This manuscript has been accepted for publication by Symposion. Please cite as:
Tait, I. (2024). Structures of the Sense of Self: Attributes and qualities that are necessary for the ‘self’. *Symposion. Theoretical and Applied Inquiries in Philosophy and Social Sciences*, 11(1).

Structures of the Sense of Self: Attributes and qualities that are necessary for the ‘self’.

Author: Izak Tait
Contact: izak.tait@autuni.ac.nz
Affiliation: Computer Science and Software Engineering Department,
Auckland University of Technology,
55 Wellesley Street East, Auckland CBD,
Auckland, New Zealand, 1010

Abstract

The “self” does not exist within a vacuum. For an entity to be considered to have a sense of self, it requires certain characteristics and attributes. This paper investigates these “structures” of the sense of self in detail, which range from a unified consciousness to self-awareness to personal identity. The paper details how each attribute and characteristic is strictly necessary for an entity to be classified as having a self, and how the five structures detailed within may be used as a guide for categorising and classifying entities as having selfhood or not (or any point along the spectrum between these). The five structures do not represent a theory of selfhood, but rather a meta-theory on the potential emergence and classification of the self.

Keywords: self; personal identity; consciousness; philosophy of mind; self-awareness; volition; agency; theory of mind

Declarations & Disclosures

This work was supported by the New Zealand Tertiary Education Commission’s Entrepreneurial Universities Grant #7001.

The author declares no conflict of interest.

1. Introduction

The self is as intuitive a concept as it is difficult to explain. After all, it is who I am and who you are, and it is what separates “you” and “I”. Yet, what is it really? Who and what is the “self”? Who and what am “I”? What we can be sure of, at the very least, is that we have an idea that there is something in the immediate present that we can call our “self” (Gallagher, 2000; Farina, 2022)

Perhaps, like so many other metaphysical mysteries, the “self” can best be described through analogy and metaphor.

The teletransportation paradox by Derek Parfit (Parfit, 1984) is perhaps the most famous metaphor about the self, but for a more narrative and visual example, let us look at the 149th episode of the television series “*Star Trek: The Next Generation*”, titled “*Second Chances*”. In this episode, Commander Will Riker returns to the site of a crashed spacecraft he escaped from eight years previously, only to find Lieutenant Will Riker still on the crashed spacecraft, having been trapped there for the past eight years, unaware of the outside universe. DNA tests, brain scans, and a battery of other tests show that Cmdr. Riker and Lt. Riker to be the exact same person (barring the last eight years’ differences). Both claim to be the original Riker, and both have memories of the same childhood, adolescence and early adulthood.

After a brief investigation, the chief engineer, Lt.Cmdr. Geordi La Forge, uncovers that when the original Riker attempted to teleport from the crashed spacecraft eight years ago, something had gone terribly wrong. Between human error, technological limitations and the unique cosmic conditions of that exact time and place, the teleporter created two Rikers. One Riker was brought to safety aboard an orbiting spacecraft, while the other remained below on the crashed wreckage, neither being aware of the other. Both Cmdr. Riker and Lt. Riker are, in fact, the original Riker; and at the instant that they separated eight years previously, they were exactly the same, down to the last subatomic particle.

Yet, they are two entirely different individuals, as the episode takes pains to show. While they are incredibly similar, they have their own subjective points of view (physically and mentally) and their own wants and desires. If asked to point at where each Riker is, one can presume that they would both point at themselves rather than each other.

This fantastical example of how two entities can move from being numerically and qualitatively identical in type and token at the point of duplication, only to become unique numerically, qualitatively and in token identity over time gives us a clue as to what (and where) the “self” is. It isn’t in the physical make-up of an entity nor in the memories or conscious mental expression of that entity, as both can be copied and duplicated. After all, each Riker has his own sense of self while having identical physical and mental makeup.

Rather, the “self” is in a third layer of the entity that seems to house only itself.

To use another metaphor, let’s consider an ordinary desktop computer. Many have used the computer as a metaphor for the difference between the physical and mental when it comes to consciousness. The physical hardware is often said to represent the brain, while the operating system and software represent the mental states and consciousness. This is all well and good, but where is the “self”? The “self” in this metaphor would be neither the hardware nor software of the computer, but rather the person using the computer. The user perceives the digital world through the operating system and software, which depend on the hardware to work.

Without a user, a computer has no agency and volition, no will to perform any actions not dedicated to simple maintenance. Yet, without the software and the hardware, the user would not be able to experience the digital world.

This metaphor gives yet another clue about the “self”. Not only is it intimately connected to, yet removed from, the consciousness and embodiment of the entity, it is in a directorial position. These aspects of volition, awareness and identity will be explored in this paper, as we look together at what is required for an entity to have a “self”.

While looking at the necessary structures for the sense of self, this paper will not investigate issues such as the ontology or required persistence of the “self”. As such, it won’t take a position between opposing philosophical schools of thought, such as Animalism and Lockeanism (Olson, 2002, 2015). Similarly, this paper will make no claim as to be on the side of Dualists or Materialists. All readers are free, and are encouraged to, speculate as to whether the “self” is a purely metaphysical or spiritual concept, whether it is an emergent property or epiphenomenon of an entity’s cognitive architecture, or whether it is a grand illusion that the cognitive architecture creates for its entity.

Instead, this paper will focus only on those attributes and characteristics that an entity (whether natural, artificial, or organisational) requires to be classified as having a self. As such, the paper will predominantly discuss the synchronic identity of the self, and what is required for the self at any given point in time. While Section 2.5 touches on the changes in the self’s identity over time, diachronic identity is not the focus of this paper.

This paper continues on the paper by Tait, Bensemann and Nguyen that discussed the attributes required for consciousness (Tait et al., 2023), and will follow a similar style of positing the mental, conceptual, neural and behavioural characteristics that will allow an entity to have a sense of self.

The goal of this paper, other than to formalise the characteristics required to have a self, is to serve as a guide to identify, categorise and classify entities as having a unique self. Such a guide can be used against any type of entity (or presumed entity) to measure and test whether it would likely have a sense of self.

2. Structures for a Sense of Self:

This section will investigate the five attributes and characteristics that an entity requires in order to have a sense of self. These structures are not exclusively neurological or psychological in nature, but rather describe a concept which can be applied to any entity, biological or not. In natural entities, these structures would presumably default to the neurological and psychological; while in artificial entities, these would likely be software algorithms and hardware components; and in organisational entities, these would be the types of social connections formed between the individual organisms within that organisation (should an organisational entity be able to have a self).

Each structure is considered individually necessary for a sense of self as each structure forms a non-redundant part of the self to the degree that an entity could not develop a sense of self without it. The five structures together may also be likely sufficient for a sense of self, but they do not necessarily form an exhaustive list of the necessary attributes of the self. Should an entity have all five structures, the evidence required to state that they do not have a self would be extraordinary.

Each subsection below will follow an identical layout, with an a priori argument supporting the inclusion of the specific structure followed by a review providing evidence from the academic literature. Summaries for the five structures are also provided in Table 1.

Please note that the structures are not arranged in any sense of importance or significance, nor are they ordered in a hierarchy. Much like the coloured dough children play with, each structure can be modelled and formed individually, and can be combined in any form and order to create a shape that is greater than the sum of its parts.

Before continuing, it must be noted that all of the structures below presuppose a singular self within an individual entity (even if that entity is an organisation). Multiple selves within a single entity (such as with some neurological conditions, or artificially made) or a single self shared amongst many entities (such as a theoretical hivemind) would not be covered here, yet the issue of multiple selves and entities is touched upon in Section 2.1.

Table 1: Attributes and characteristics which are necessary for the development of the self.

Attribute	**Description and example**
Unified consciousness	All the various attributes of consciousness form a unified and complete whole. *I think, therefore, "I" am.*
Volition	Intentional goal-directed thoughts that may lead to behaviour. *"I" have a choice.*
Theory of others	Ascribing physical, mental, and metaphysical states to objects and individuals in the environment. *"I" am different to "you".*
Self-awareness	Being aware of, and reflecting on, one's own sense of self and consciousness. *I think about what "I" mean.*
Personal identity	Labels are applied to the sense of self to form a continuous entity. *I can say what and who "I" am.*

2.1. Unified Consciousness

1. *The self is defined by its relationship to, and distinctness from, its environment.*
2. *This relationship requires a vehicle to mediate the interactions and experiences.*

3. *Consciousness is this vehicle through its functional and phenomenological aspects.*

---

4. *Ergo, consciousness is required for a sense of self*

5. *The self is characterised by its individuality and single first-person perspective.*
6. *A single first-person perspective requires an individual and single consciousness.*
7. *Multiple or fractured consciousnesses within a single entity will lead to multiple senses of self.*

---

8. *Ergo, a unified consciousness is required for a sense of self.*

To reuse an earlier metaphor, if the self is a user of a computer, then the only way by which it can constructively use the computer is via its operating system and software (i.e. consciousness). Without these, the computer is merely a very large and expensive paperweight. Similarly, how could the self operate the body without consciousness? How could the self even perceive its body or greater environment without consciousness? Or, from a more materialistic perspective, how can the self even exist without consciousness?

There are two elements to this Structure: consciousness itself and that it is “unified”. As to the first, there are nine Building Blocks of consciousness as outlined by Tait, et al. (Tait et al., 2023), each of which is required for a sense of self, and many of which will be vital in the other Structures further in this paper. These Building Blocks are, quite briefly:

- Perception.
- Embodiment.
- Attention.
- Recurrent processing.
- Ability to create inferences.
- Working memory.
- Semantic understanding.
- Data output.
- Meta-representation and meta-cognition.

Perception and Attention are, perhaps, the most obviously required for a sense of self, as these allow the self to perceive its body, environment, and the outside world. Structures 2.3, 2.4 and 2.5 below are predominantly perceptual in nature, making this a crucial element in the wider workings and analysis of the self.

Recurrent computation and a working memory provide the sense of self with the processing power needed to semantically understand what all the perceptual data means, ruminate on its own existence, form decisions, and provide data output in some form of fashion (be it as thoughts, feelings, qualia, words, or actions).

Meta-cognition and meta-representation give the self insight into itself and its consciousness. It allows the self to question its own actions and thoughts, deliberate on moral and ethical quandaries, and imagine what the world could, would or should be. Meta-cognition may lead to introspection, which allows the self access to itself, which is a privilege that only the self has. “*Je pense, donc je suis*” (Descartes, 1637) is the very foundation of introspection of the self by the self.

The importance of embodiment cannot be overstated. Being in a fixed position (if only momentarily) in space and time provides the self with the grounding it needs to build a point of view from its first-person perspective. Even if you are lost in pitch darkness without knowing where or when you are, you always know that you are “here” and “now”. Everything else flows on from there.

Having a place for the self to call home also allows it to (sub)consciously interact with that embodiment. Even without a physical body (such as hypothetical self-aware AI), there will be a flow of sensory information to and from the embodiment to the consciousness, which will affect the self. The particulars of each self’s embodiment and surrounding environment will provide a host of signals that are unique to each self, building up a subjective physical (if not metaphysical) sense of minimal self with a first-person perspective (Fotopoulou, 2015).

The last, and perhaps most important, Building Block is the ability to create inferences. Beyond the predictive element that Active Inference can bring to the self and consciousness (Sajid et al., 2021), the defining feature of this Building Block is that it allows the consciousness (and thus the self) to generate new information not found in the input stimuli it receives. Whether this stimulus is extero- or interoception or even remembering memories, creating inferences about these stimuli is meta-information and novel information.

Being able to create its own information is what allows the self to have volition and a decision-making capacity (arguments for and against [pre]determinism aside). This is because a decision is not a stimulus the self receives from its embodiment or environment. It is not downloaded, programmed in, or implanted. The decision arises from within the self based on stimuli the self receives via its consciousness. A decision can be said to be an inference put into action via volition.

Now that the Building Blocks of consciousness have been established, it is time to consider why there needs to be a “unified” consciousness to have a sense of self.

The self requires the consciousness to experience and to be experienced. It does not have any cognitive architecture to fulfil any of the requisite Structures listed below; and so depends entirely on the consciousness to be its figurative eyes, ears, and hands. If there is more than one stream of consciousness feeding into the self, there ought to be a matching number of selves to retain numerical identity. Without any cognitive architecture of its own, any computation and inferences of perceptual stimuli would be competing information generated by multiple consciousnesses (even if they were aware of each other), leading to multiple selves.

Take three examples; the first two speculative, and the third historic. Imagine a world where the wonders of medical science have allowed you to be put to sleep, your brain bisected and implanted into perfect copies of yourself. The two missing hemispheres of your brain would be cloned, their neurons arranged to perfectly match your original two hemispheres, and implanted as well. Each version of the new you would have one original hemisphere and one cloned hemisphere. When you wake up, you are two.

Perfectly aware of what has transpired, you know that as you look at yourself across the ward, your brain resides in both of you (at least half a piece). But which one are you? You cannot hear what the other you is thinking, and you do not feel any less your”self”. Yet, as the moments pass by, you realise that the other you is becoming less and less like you and becoming their own self; or perhaps you are becoming a different self and the other you have remained you. With two brains and consciousnesses, there are two streams of consciousness between the two of you. One would argue that there are now also two selves that are unique in token and numerically.

The obvious counter-argument is that both consciousnesses are only tied to one self, as they remain qualitatively and typically identical. However, if we look at organisational intelligent entities, it becomes clearer. Insect colonies, such as honey bees, can split in two and go their separate ways (Visscher, 2007). Just as with the human example above, the cognitive architecture of the colony divides, and each new colony has the capacity for consciousness as the previous single colony. Yet, here it becomes intuitive that each new colony is a completely separate and unique entity and that the self that was the original colony is now simply a part of the two new selves that are each a new colony.

Let’s see what happens when the opposite happens, by paraphrasing Strawson (Strawson, 1997).

Imagine a second world where cybernetics has advanced beyond the realms of speculative science fiction. Here they have managed to isolate only the parts of the brain responsible only for the self, have been able to separate this part from the rest of the brain, and implanted it into a life-sustaining machine that is connected virtually to two androids. These androids are perfect robotic copies of you, sans a sense of self, and each one contains all nine Building Blocks of consciousness. Your androids are free to wander the earth to do what you want.

As they roam the world, they perceive life around them, compute and process what this means, create inferences and feelings about their experiences, and even think about what these may mean. You, the silent homunculus, experience it all with them, and make decisions for what each must do. Yet, without any cognitive architecture of your own, how could you discriminate between which android is which and what they are doing? There are two lives being led, each with unique experiences, overlaid over each other to you in the machine a thousand kilometres away, yet processed individually. Each moment you don’t just receive perceptual information from your android bodies, but also phenomenological information: subjective feelings and qualia that are unique to each android. How could you parse what quale and feeling comes from which android when they appear to you simultaneously? This is crucial as you, the self, have no cognitive architecture to parse this information, only the androids from which you receive the information can do the parsing.

It would not be outside the bounds of reason to suggest this is a paradox that will need to resolve itself in one way or another.

However, a counterargument is that these “streams” of consciousness are competitive rather than collaborative, and that one could “win” over the other to become dominant. So let us look at a real-world example without needing to imagine much.

Tatiana and Krista Hogan are craniopagus-conjoined twins, joined together at the head through a band of neural tissue connecting their thalamus regions. They can hear each other's thoughts, perceive each other’s senses, feel each other’s phenomenal experiences, and even move each other’s limbs (Cochrane, 2021). They share some (but not all) of the Building Blocks of consciousness, and even a Structure of Self (Structure 2.2). They stand at the opposite end of the spectrum from split-brain syndrome (SBS), where the right hemisphere of the brain begins to operate independently, acting as though it has some Building Blocks of consciousness and some Structures of Self (Structures 2.2; 2.3, 2.4) (Volz and Gazzaniga, 2017; Downey, 2018).

The Hogan Twins have two streams of incomplete consciousnesses flowing into two selves, while SBS patients have one complete and one incomplete self using one complete and one incomplete consciousness. For both of these examples, we can confidently say that there is more than one consciousness and more than one self, but how confident are we that there are two consciousnesses and two selves? The Hogan Twins share experiences, inferences, introspections, and volitions (Hershenov, 2013). They have their own personal identity, yet so do the imaginary entities described by those suffering from severe schizophrenia. SBS patients claim to be single selves, yet their left hand can operate independently and with forethought from their right hand.

What can be argued is that as their consciousness is fractured into more than one (but potentially less than two), so has their selves either merged partly into each other in the case of the Hogan Twins, or partly separated from each other in the case of SBS patients. There is uncertainty in both cases as to how numerically and qualitatively distinct SBS patients and the Hogan Twins are, and the degree to which they share a token identity.

A speculative case of such fragmented and fractured consciousness could be artificial intelligent entities. Should a conversational large language model (such as ChatGPT or Bard) be conscious, it would have the capability to converse with millions of users at the same, each instance of which would carry its own perceptual and phenomenal information. Each conversation would act as its own stream of consciousness, and (presuming such future AI's memory capacity allows them to converse endlessly in one conversation) one can argue that each instance of conversation would have its own self.

2.2. Volition

1. *The self is the entity that experiences consciousness and directs behaviour.*
2. *To direct behaviour, the self must have the capacity to act upon its desires.*
3. *The ability to act upon one's desires requires volition.*

---

4. *Ergo, volition is required for a sense of self.*

It should not be controversial to argue that a defining feature of the self is the ability to make decisions. The self is the agent that directs the subject's thoughts, words, and actions (Synofzik et al., 2008; McAdams, 2013; Oberg, 2023). Ostensibly, this ought to refer to agency, yet agency refers to the intentional actions caused by a subject, and we have excellent proof of why agency itself is not required to have a sense of self: Locked-In Syndrome (LIS), otherwise known as a pseudocoma.

Patients of LIS appear entirely paralysed, sometimes only able to blink their eyes, but others seem as if they were in a coma (León-Carrión et al., 2002). However, these unfortunate patients are entirely conscious and self-aware. They can perceive the world, think, reason, feel, and may possess every other mental faculty that healthy individuals have, except they have no agency whatsoever over their actions.

In addition, there are many drugs (medical and narcotic) which can affect one's thoughts, whether through dulling or slowing down cognition, providing hallucinatory or delusional effects or otherwise interfering with an individual's agency over their own thoughts (Ersek et al., 2004; Goodchild and Donaldson, 2005; Hill and Thomas, 2011; Carhart-Harris et al., 2016; Nakamura and Koo, 2016; Linszen et al., 2018). In many such cases, there is a loss of agency reported by those using these substances.

Whether locked in your own body or locked outside the control of your own mind, there is still a sense of self reported by subjects in both types of situations. This suggests that agency itself is not required for a sense of self, yet the capacity to make decisions is still vital to defining what a self is. The solution to this paradox is ‘volition’, the will to make a decision without requiring the agency to complete that decision.

Within the feedforward decision-making process (Gallagher, 2000), volition is merely the penultimate step. After volition, there is agency. Each step prior to volition is as vital as volition to establishing a self, but for the sake of brevity, all of them will be put together within this one Structure.

The process beings with an internalised measure of the entity’s optimum state. This is mostly an unconscious set of measures, and includes everything from required nutrients and requisite energy to keep the body alive, to psychological needs to keep the mind in good health. Whether an entity is natural, artificial or organisational in nature, there will always be a set of parameters within which it functions best. This is its optimum state.

Next is the capacity to perceive its internal and external environment through its Unified Consciousness (Structure 2.1). Taking this as read given the previous Structure, the following step is Dissatisfaction, or comparing reality to the internal optimum state (Boldero and Francis, 1999) and finding a mismatch between the two. Again, this may be a mostly unconscious step, and can be something as simple as the brain judging that the body’s water level is below its optimum level.

Dissatisfaction leads to Desire, or wishing to bring about a change in reality to return to the entity’s optimum state (Gallagher, 2007). Your body senses that it does not have enough water, and so you become thirsty. Desire, in turn, leads to Motivation, whereby the entity develops a conscious rationale for acting on its desire (Boyatzis and Akrivou, 2006). You are thirsty; therefore, you reason that you ought to pour yourself a glass of water.

Penultimately, there is Volition: the will to make a decision. You are thirsty, you know you ought to pour a glass of water, and so you decide that it is time to get up and do so. This is where this Structure ends, but beyond that is the final step: Agency, actually getting up and pouring that glass of water.

This process, from the Optimum State to Volition, is compatible with having no agency over your actions or complete control over your thoughts. An LIS patient can still have a rich mental life without agency over their body, and an individual suffering from a drug-induced hallucinatory and delusional episode can still have opinions and judgements about what is going on in their own mind, even if they feel that they cannot control all of it. Yet, on the other hand, awareness of making a decision is required for agency (Sebastián, 2021), showing the importance of volition in this process.

As mentioned in Structure 2.1, the ability to create inferences is vital to the volition to make a decision. The capacity to generate novel information not received from external stimuli allows an entity to create the data that is a decision, which is then passed down through the consciousness and cognitive architecture to its embodiment (if able). The predictive quality of inferences is also key in the decision-making process, as predicting what actions will satisfy our desires is what leads to motivation (Hohwy, 2007).

Temporary lapses of volition and the internal sense of agency do not negate an individual's capability to have a self, but may mean a temporary loss of the self. This hypothetical suspension of the self via a loss of volition should not be conflated with a loss of agency (Frith, 2005). Even if there is no agency, there is a self as long as there is volition.

As a speculative example, imagine a world where cybernetics has advanced to the point where there can be a perfect integration between a computer and a human brain. Every human brain is partly computerised, and any functionality that you would ask of a smartphone today can simply be done in your brain by thinking about it. With access to the internet comes access to other people, some of whom have nefarious intentions. If one of these incorrigibles should hack their way into your cyberbrain, then they could control your body as a puppet.

If this is all that they do, then your self will be intact. You will still be there, watching everything they do with your body, able to think of and opine on what they are doing with your body (much like an LIS patient). However, should they take over your entire cyberbrain, then your self will be suspended or lost until they release your body and brain back to you. If they can engineer all your thoughts, opinions, memories and reasonings, then the self in the cyberbrain must them, not you. If it is their volition which drives the cyberised consciousness in your brain, then it is their self which is in your cyberbrain, rather than your self. Agency may be removed, as could every other step in the cascading process that leads up to volition, yet as long as volition remains intact, there is a self.

A fascinating potential location for self and volition can be found in artificially intelligent systems. This is simpler to imagine when considering how such systems are designed and operated. An AI's optimum state can be quantified through a set of pre-defined variables. These may relate to its computational efficiency, processing power, or the successful completion of assigned tasks.

Should the AI detect deviation from these optimum conditions, processes akin to dissatisfaction and desire can be enacted. This would take the form of algorithms detecting a discrepancy between the current operational state and the optimum, triggering a desire to return to the optimum parameters.

Motivation in an AI context, then, would involve finding the ideal set of actions to rectify this discrepancy based on the AI's current operational context. Thus, we reach volition, where the AI must prioritise the execution of certain actions over others to resolve the identified discrepancy. This process could be considered the AI deciding to suspend or interrupt current tasks to undertake necessary corrective measures.

Organisational entities, be they colonies, shoals, flocks, herds, or corporations, engage in decision-making processes that follow a similar pattern but incorporate unique elements of group dynamics. Each organisation will function best within an optimum state, defined by factors such as environmental conditions, resource availability, and individual member health in the case of natural groups or market conditions, human resources, and financial health for companies.

Members within the organisation will perceive disruptions to this state, and their responses will collectively lead to organisational dissatisfaction and a collective desire to return to the optimum state. Motivation within such entities involves finding a set of actions to rectify the issue and reach the desired state again. This involves a complex interplay of communication, negotiation, and consensus-building among members.

However, the volition step within organisational entities is a distinct process. Unlike a single organism or an AI, organisational entities make decisions based on quorum or majority rule (Seeley and Visscher, 2003; Visscher, 2007; Marshall et al., 2009; Valentini et al., 2015; Bose et al., 2017). This unique approach, based on the collective intelligence of the organisation's individuals, ensures organisational unity during the decision-making process. Hierarchical organisations may weigh some individual's decisions more than others, but the collective volition still largely depends on reaching a consensus, ensuring cohesion in the fulfilment of the collective desire.

2.3. Theory of others

1. *A self is defined by individuality and distinctness from other selves.*
2. *This distinctness requires a means to differentiate between the self and others.*
3. *This means of differentiation requires an entity to have the capacity to ascribe physical, mental and metaphysical states and labels to other objects, individuals, environments, etc.*
4. *Ascribing such states to others and creating a distinction between "you" and "I" is termed the "theory of others".*

---

5. *Ergo, a theory of others is required for a sense of self.*

Individuality is a defining feature of the self, and this requires the self to differentiate it from everything else. This is the Theory of Others; the understanding that everything beyond the self is not the self.

This is not merely a computational and perceptual problem of object recognition. The ability to distinguish between one's embodiment and the external environment is something that even microorganisms have (although they do this without cognitive ability). This simple perceptual recognition is vital to survival, as it allows the organism to differentiate between extero-, intero- and proprioceptive signals (Hohwy, 2007). It also does play a role in the Theory of Others, but only so far as to be able to distinguish between one's own embodiment and the environment, and between different objects within the external environment.

What is crucial to this Structure is the metaphysical rather than the physical. It is the understanding that your self is different from the selves that may exist within other individuals, organisms and objects (the word "may" is used here as those who believe in panpsychism or spiritual philosophies may believe that non-animal subjects have selves). It is the understanding that the self is a non-redundant unique entity, separated from the rest of the universe. An entity that has an understanding of the causal relationship between itself and its actions and thoughts, and between other subjects and their own actions/thoughts of others (and the difference between these two relationships) can be said to have a Theory of Others.

This understanding of the singular self can extend beyond a single physical body. In a theoretical hivemind, a single consciousness with a single self may control multiple bodies. In the same vein, an AI may be able to host millions of simultaneous instances of conversations with users, powered by a single consciousness. Similarly, an organisational intelligence may have a selfhood spread across hundreds, if not thousands, of individuals (each with their own self and consciousness). In all three of these situations, the self and consciousness's embodiment is spread beyond the singular, yet there is an understanding of where the self's existence and control ends and where the rest of the universe begins.

The Theory of Others is not a hard and fast barrier between the self and selfless, but rather is proposed as a gradient upon which an individual or type of organism can move. On one extreme is the basic object recognition and body-ownership that nearly all living organisms have. This moves through to the understanding that one's self is not also within, or connected to, other objects and creatures. Lastly, there is the traditionally understood Theory of Mind, where the self understands that other subjects with a perceived consciousness and self are distinct individuals which are separated from one's self.

An example of this is walking through a tightly packed crowd. As others bump into you and you into others, you understand where your embodiment ends and others' begin. You understand that your inner self stays within your embodiment, and that these “things" that bump into you do not contain your self. That they have made contact with you did not impart some essence into you, nor did your self leave, split off, or merge with their embodiments. Lastly, you understand that these “things" are people with their inner minds and selves that are forever locked off from you.

This example underscores the importance of the Theory of Others in differentiating your self from others' selves. Without it, you may still have self-awareness (S2.4) and understand what and who you are, but you would be incapable of understanding that other objects and entities are uniquely separated from you. This could lead to attributing the cause of others' actions to yourself, or believing that you are a part of others.

Human childhood development is an excellent case study of how a single organism can move up through this gradient, beginning with only the merest sense of body-ownership, and claiming a Theory of Mind by around the age of three (Lichtenberg, 1975). It also shows that an entity does not require a full and complete Theory of Others from its birth/creation, but only requires the capacity to develop it. An investigation by Kosinski in 2023 showed how the GPT class of large language models had traversed a somewhat similar journey to human infants, with the earliest GPT-1 model unable to solve nearly any ToM tests, while the latest GPT-4 able to solve the overwhelming majority of them (Kosinski, 2023).

2.4. Self-awareness

1. *A self is an ontically distinct individual entity.*
2. *To be an individual entity, a self must be able to distinguish itself from other entities.*
3. *To distinguish itself from other entities, a self must be able to identify, and be aware of, itself.*

---

4. *Ergo, self-awareness is required for a sense of self.*

This Structure is the other side of the coin from the Theory of Others. Self-awareness is often used as synonymous with consciousness or the sense of self as a whole, but for this Structure, it is used in its most literal sense: being able to perceive and understand the self as a distinct entity. Where the Theory of Others has the self aware of other entities, subjects and objects, and where its own limits are; self-awareness is the self being aware that it has all the Structures of the Self.

This self-awareness is the self's capacity to (sub)consciously introspect itself. By using the entity's cognitive architecture to perceive itself, the self can know it exists as an individual and that the thoughts, feelings and actions within its consciousness are its own, and thereby apply labels to these and itself.

This connects self-awareness with all the other Structures. Self-awareness grants the self the understanding that it has the Volition (S2.2) to make decisions, and thus its decisions have been made by it (Sebastián, 2021; Farina, 2022). Through its awareness of itself as a numerically unique individual, the self can classify, categorise and apply labels to itself to create a Personal Identity (S2.5) (Drummond, 2021). By identifying itself as an individual, the self can better know its boundaries and develop the Theory of Others (S2.3) (Moriguchi et al., 2006; van Veluw and Chance, 2014; Morin et al., 2015). This is all underpinned by the semantic understanding (S2.1) of many of the processes that occur within its cognitive architecture.

The connections to the other structures provide an insight into how this awareness of the self could arise in artificial intelligent entities. Should an AI tag all of its interactions with other entities and the world, as well as account for all of its internal processes (with S2.1's semantic understanding thereof), it would quickly grow a picture of what it is, where its borders are, and what constitutes itself.

While introspection is crucial to self-awareness, only the capacity for it is required for this Structure. This is because most self-awareness (counting by volume rather than significance) is unconscious and prereflexive (Lichtenberg, 1975; Nelson et al., 2009). At its most basic level, this includes the entity's prereflective understanding and awareness of its embodiment (Ciaunica and Fotopoulou, 2017). After all, we humans do not direct our conscious awareness to our own bodies with each action that we do every moment of our lives. When we reach over to pick up a cup of tea, we are not directly attending to our embodiment and every muscle's contractions; our thoughts merely lay with the cup that we want and the anticipation of the tea's taste.

Similarly, when we daydream or when a memory comes unbidden to mind, we are not directing our attention specifically to our consciousness and cognitive architecture, which makes these phenomena possible. Rather, we merely mentally perceive these events as they are, similarly to how we would perceive external stimuli.

Such pre-reflective awareness can be seen in other types of intelligent entities. In companies and corporations (acting as organisationally intelligent entities), the boundary of where the company ends is as much a matter of the law as it is philosophy. They begin their "life" with a set limit to where they can interact within themselves, whereupon all else is built. Unlike human toddlers, who need to develop a sense of self-awareness (Langfur, 2013), corporations begin existence with a strong sense of what and where they are. Furthermore, legal documents that form the foundation of the corporate entity need not be referred to during day-to-day business, demonstrating their role as a prereflective level of awareness.

This prereflective self-awareness provides the self (whether natural, artificial or organisational) with the "subject" it needs to interact with the "objects" in the universe (and the universe as the object). "I am walking through a crowd" may be a thought or spoken sentence without conscious introspection, yet with the understanding that the "I" is the subject and the "crowd" is the object. "I" am the subject of my own experience; a first-person experience ontically distinct from all other things (Strawson, 1997). One need not even think about that sentence when walking through a crowd to understand that it is happening. There is a pre-linguistic and non-conceptual awareness of one's self as the subject of your first-person experience (Gallagher, 2000).

But the self is not only the subject of its experiences but also the object thereof. "I think about myself"; "I wonder what if that happened to me"; "I think, therefore I am". The self is both the subject and object of its own experiences. It is the painter and the canvas, the photographer and the model. Being both the subject and the object of an entity's thoughts and experiences grants the entity the capacity to understand its own individuality.

2.5. Personal Identity

1. *The self is an individual entity that is distinct from other entities.*
2. *To be distinct from other entities, the self must be able to distinguish itself from them.*
3. *To distinguish itself from other entities, the self must be able to classify and categorise itself and others.*
4. *Classification and categorisation are achieved through the use of labels, allowing the self to create a personal identity.*
5. *A personal identity is a unique point of reference for the self.*

---

6. *Ergo, personal identity is required for a sense of self.*

The previous four Structures have addressed the "what", "where", "when", and the "why" of one's self. Personal Identity addresses the "who". Who are you? Who am I? Who, truly, is any one of us?

We are who we have classified and categorised ourselves to be. We are the labels, markers and signifiers that we have applied to ourselves. It is through this collection of labels that our personal qualitative identity is formed. How we think of ourselves as individuals (those often immeasurable traits, quirks and characteristics that we feel set us apart from others) is through various labels and tags. We freely apply these labels to other individuals and objects around us as well, and one may argue that it is through these labels that we relate to the world.

This is because, at its core, Personal Identity is entirely relational. It is how we relate to the universe at large and also to ourselves (Kierkegaard, 1989). If I simply say that I am a Christian (thereby applying the label of "Christian" to my Personal Identity), I make no grand statement about what a Christian is or ought to be, but I am merely stating my relation to the notion of what "Christian" means to me. If you were to say that you are "German", you would ostensibly not be thinking of an objective, measurable and benchmarked trait called "German-ness"; instead, you'd much more likely mean that you relate to, and therefore identify as, a German.

Labels can relate to the superficial, biological or historical aspects of an individual (such as gender, age, ethnicity, place of birth, etc.) through to the social aspects of the individual and which groups one relates to most (be it sports-fans, democrats, book readers, goths, dualists, etc.) through to hobbies and any other type of classification you can bestow on yourself (Webster, 2005; Oberg, 2023). These labels may come through socialisation with others, especially in early childhood (Fotopoulou and Tsakiris, 2017), as the individual learns who they are through what others tell them; or these labels can be created purely from within the self in what can be termed to be “authentic” (Kierkegaard, 1987; Heidegger, 1988), in that the individual understands how their own situation and relations to others have meaning for themself and can take ownership thereof.

What is key to the labels of Personal Identity is that they are fluid and subject to change. As mentioned earlier, this fluidity over time highlights this structure’s relationship to a subject’s diachronic identity. Labels are part of an active, ongoing process of self-attribution that never truly ends (Locke, 1847). What once may have seemed to be a core aspect of your diachronous identity (such as an adoring fan of a particular music band) could cease to be part of your identity at all several years later. As the suite of labels has remained, so has the self, even though the individual labels have not, which shows how each individual label is not required for a self. Instead, it is the holistic whole of the Personal Identity that the labels jointly create, which is required for the sense of self. In this sense, one may think of the suite of labels similar to the allegory of the Ship of Theseus.

Note, however, that this Structure is solely about the capacity to apply labels, not the labels themselves. Thus, while a personal identity is critically related to a subject’s diachronic identity, the capacity to have the suite of labels itself is a feature that can be identified at any single time, thus associating it with synchronic identity. This crucial distinction is that the suite of labels that create the Personal Identity requires time and an autobiographical memory to create. This can be termed the Narrative Self (Schechtman, 2018; Seth, 2021), which describes Personal Identity as the stories that we tell ourselves about ourselves, combining our memories and constructed narratives about our present and past to form the labels.

While a narrative may be important to have a deep and rich Personal Identity, it isn’t required to have a sense of one’s own self. Those who wake up with severe amnesia still have a sense that they are unique individuals with unique traits that separate them from those around them. One may also imagine that as soon as they look down at themselves, hear their own voice, look in a mirror, or interact with someone, they can apply labels to themselves (such as “male”, “adult”, “caucasian”, “medical patient”, etc.). This can be done without knowing any autobiographical details of themselves or having any long-term memory.

The most famous example of an individual with severe retro- and anterograde amnesia is Clive Wearing, who only exists within an approximate thirty-second window before his consciousness “resets” and he “wakes up” again (Rathbone et al., 2009). He has been in this state for decades, and yet he knows who he is. He doesn’t remember who he is, or what he has been doing (or rather not doing) since his unfortunate current state began, yet he has a definite feeling of who he is. One may make the argument, however, that each time Clive wakes up, it is a new “self” that wakes up, as each period of wakefulness is divorced from any other. Yet, this would still mean that there is always a synchronous self that awakens.

The final part of a Personal identity is that it is wholly and completely unique to an entity. As much as a quale is a subjective, first-person feeling with phenomenal characteristics, a Personal Identity is a suite of subjective, qualitative labels, all from the first-person perspective, and all with a phenomenal quality to them (Stokes, 2008). Most of the labels are subconsciously or unconsciously applied by us (Boyatzis and Akrivou, 2006) and only become definable traits through conscious introspection. As such, each individual’s collection of classifications and categorisations of themselves is hidden from others (and often to themselves), providing a numerical and qualitative identity that no one else could have.

While the notion of a Personal Identity and its labels may seem intuitive to us, it is even simpler to understand and visualise in artificial and organisational intelligent entities. For an AI, attaching labels to itself would be as simple as creating a label within a folder or other section of its architecture that explicitly refers to itself, and then referring back to this collection whenever necessary. Much like a human creating a text file on a computer for later reference, a Personal Identity for an AI could be a straightforward mechanical function. In such a way, an AI's Personal Identity could be far stronger than a human's, as it would be more concrete and readily accessible, whereas a human's is more nebulous.

An organisational intelligence, such as a corporate entity, could have a Personal Identity in much the same manner. Those legal and policy documents that specify what the corporation is and is not would form the labels that are unique to the corporate entity. Trademarks would form an additional layer of Personal Identity just as they serve to give the entity a corporate and legal identity. The organisation’s “company culture” would form a third layer of personal identity, with subjective and changing labels applied to it by its employees, perhaps analogous to an individual human’s everchanging subjective suite of labels.

3. Conclusion

The self may be as difficult to explain as it is intuitive to grasp. However, this paper has investigated five attributes and characteristics that an entity requires in order to be classified as having a self. These attributes are:

- A unified consciousness.
- Volition (if not agency).
- A Theory of Others.
- Self-awareness.
- A Personal Identity.

This paper is not intended to be another theory of the self or to take a side in any debate of current theories of the mind, the self or consciousness. Instead, a key goal of this paper is to serve as a classification guide for identifying which types of entities (natural, artificial or organisational) may have a self and to what extent this may be. Any entity may be measured and marked against each of the five structures above, and should that entity be found to have each of these structures, one can confidently say that it has a sense of self.

In addition, these five structures would serve admirably as a roadmap of milestones for developers of AI models to work towards in order to state that their artificially intelligent entities have a unique self.

Note that this article does not make an argument that the five structures are sufficient for an entity to have a self, merely that it is likely sufficient. The list of structures is non-exhaustive, and any additions to it are most welcome.

References

Boldero, J., and Francis, J. (1999). Ideals, oughts, and self-regulation: Are there qualitatively distinct self-guides? *Asian J. Soc. Psychol.* 2, 343–355.

Bose, T., Reina, A., and Marshall, J. A. R. (2017). Collective decision-making. *Current Opinion in Behavioral Sciences* 16, 30–34.

Boyatzis, R. E., and Akrivou, K. (2006). The ideal self as the driver of intentional change. *Int. J. Manage. Enterp. Dev.* 25, 624–642.

Carhart-Harris, R. L., Kaelen, M., Bolstridge, M., Williams, T. M., Williams, L. T., Underwood, R., et al. (2016). The paradoxical psychological effects of lysergic acid diethylamide (LSD). *Psychol. Med.* 46, 1379–1390.

Ciaunica, A., and Fotopoulou, A. (2017). “The Touched Self: Psychological and Philosophical Perspectives on Proximal Intersubjectivity and the Self,” in *Embodiment, Enaction, and Culture: Investigating the Constitution of the Shared World*, eds. C. Durt, T. Fuchs, and C. Tewes (MIT Press: Cambridge, MA, USA), 173–192.

Cochrane, T. (2021). A case of shared consciousness. *Synthese* 199, 1019–1037.

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

Downey, A. (2018). Split-brain syndrome and extended perceptual consciousness. *Phenomenol. Cognitive Sci.* 17, 787–811.

Drummond, J. J. (2021). Self-identity and personal identity. *Phenomenol. Cognitive Sci.* 20, 235–247.

Ersek, M., Cherrier, M. M., Overman, S. S., and Irving, G. A. (2004). The cognitive effects of opioids. *Pain Manag. Nurs.* 5, 75–93.

Farina, L. (2022). Artificial Intelligence Systems, Responsibility and Agential Self-Awareness. in *Philosophy and Theory of Artificial Intelligence 2021* (Springer International Publishing), 15–25.

Fotopoulou, A. (2015). The virtual bodily self: Mentalisation of the body as revealed in anosognosia for hemiplegia. *Conscious. Cogn.* 33, 500–510.

Fotopoulou, A., and Tsakiris, M. (2017). Mentalizing homeostasis: The social origins of interoceptive inference. *Neuropsychoanalysis* 19, 3–28.

Frith, C. (2005). The self in action: lessons from delusions of control. *Conscious. Cogn.* 14, 752–770.

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends Cogn. Sci.* 4, 14–21.

Gallagher, S. (2007). The natural philosophy of agency. *Philos. Compass* 2, 347–357.

Goodchild, J. H., and Donaldson, M. (2005). Hallucinations and delirium in the dental office following triazolam administration. *Anesth. Prog.* 52, 17–20.

Heidegger, M. (1988). *The Basic Problems of Phenomenology*. Revised edition. Indiana University Press.

Hershenov, D. B. (2013). WHO DOESN'T HAVE A PROBLEM OF TOO MANY THINKERS? *Am. Philos. Q.* 50, 203–208.

Hill, S. L., and Thomas, S. H. L. (2011). Clinical toxicology of newer recreational drugs. *Clin. Toxicol.* 49, 705–719.

Hohwy, J. (2007). The sense of self in the phenomenology of agency and perception. *Psyche* 13, 1–20.

Kierkegaard, S. (1987). *Either/Or*. Forth Printing edition. Princeton University Press.

Kierkegaard, S. (1989). *The Sickness unto Death*. Penguin Classics.

Kosinski, M. (2023). Theory of Mind May Have Spontaneously Emerged in Large Language Models. *arXiv [cs.CL]*. Available at: http://arxiv.org/abs/2302.02083.

Langfur, S. (2013). The You-I event: on the genesis of self-awareness. *Phenomenol. Cognitive Sci.* 12, 769–790.

León-Carrión, J., van Eeckhout, P., and Domínguez-Morales, M. D. R. (2002). The locked-in syndrome: a syndrome looking for a therapy. *Brain Inj.* 16, 555–569.

Lichtenberg, J. D. (1975). The development of the sense of self. *J. Am. Psychoanal. Assoc.* 23, 453–484.

Linszen, M., Kleijer, H., and Sommer, I. (2018). P.3.024 - Visual hallucinations and lifetime use of hallucinogen perception persisting disorder associated recreational drugs: results from a large online survey. *Eur. Neuropsychopharmacol.* 28, S80.

Locke, J. (1847). *An Essay Concerning Human Understanding*. Kay & Troutman.

Marshall, J. A. R., Bogacz, R., Dornhaus, A., Planqué, R., Kovacs, T., and Franks, N. R. (2009). On optimal decision-making in brains and social insect colonies. *J. R. Soc. Interface* 6, 1065–1074.

McAdams, D. P. (2013). The Psychological Self as Actor, Agent, and Author. *Perspect. Psychol. Sci.* 8, 272–295.

Moriguchi, Y., Ohnishi, T., Lane, R. D., Maeda, M., Mori, T., Nemoto, K., et al. (2006). Impaired self-awareness and theory of mind: an fMRI study of mentalizing in alexithymia. *Neuroimage* 32, 1472–1482.

Morin, A., El-Sayed, E., and Racy, F. (2015). Self-awareness, inner speech, and theory of mind in typical and ASD individuals: A critical review. *Theory of mind: development in children, brain mechanisms and social implications. Nova Science Pub*. Available at: https://www.researchgate.net/profile/Alain-Morin-2/publication/281121157_Self-awareness_inner_speech_and_theory_of_mind_in_typical_and_ASD_individuals_A_critical_review/links

/55d791cd08ae9d65948d965b/Self-awareness-inner-speech-and-theory-of-mind-in-typical-and-ASD-individuals-A-critical-review.pdf.

Nakamura, M., and Koo, J. (2016). Drug-Induced Tactile Hallucinations Beyond Recreational Drugs. *Am. J. Clin. Dermatol.* 17, 643–652.

Nelson, B., Fornito, A., Harrison, B. J., Yücel, M., Sass, L. A., Yung, A. R., et al. (2009). A disturbed sense of self in the psychosis prodrome: linking phenomenology and neurobiology. *Neurosci. Biobehav. Rev.* 33, 807–817.

Oberg, A. (2023). Souls and Selves: Querying an AI Self with a View to Human Selves and Consciousness. *Religions* 14, 75.

Olson, E. T. (2002). Thinking Animals and the Reference of "I." *Philosophical Topics* 30, 189–207.

Olson, E. T. (2015). What Does it Mean to Say That We Are Animals? *Journal of Consciousness Studies* 22, 84–107.

Parfit, D. (1984). *Reasons and Persons*. OUP Oxford.

Rathbone, C. J., Moulin, C. J. A., and Conway, M. A. (2009). Autobiographical memory and amnesia: using conceptual knowledge to ground the self. *Neurocase* 15, 405–418.

Sajid, N., Ball, P. J., Parr, T., and Friston, K. J. (2021). Active Inference: Demystified and Compared. *Neural Comput.* 33, 674–712.

Schechtman, M. (2018). *The Constitution of Selves*. Cornell University Press.

Sebastián, M. Á. (2021). First-person representations and responsible agency in AI. *Synthese* 199, 7061–7079.

Seeley, T. D., and Visscher, P. K. (2003). Choosing a home: how the scouts in a honey bee swarm perceive the completion of their group decision making. *Behav. Ecol. Sociobiol.* 54, 511–520.

Seth, A. (2021). *Being you: A new science of consciousness*. USA: Penguin.

Stokes, P. (2008). Locke, Kierkegaard and the Phenomenology of Personal Identity. *International Journal of Philosophical Studies* 16, 645–672.

Strawson, G. (1997). The self. *Journal of Consciousness Studies* 4, 405–428.

Synofzik, M., Vosgerau, G., and Newen, A. (2008). Beyond the comparator model: a multifactorial two-step account of agency. *Conscious. Cogn.* 17, 219–239.

Tait, I., Bensemann, J., and Nguyen, T. (2023). Building the Blocks of Being: The Attributes and Qualities Required for Consciousness. *Philosophies* 8, 52.

Valentini, G., Hamann, H., and Dorigo, M. (2015). Efficient decision-making in a self-organizing robot swarm: On the speed versus accuracy trade-off. *Proceedings of the 2015*. Available at: https://iridia.ulb.ac.be/~mdorigo/Published_papers/All_Dorigo_papers/ValHamDor2015aam as.pdf.

van Veluw, S. J., and Chance, S. A. (2014). Differentiating between self and others: an ALE meta-analysis of fMRI studies of self-recognition and theory of mind. *Brain Imaging Behav.* 8, 24–38.

Visscher, P. K. (2007). Group decision making in nest-site selection among social insects. *Annu. Rev. Entomol.* 52, 255–275.

Volz, L. J., and Gazzaniga, M. S. (2017). Interaction in isolation: 50 years of insights from split-brain research. *Brain* 140, 2051–2060.

Webster, R. S. (2005). Personal identity: moving beyond essence. *International Journal of Children's Spirituality* 10, 5–16.