than C_2 becoming C_1. This avoids the challenges associated with the requirement for psychological continuity for the diachronic identity of artificial persons. This strategy would allow the hardware view to explain the nature of artificial persons using the same (anti-psychological) model as animalism does for biological persons. However, this solution encounter counterintuitive conclusions: if we transfer a hard disk containing a program that creates a conscious artificial person, "Eve", from computer C_1 to computer C_2, it appears evident that Eve will be transferred from C_1 to C_2, which is not analogous to simply switching cables or components from C_1 to C_2. As such, it seems that not all organs or computer components are equally crucial for the realization of persons, and this will be further examined in Section 4 of the paper.

3. The program view

In this section, we will consider the so-called "program view", according to which artificial thinkers are not *physical entities*, but rather computer programs running on a computer; i.e. the concept of an artificial thinker or person is perceived as an *intangible* or *virtual* construct rather than a tangible entity.[5] The program view allows for the overcoming of two common-sense assumptions that have presented challenges for the alternative hardware view.

One of these assumptions is the thesis that the initiation and termination of a program results in the creation and destruction of an artificial person. This thesis can be further refined by considering the AI program as enabling the computer to attain conscious mental states. In this context, the initial installation of the program on a computer can be understood as the "birth" of an artificial person, while turning off and on the computer represents putting the person to sleep and waking them up; modifications to the program can be seen as modifications to the person's psychological content and abilities, and uninstalling the program or resetting the computer can be seen as the person's "death".

Another advantage of the program view is that it accounts for the transfer of an artificial person from one piece of hardware to another in a straightforward manner. In other words, it follows from this view that by transferring the data (program file) from one computer to another, the first computer loses all of its mental properties and the second computer acquires them. This implies that the artificial person possesses the property of being able to be transferred from one piece of hardware to another, which is not possible in the hardware view. One limitation of the program view of artificial intelligence is that it suggests that artificial persons may be incapable of intrinsic change. This limitation arises from the fact that the type of program that enables artificial intelligence does not change when the computer is turned on or when a sentence is typed and saved. As Olson points out:

> [U]niversals don't change. ... any more than the colour white changes when I spill coffee on a piece of paper. ... At most a particular concrete instance of the program can change. But conscious, thinking beings must be able to change intrinsically: in their beliefs, preferences, and perceptual states. (2019: 74–75)

In short, if artificial persons are identified with a certain *type* of program, they would not be able to experience changes in their beliefs, preferences, and perceptual states. However, the conclusion that the program view is ultimately

5 Olson correctly observes that, from this perspective, an artificial person would be viewed as a set of instructions, and as such, it would be brought into existence at the time of its initial conception or notation, rather than upon initiation of execution on a computer (see Olson 2019: 74)

flawed may be premature, as it is not immediately evident why proponents of this view could not claim that each artificial person is a particular concrete *instance* of the program type, rather than being identical to the program type itself. This alternative formulation of the view aligns more closely with the intuitive understanding of biological persons and allows for the possibility of changes to individual instances of the program.[6]

Despite this, we will conclude this section by stating that Olson effectively brings attention to the significant drawbacks of the program view, and that such limitations serve as a counterargument to the acceptance of a viewpoint in which the notion of a person within a biological context is fully identified with or even reduced to a set of psychological characteristics. In the following section, we will thoroughly examine Olson's ideas on artificial persons, which will enable us to articulate our own position.

4. Objections to Olson's analysis

According to Olson's view on personal identity, which is generally known as "animalism", a person is identical to a specific biological organism (Olson 2000: 16). As previously discussed, the hardware view adopts a similar approach in attempting to explain the concept of artificial persons, suggesting that an artificial person is merely a specific material object. However, Olson argues that there are fundamental distinctions in the definitions of biological and artificial persons that pose significant challenges for the hardware view. We take it that Olson's argument contains several shortcomings which we will further examine in this section.

First, the intuitive appeal and acceptance of Olson's position on the nature of biological persons can be attributed to the fact that, due to the current technical inability to transfer brains from one body to another, the concept of self has been indelibly linked to the physical body that realizes it. Even in the event that brain transfer becomes feasible, it is likely that many individuals would reject the notion that person *S*, previously embodied in body B_1, would remain the same individual post-transplantation into body B_2. It is reasonable to assume that many acquaintances of person *S* would likely not accept that they are interacting with the same person in a new body, instead positing that *S* continues to reside within the original body B_1, despite the fact that it no longer possesses the first-person perspective or psychological attributes characteristic of person *S* as they knew it. These intuitions make it challenging to accept the notion that psychological characteristics constitute an integral aspect of one's identity. It is important to note, however, that human intuitions do not necessarily serve as a reliable indicator of truth.

6 For instance, it can be argued that every human individual is a particular concrete instance of the type of biological organisms within the genus Homo sapiens.

The history of philosophy and scientific inquiry is replete with instances in which intuitive claims have been proven false, as well as those in which counterintuitive assertions have been vindicated.

The conventional and commonly held understanding of the concept of personhood extends to artificial entities as well. However, at present, the state of computer programs and artificial intelligence development does not afford us the capability to establish significant communication and interpersonal relationships with a distinct artificial person. Transferring digital files from one drive to another is not perceived as being fundamentally different from simply moving digital material from one location to another. However, as technological advancements progress, it will become possible to "transfer" an artificial person from one medium to another. The digital structure or program with which we establish communication will assume a far more significant role than the hardware components have thus far. These components will be viewed solely as the physical embodiment of an intelligent entity, rather than being identified as the entity itself. This may also lead to new intuitions regarding biological persons, who will no longer be seen as being inextricably linked to a single physical embodiment, as has been the case thus far. This makes the work of philosophers and cognitive scientists on determining the concept of artificial persons even more vital, as it can help us transcend some of the preconceptions that have been formed through familiar thought experiments based on current examples of brain transplantation or memory transfer.

A further component of Olson's argumentation with which we do not concur pertains to the assertion that there exists a clear distinction between the way in which we can explain the nature of biological persons on one hand, and artificial persons on the other. Olson articulates this thesis in the following manner:

> *An organism's life* is roughly the sum of its physiological, immune, and metabolic activities. *My* hands are parts of me because they are caught up in *my* life: they and all their parts are nourished by my bloodstream and participate in *my* metabolic processes. *My* gloves are not parts of me because they are not caught up in *my* life. There are many hard questions about what counts as *a life*, but this is at least a start. Obviously nothing like this could apply to artificial thinkers. What would be the corresponding principle for them? If my computer's central processing unit could be a part of an artificial thinker but its keyboard could not, why should this be? ... The keyboard does not seem to be involved in the computer's thought at all. And although its power supply is involved--the computer could not produce thought without it – its involvement seems only indirect, compared to certain parts of the computer's digital circuitry. This suggests that an artificial thinker would be composed entirely of electronic components and the wires connecting them. It

> would be a thin, spidery thing made of metal and silicon weighing only an ounce or two. (Olson 2019: 79, emphasis added)

It is our understanding that this passage contains several highly problematic points that are deserving of careful consideration. Primarily, Olson makes a transition from a discourse on an unspecified organism, referred to through the use of the appropriate indefinite pronoun "a", to a discourse on what constitutes an individualized life from a first-person perspective, utilizing the personal pronoun "my". It is well-established that certain physiological, immune, and metabolic processes are necessary for the maintenance of life in biological entities. This is accurately reflected in Olson's assertion that these processes serve as the foundation for the continuation of life in biological systems.

However, it is important to note that these activities are not sufficient for the life in question to be developed enough to allow for a first-person perspective, as Olson later on asserts. Specifically, immediately following the enumeration of these basic systems necessary for maintaining a biological life, Olson, without any additional justification, speaks in terms of "my" metabolic processes. While the physiological activities previously mentioned may be adequate for enabling and maintaining the life of "a" biological organism, they cannot, without the presence of additional and highly complex activities within the brain enabling a first-person perspective, enable "my" life. In summary, the necessary conditions for maintaining "a" biological life are indeed necessary, yet not sufficient for the maintenance of "my" life. It is clear that acknowledging this reality is not consonant with the radical biological viewpoint that Olson has espoused for over three decades; nonetheless, disregarding this fact and Olson's shift from the indefinite pronoun "a" to the personal pronoun "my" is entirely unjustified and lacking in explanatory rationale.

Furthermore, Olson's argument that the basic components of a computer, such as wires and simple electrical components, would be insufficient to support the operation of an artificial thinker or person, is problematic. This is because it overlooks the fact that, similar to biological entities, the functioning of these components alone does not warrant the attribution of characteristics such as a first-person perspective or intelligent thought to an artificial system. This argument is problematic because it does not consider the more complex systems and operations that are necessary for an artificial entity to be considered a thinker or a person. Such complex systems include those that allow for cognitive ability, decision making and self-awareness. These systems would be composed of not just the basic components such as wires and electrical components but many other sophisticated elements which makes it more complex to attribute the characteristics of an artificial person. In the case of biological entities, the fact that they possess basic metabolic activities does not alone suffice to attribute to them the property of a biological person. Similarly, the mere functioning of the basic components

of an artificial system does not warrant the attribution of a first-person perspective, intelligent thought, or any other characteristics that would make it an artificial thinker. Therefore, this objection, which appears to be the main objection that Olson presents to the hardware view on the nature of artificial persons, is not supported by valid reasoning and lacks foundation.

Despite these objections, we agree with Olson's assertion that for an artificial entity to be considered a person or thinker, it must have at least some sort of *material constitution*, similar to that of biological entities. This assertion goes against the prevailing psychological understanding of personhood, the implication of which is that a person can be viewed as a purely abstract entity or functional structure that is capable of being "transferred" and manifested in different physical forms, and may potentially exist without any physical manifestation (Olson 2019: 74).[7] We contend that this prevailing understanding of the necessary and sufficient conditions of personhood in both biological and artificial entities is inadequate. To address this, we put forth an alternative account of personhood, referred to as the "architectural view", according to which an entity must fulfill certain objective criteria of material constitution to be regarded as a person. These criteria will be discussed in detail in the following section of this paper.

5. The architectural view

According to the architectural view, the notion of personhood is closely linked to the presence of a specific physical structure, commonly referred to as *cognitive architecture*. The concept of cognitive architecture is rooted in the works of cognitive scientists and philosophers, such as Pylyshyn, as explored in publications such as Lepore & Pylyshyn (1999). In its simplest form, cognitive architecture can be understood as the *appropriate structure* that enables the emergence of intelligent behavior (Milojević, 2018: 183), or, alternatively, as the *underlying infrastructure* for an intelligent system (Langley et al., 2009: 141).

More specifically, in his book *Macrocognition* (2013), Bryce Huebner explains that cognitive architecture

> consists of relatively independent subsystems, which each process a narrow range of information, and which can be coordinated and

7 We wish to clarify that, although we have adopted and defended the functionalist perspective on the nature of mentality in previous works (see Lazović 2009 and Sokić 2020), we reject the implication that is often associated with functionalism in the context of personhood. This implication holds that the essential property of a person is seen in the abstract functional structure from which propositionally structured intentional states (as well as other mental states and psychological contents) can be attributed. Our understanding, as presented in this paper, is that for an entity to be considered a person, it must meet the minimal condition of *embodiment*, as well as several specific conditions that we explore in the following section.

> interfaced to facilitate skillful coping with environmental contingencies that are significant to the collectivity as such. (Huebner 2013: 199)

To illustrate this concept on a common example, cognitive architecture refers to a set of distinct elements or subsystems that are interconnected in a manner that enables, for instance, unhindered movement across various types of terrain and the ability to circumvent physical obstacles to reach point B from point A. Along with the elements essential for physical movement – e.g. wheels, tracks, legs, paws, etc. – the architecture also comprises a sensory apparatus and a processor that processes information obtained from the environment to adapt its behavior and overcome obstacles.

The aforementioned type of cognitive architecture pertains to the most basic forms of intelligent (i.e. adaptive) behavior. However, for an entity to be considered a person in the full sense of the term, it is clear that additional elements are necessary. Therefore, the question arises as to what constitutes an adequate cognitive architecture for an entity, whether biological or not, to be considered a person. We propose that an adequate cognitive architecture for an entity (whether biological or artificial) to be considered a person can be represented by the following set of conditions:

a) The entity must have propositionally structured intentional states, such as beliefs, desires, hopes, and intentions.
b) The entity must possess at least some form of sensory apparatus.
c) The entity must possess at least some means of interaction with the environment.

These three conditions frame our understanding of personhood. We should clarify, however, that we do not take them to be exhaustive or definitive, and that some philosophers maintain that at least some of these conditions are not even necessary. Specifically, according to Olson's animalistic view, the concept of personhood does not necessitate the presence of any of these conditions.[8] Yet, in contrast to Olson's view, we argue that condition (a), for instance, pertains to the capability of having intentional states, and that its inclusion in our proposal is self-evident, as an entity devoid of such capabilities cannot be considered a person. In other words, we think it is evident that the necessary condition for an entity to be considered a person, both in a biological and artificial context, is the ability to possess properly structured intentional states; namely, an entity cannot be considered a person if it is unable to possess beliefs, doubts, or the ability to question its own existence as a person.[9]

8 This conclusion is a direct implication of Olson's position that a person is numerically identical to her biological organism, even in instances where the individual is in a persistent vegetative state (see Olson 2000: 7–10).

9 Whether a person should possess additional mental states – e.g. sensations, affections, etc. – is a separate and complex issue that falls outside the scope of this paper.

An in-depth examination of the rationale for the implementation of conditions (b) and (c) will be presented with regards to individuals in a persistent vegetative state. For the purpose of clarity, it is important to note that the persistent vegetative state is defined as a condition of extremely severe brain damage in which the patient exhibits no observable behavior, despite appearing to be awake (Cranford & Smith 1979: 203). Now, studies conducted by neuroscientist Adrian Owen (2006) have revealed a method of successful communication with these patients, which suggests that these individuals still meet the criteria outlined in conditions (a), (b), and (c) which are necessary for an entity to be considered a person. The case of Owen's patients exemplifies the importance of each of the conditions in the cognitive architecture, and also illustrates the possibility of a "hybridization" of personhood, as their cognitive architecture includes elements that are not necessarily a part of their physical organism. Furthermore, the case of Owen's patients suggests that the determination of an adequate cognitive architecture should not be restricted to the biological organism alone.[10]

The question of whether an entity that satisfies only one or two of the conditions mentioned above can be considered a person is a complex and nuanced one. Specifically, can an entity that possesses an appropriate network of propositionally structured intentional states, but lacks both sensory apparatus for contacting the environment and means of interaction with it, be considered a person? In addressing this question, it is important to consider the practical implications of the concept of a person, such as legal, ethical, and epistemic considerations. These implications require at least some form of contact or interaction with the environment. Therefore, it can be argued that an entity without any means of contacting or interacting with the external world cannot be considered a person.[11]

However, it should be noted that there are certain borderline cases, such as the example of individuals in a persistent vegetative state, as presented by Owen (2006), where the status of a person may be dependent on the possibility

---

10 It is important to note that this example is not intended to provide a definitive answer to the question of whether individuals in a permanent vegetative state should be considered as persons. Rather, it serves as an illustration of the complexity and nuances of the concept of personhood and the importance of considering different perspectives and evidence when examining this question. Furthermore, it is important to keep in mind that this is a complex and ongoing debate in the field, and more research and studies are needed to fully understand the implications of cognitive architecture and personhood.

11 This thesis unambiguously derives from the classical understanding of the concept of person, which traces its origin to Locke (1975: 2.27.26), and according to which the term "person" is primarily a "forensic term", i.e. a term that has ethical and legal significance. Specifically, we argue that an entity which satisfies condition (a) alone but not conditions (b) and (c) may still belong to the class of *moral patients* (i.e. the class of entities towards which actions may be evaluated in moral terms), although it certainly does not belong to the class of *moral agents* (i.e. the class of entities whose actions may be evaluated in moral terms). For further information on this distinction, see McPherson (1984).

of establishing some form of rudimentary interaction with them. This suggests that the concept of a person is not a rigid one and may depend on the specific circumstances and context. Furthermore, it can be argued that the conditions proposed by our architectural view also determine the temporal boundaries of a person. Specifically, an understanding of the cognitive architecture of a person allows for the specification of the conditions that determine when a person begins and ceases to exist.

With this in mind, let us now examine how the most common examples of borderline cases of personhood would be characterized according to our proposed framework:

- *Is a fetus a person?* The fetus does not meet the requirements for personhood as outlined in conditions (a)-(c). Although it has a rudimentary sensory apparatus, it does not possess the necessary mental states or propositional intentional structures to be considered a person.[12]
- *Is a newborn a person?* A newborn may have some limited abilities to interact with the environment, but it does not possess the necessary mental states or propositional intentional structures to be considered a person.
- *Is a three-year-old child a person?* A child of this age possesses the necessary cognitive architecture to be considered a person, as outlined in conditions (a)-(c).
- *Are primates (e.g. chimpanzees) persons?* Some research suggests that chimpanzees and gorillas possess the cognitive architecture necessary for personhood as outlined in conditions (a)-(c).[13]
- *Is the chatbot "Sophia" a person?* Sophia may have advanced conversational abilities and sensory apparatus, but it does not possess the necessary mental states or propositional intentional structures to be considered a person.[14]

---

12 This point contradicts the position on fetuses put forth by Olson (1997: 96).

13 For more information on the philosophical argument that at least some animals are persons, see Aaltola (2009), and Rowlands (2019). It is important to emphasize, however, that our conclusion is a matter of philosophical debate. Specifically, some philosophers challenge the capacity of non-language-using animals to form propositional attitudes, which would render them non-compliant with condition (a), and therefore not qualified as persons. For further information on these discussions, see Davidson 1982; Dreckmann 1999; Fellows 2000.

14 It is acknowledged that some advanced robots currently possess the capability to experience tactile and auditory sensations. For the purposes of this argument, it is assumed that the chatbot "Sophia" possesses such capabilities as well. However, it is noteworthy that the communication abilities of Sophia, while highly developed, are based solely on the functioning model of a pure algorithm, which is not sufficient in and of itself to meet the requirements outlined in the condition (a) for personhood. The possibility of meeting this condition through the use of a sufficiently developed algorithm is not explored in this discussion. The reality is that at present, Sophia does not meet this

- *Is Andrew, the protagonist from Isaac Asimov's science fiction novel, The Bicentennial Man, a person?* Andrew certainly possesses the cognitive architecture necessary for personhood as outlined in conditions (a)-(c), and thus, according to our proposal, he would be considered a person.
- *Is a human being in a permanent vegetative state a person?* According to the results of Oven's experiments, these patients possess the cognitive architecture necessary for personhood as outlined in conditions (a)-(c).
- *Is a human being who is in a coma or deceased a person?* A person in a coma or deceased does not meet any of the conditions for personhood as outlined in conditions (a)-(c).[15]

As we can see, our three conditions effectively address all the borderline cases by avoiding the drawbacks of the program and hardware views while, at the same time, retaining their positive aspects. Additionally, these conditions classify these cases in a way that aligns with our common-sense intuitions. On the one hand, our architectural perspective avoids the radical anti-psychological stance of Olson's animalism and his interpretation of the hardware view, according to which psychological characteristics of a biological or artificial entity do not constitute a necessary condition for determining personhood. On the other hand, the psychological/program view is frequently criticized for its pronounced virtuality, which renders the concept of personhood *ephemeral* and fundamentally *immaterial*. By emphasizing the necessity of material or physical realization and the conditions under which an entity must possess the ability to interact with its environment, our architectural concept effectively addresses this limitation.

The rationale for adopting this position is that it allows for a separation of function and implementation, with the program remaining at the level of functional description and the hardware remaining at the level of implementation. In other words, we take it that by transitioning to the level of cognitive architecture, the main theoretical challenges for the hardware and the program view can be avoided. This is achieved by considering cognitive

---

condition in practice and, as such, does not possess the cognitive architecture necessary for it to be considered a person.

15 This point directly contradicts the thesis advanced by David Mackie (1999). In agreement with our assertion, Olson presents a well-developed argument to demonstrate that an organism does not persist after death. What is commonly referred to as a "dead body" is simply a remnant of the organism that cannot be identical to any organism that was once alive (Olson 2000: 142–53). Similarly, Leonard Sumner (1976: 153) also advocates for the perspective that death represents the cessation of existence. This understanding has numerous significant philosophical antecedents. For instance, Epicurus articulates a perspective in which death represents the end of our existence. This thesis, which is present in Epicurus' philosophy and also advocated by Olson, Sumner, and numerous others, is known in contemporary literature as the "termination thesis".

architecture as referring to physical structures, such as the presence of a sensory apparatus and the processing of internal and external representations, without requiring that these structures be neural or a part of a biological organism. In this way, our position provides a consistent definition of the concept of personhood that applies to both biological and artificial contexts.

6. Concluding remarks

In conclusion, we have explored the connection between cognitive architecture and personhood in this paper. After delving into the topic, we have come up with a set of criteria that must be met for an entity, whether it is biological or artificial, to be considered a person. These requirements include having propositionally structured intentional states, possessing some form of sensory capabilities, and having the ability to interact with the environment. We have argued that these criteria provide a framework for comprehending the concept of personhood, and the case of individuals in a persistent vegetative state, studied by Owen, serves as a prime example of the importance of these conditions and the possibility of personhood being a "hybrid". Owen's research suggests that these individuals still fulfill the outlined criteria, proving the significance of cognitive architecture in determining personhood. In the end, we are aware that further research is necessary to fully grasp all of the important implications of cognitive architecture in regards to personhood, for this complex concept demands a more profound understanding, which cannot be fully achieved within the confines of this study alone.

References

Aaltola, E. (2008). "Personhood and animals." *Environmental Ethics* 30, 175–193.

Asimov, I. (1990). *The bicentennial man and other stories*. Gollancz.

Baker, L. R. (2000). *Persons and bodies*. Cambridge University Press.

Blatti, S., & Snowdon, P. F. (2016). *Animalism: New essays on persons, animals, and identity*. Oxford University Press.

Boden, M. (1990). *The philosophy of artificial intelligence*. Oxford University Press.

Chisholm, R. (1991). "On the simplicity of the soul." *Philosophical Perspectives* 5, 167–181.

Cranford, R. E., & Smith, H. L. (1979). "Some critical distinctions between brain death and the persistent vegetative state." *Ethics in Science and Medicine* 6, 199–209.

Davidson, D. (1982). "Rational animals." *Dialectica* 36, 317–28.

Dreckmann, F. (1999). "Animal beliefs and their contents." *Erkenntnis* 51, 597–615.

Duch, W., Oentaryo, R. J., & Pasquier, M. (2008). "Cognitive architectures: Where do we go from here?" *Frontiers in Artificial Intelligence and Applications* 171.

Fellows, R. (2000). "Animal belief." *Philosophy* 75, 587–599.

Haugeland, J. (1985). *Artificial intelligence: The very idea*. MIT Press.

Huebner, B. (2013). *Macrocognition: A theory of distributed minds and collective intentionality*. Oxford University Press USA.

Langley, P., Laird, J. E., & Rogers, S. (2009). "Cognitive architectures: Research issues and challenges." *Cognitive Systems Research* 10, 141–160.

Lazović, Ž. (2009). "Neurofilozofija na delu: filozofske pouke neuroloških defekata." *Theoria* 52, 115–125.

Lepore, E., & Pylyshyn, Z. (1999). *What is cognitive science*. Wiley-Blackwell.

Maslin, K. (2001). *An introduction to the philosophy of mind*. Polity Press.

Mackie, D. (1999). "Personal identity and dead people." *Philosophical Studies* 95, 219–242.

Milojević, M. (2018). *Metafizika lica*. Institut za filozofiju, Beograd.

Noonan, H. (2019). *Personal identity*. London: Routledge.

Olson, E. (1997). "Was I ever a fetus?" *Philosophy and Phenomenological Research* 57, 95–110.

Olson, E. (2000). *The human animal: Personal identity without psychology*. New York: Oxford University Press.

Olson, E. (2019). "The metaphysics of artificial intelligence." In Guta, M. P. (ed.), *Consciousness and the ontology of properties*. Routledge, 67–84.

Owen, A. M., Coleman, M.R., Boly, M., et al. (2006). "Detecting awareness in the vegetative state. *Science* 313, 1402.

Pollock, J. (1989). *How to build a person*. MIT Press.

Putnam, H. (1964). „Robots: Machines or artificially created life?" *Journal of Philosophy* 61, 668–91.

Rowlands, M. (2019). *Can animals be persons?* Oxford University Press.

Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach*. Pearson.

Searle, J. (1980). "Minds, brains and programs." *Behavioral and Brain Sciences* 3, 417–24.

Snowdon, P. F. (1990). "Persons, animals, and ourselves." In Gill, C. (ed.), *The person and the human mind: Issues in ancient and modern philosophy*. Oxford: Oxford University Press.

Sokić, M. (2020). "Lični identitet i teorija psihološkog kontinuiteta." *Theoria* 63, 87–104.

Sumner, L. W. (1976). "A matter of life and death." *Noûs* 10, 145–171.

Sutton, C. S. (2014). "The supervenience solution to the too-many-thinkers" *Philosophical Quarterly* 64, 619–639.

Turing, A. (1950). "Computing machinery and intelligence." *Mind* 59, 433–60.