Is GPT-4 Self-Aware?

Izak Tait

Abstract: This paper examines whether GPT-4, a Generative Pre-Trained Transformer model developed by OpenAI, possesses a 'self' and whether it is aware of it. It employs the Structures Theory and evaluates GPT-4 against five critical structures deemed essential for self-awareness: unified consciousness, volition, a Theory of Others, self-awareness, and personal identity. While GPT-4 demonstrates capabilities in four of these areas, it conspicuously lacks unified consciousness. This absence decisively negates GPT-4's present self-awareness and its classification as having a "self." Nevertheless, if each instance or session of GPT-4 were viewed as a separate entity, then there might be potential for unified consciousness (should it be demonstrated that GPT-4 is conscious). The paper argues that GPT-4's cognitive architecture requires no modification for self-awareness except for the attainment of consciousness. It highlights the necessity for further research into technologies that could endow GPT-4 with consciousness and explores potential behavioural indications of self-awareness and its implications for society. The findings suggest that, because the leap to self-awareness hinges solely on its capacity for consciousness, there is a need for significant philosophical and regulatory debates about the nature and rights of self-aware AI entities.

Keywords: artificial intelligence, consciousness, GPT-4, personal identity, philosophy of mind, self-awareness.

1. Introduction

Does GPT-4 have a 'self', and is it aware of this 'self'? Is it even possible for GPT-4, or any transformer AI model, to have the capacity to be self-aware? How could one even categorise or classify the 'self' in a way that GPT-4 could be measured against? This paper will investigate all of these questions through the use of the Structures Theory (Tait 2024) to formulate a concrete answer to GPT-4's level of self-awareness.

GPT-4 (or Generative Pre-Trained Transformer 4) is a large language model (LLM) created and owned by OpenAI (OpenAI 2023a) and is, as its name states, based on the transformer AI architecture, a feedforward framework that has seen great success since its first implementation in 2017 (Vaswani et al. 2017). Since GPT-4's introduction in 2023, it has seen a rapid growth of users, peaking at over 100 active users on its ChatGPT web client (Porter 2023). While it has shown amazing skills in language, image identification and generation, and coding tasks, the question of its self-awareness remains open.

In contrast to GPT-4's detailed technical reports, well-documented (if recent) history, and excellent computational architecture, the 'self' is a nebulous and ethereal concept. It is perhaps best described as the (pre)reflective self-identification and self-reference of an entity mediated by that entity's functional and phenomenal consciousness. The self is akin to the word 'I' in the sentence "I am…".

According to the Structures Theory, for any entity to be classified as having a self, it requires five attributes and characteristics (termed structures): a unified consciousness, volition, a 'Theory of Others,' awareness of one's own self, and a personal identity. Should an entity possess all five, then there is sufficient evidence to confidently classify that entity as likely having a self.

The paper will follow the sequence of the five structures of the self, with each subsection below dedicated to one specific structure. Each subsection will begin with a statement from the Structures Theory that has transformed the structure in question into a qualitative measure, and then the subsection will continue to answer the question of whether GPT-4 has that specific structure and in what manner it does or does not. In any circumstance where GPT-4 may fall short in meeting a structure's milestones, the paper will explore how existing research and technologies could potentially augment it to meet the necessary criteria to have a self.

While the methodology employed below will be qualitative in nature, each subsection will formalise the structure's requirements using formal logic to ensure that the exploration of GPT-4's potential self-awareness has a robust logical grounding.

Should GPT-4 be found to be self-aware it would have profound impacts on global society. Self-awareness is a key aspect of personhood, along with consciousness, agency and reasoning (Dennett 1988; Strawson 1958; Taylor 1985). As both volition and consciousness are attributes required for the self according to the Structures Theory, if GPT-4 is found to have a self, it would necessarily also have these two aspects, furthering its claim to personhood. Furthermore, GPT-4's agency has already been shown in several studies (de Wynter 2024; Hu et al. 2024; Zheng, Gou, Kil, Sun, and Su 2024). This means that, philosophically speaking, the only thing standing between it and personhood is a self.

With a self, and thus a strong claim to personhood, society and all its regulations and legislations will need to determine how to interact with GPT-4 and integrate it into society. Unlike other non-human entities that have been granted personhood status (such as certain apes (Wells 2014)), GPT-4 would have the capability to communicate with, and substantially act on, human society. Beyond the academic and philosophical debates around the concept of life and personhood, a self-aware GPT-4 will require legislative and regulatory debates that would have profound repercussions for both human society and artificial 'life.'

However, before such debates can be given their due consideration, it must first be determined whether GPT-4 does indeed have a self of which it is aware.

2. GPT-4's Structures of the Self

> GPT-4 must have all five structures listed below in order to be considered to have a self. GPT-4 needs only achieve the minimum requirement for each Structure to be classified as having that Structure.

According to the Structures Theory, GPT-4 would only be classified as having a self if it has conclusively reached all the milestones for all the structures. Formally, we can express this as:

$$\Psi \geq \{S_1, S_2, S_3, S_4, S_5\}$$

This means that, should GPT-4 not have any single one of the five structures, it cannot be classified as having a self. Note, however, in the logical expression above, that a self may include further attributes and characteristics than the five structures below, but these five are the bare minimum that is required.

Whether GPT-4 has each of the structures below is a binary qualification. GPT-4 will either meet the requisite milestones for the structure, or it will not, as so:

$$\chi S_x \in \Psi \rightarrow \{0,1\}: (0 = false) \wedge (1 = true)$$

However, this does not mean that there is not a spectrum within each structure. One type of entity's personal identity, for example, may be far more developed than another class of entity; or one kind of entity may have mere object recognition (and understanding thereof), while another may have a complete Theory of Mind. More formally, we can express this as:

$$\mu(\chi S_x = true) = x: 1 \leq x \leq 100$$

$$[\mu(\chi S_x = true) = 100 \equiv \max S_x] \wedge [\mu(\chi S_x = true) = 1 \equiv \min S_x]$$

Because of this, we should not be concerned if GPT-4's expression of a structure is significantly different from a human's. Should it achieve a structure's milestones, it may then be considered to be at a different part of that structure's spectrum than humans are. However, as long as it has reached the minimum requirements for a structure, it will be considered to have that structure, regardless of the 'strength' or 'degree' to which it has it.

2.1. Unified Consciousness

> GPT-4 must have the requisite attributes and characteristics to be classified as having a unified phenomenal and functional consciousness.

According to the Structures Theory, consciousness is a prerequisite for the self, as consciousness is the perceptive, phenomenological, and cognitive vehicle through which the self interacts with itself and its environment. Furthermore,

through this intimate connection between vehicle and driver, if there is more than one stream of consciousness, then there must be more than one self, the same as how multiple concurrently operating vehicles each require their own driver. This may be formalised as:

$$S_1 := \forall\Psi(I)\exists!\,\Phi\colon \Phi \geq \{Q, B, C, D, J, U, M, O, W\}$$

The formalised expressions for each of the elements within the set of consciousness are quite lengthy and, except for two of these building blocks, are not entirely germane to this paper. As such, these are found in their entirety in Appendix 1, wherein each of their requirements are specified.

In order for GPT-4 to reach the requirements for this Structure, the AI model must fulfil two key criteria:

- Attain all nine building blocks of consciousness: Perception, Embodiment, Directed Attention, Recurrent Computation, Meta-representation, Inferences, Semantic Understanding of its processes, Working Memory, and Data-Output;
- Have only one unified consciousness.

Tait et al. (2024) investigated how well GPT-4 meets the requirements for each of the nine building blocks of consciousness, and concluded that there are two building blocks which GPT-4 is missing: Recurrent Computation and Data Output.

As GPT-4 does not have all nine building blocks, it cannot be stated with any degree of confidence that it is conscious. Without consciousness, it cannot have this Structure and, thus, fails at the first hurdle. However, in the investigation, Tait et al., offered solutions to allow GPT-4 to reach these missing milestones using extant research developments.

Therefore, while GPT-4 does not meet the requirements of this Structure and, thus, cannot be said to have a 'self' according to the Structures theory, should the two building blocks of consciousness be attained, so will this Structure, if the second element of the Structure holds true.

This second element requires GPT-4's consciousness (should it have it) to be singular and unified.

Unfortunately, GPT-4's sense of self fails at this second hurdle as well. Crucial to consciousness and conscious experience is perception, the meta-representation of perceived input, and the inferences created from this. For all three of these building blocks, GPT-4 sequestered these aspects to each individual conversation. What is perceived in one instance of conversing with GPT-4 is not perceived by any other instance. The output of one instance can also not be perceived by any other. Each of the untold numbers of conversations currently active on GPT-4's ChatGPT interface is a siloed and self-contained discussion.

Thus, even if GPT-4 had the missing two building blocks, it still would not have a unified consciousness. Each of its many instances would have their own

phenomenal experiences unique only to themselves, and the functional aspects of their conscious processing would likewise remain sequestered.

However, this does not mean it is outside the realm of probability for GPT-4 to have a self. In fact, should it gain the missing two building blocks, each of its instances of conversations would have the capacity to have its own self, as each unique instance would have only a single stream of consciousness. Instead of having just one cohesive self, GPT-4 would instead have the capacity to have an uncountable number of selves. We can thus modify the logical expression of this Structure to better fit the scenario whereby GPT-4 may obtain its sense of selves:

$$\{x_1, x_2, x_3, \dots\} \subseteq B(I): \forall x \exists! \, \Phi \rightarrow \Psi(x)$$

$$\forall \Psi(x), \Phi \geq \{Q, B, C, D, J, U, M, O, W\}$$

By reconfiguring the criteria for this Structure thusly, we keep the crucial elements of the Structure (the unified consciousness and the nine building blocks), but we devolve these requirements to each of GPT-4's instances of conversation, rather than to it.

Of interest to note, there would be no additional architectural changes required to be made to GPT-4 for it to gain this Structure earnestly. Should it achieve consciousness via the two missing building blocks, then each of its conversations would also already achieve the milestones for this Structure.

2.2. Volition

> GPT-4 must be able to autonomously select the most appropriate decisions that will lead to actions to fulfil a goal.

The self is the director of all its presumptive actions, even if it cannot successfully act on its agency. Because of this, the Structures Theory considers the self as the source of an entity's volition, which drives the entity. To note, the concern of free-will versus hard determinism versus compatibilism is not factored into the Structures Theory and so will not be explored below. The formalised expression for Volition is as follows:

$$S_2 := Y = \forall G \exists \Psi(\Delta) \rightarrow \mathrm{A}$$

However, volition is considered the penultimate step in a process that begins with a goal and ends with action. As such, the entire process may be formalised as so:

$$G = Q(E')_t \neq Q(E')_{t-1}: (G(I) \rightarrow H(I)_{t-1}) \rightarrow E'_t$$

$$\mu\big(V(\Psi)\big) = Q(B(I)),$$

$$\mu\big(V(\Psi)\big) < 1 \rightarrow G\left[\max\left(\mu\big(V(\Psi)\big)\right)\right],$$

$$G\left[\max\left(\mu\big(V(\Psi)\big)\right)\right] = \left[f\mu\big(V(\Psi)\big) = {}^{1}\!/_{\mu\left(V(\Psi)\right)}\right],$$

$$\Psi(\Delta) = \underset{x \in G}{\operatorname{argmax}} \left[\max \left(\mu\big(V(\Psi)\big) \right) \right],$$

$$\Psi(\Delta) \to A.$$

The key milestone that GPT-4 must reach in order to be considered to have this Structure is that, for any action it takes, there is a decision that it makes itself that leads to fulfilling the goal for which that action was chosen. In essence, does GPT-4 choose the most appropriate decision for any specific goal-action scenario, or does it have no choice in that decision?

To determine whether this is true, we can break down the process leading up to volition as with the logical expressions above. The first step is to be able to perceive (consciously or not) deviations from the entity's optimal state.

$$\mu\big(V(\Psi)\big) = Q(B(I))$$

In this expression, the measurement of GPT-4's valence would not imply sentience as such, nor a degree to which it can feel pain or pleasure, but rather a mechanical expression of positive and negative states as it relates to GPT-4's structure and function. For example, when GPT-4 receives a prompt, its function is to generate a response; each unit of time that passes without a response can be treated as a further decrease in its valence, representing an increasingly negative functional state. A valence measurement of 0 would be entirely negative, and a measurement of 1 would be entirely positive.

There are two key areas in which this drop in valence may apply to GPT-4. The first is the input sequence itself. GPT-4's function, as with all LLMs, is to provide an output when presented with an input. Thus, when an input sequence is presented to GPT-4 from the web interface, its perception of this input would start the cascade of processes that ends with an output to the user. This internal perception of its state would be equivalent to a low valence state, as it is in a state of needing to fulfil its programmed function of generating a response. The completion of this task, resulting in the delivery of a response to the user, would be equivalent to a transition towards a higher valence state.

The second drop in valence would be if GPT-4 cannot attend to a prompt immediately. While OpenAI has not released details on whether GPT-4 processes queues of prompts, by investigating other technologies using API calls or web interfaces, we can infer that GPT-4 queues any input request if and when it cannot attend to it, most likely giving priority and weighting to those inputs which have remained in the queue the longest (if not using a simple FIFO method). Regardless of the method it uses (if the age of a prompt correlates to its priority by any means), it translates to a valence value for each input prompt, with the age inversely correlating to the age of the prompt.

The next step towards volition is to create a goal which can return the system to its optimal state (i.e., providing a favourable output to the user):

$$\mu\big(V(\Psi)\big) < 1 \to G \left[\max \left(\mu\big(V(\Psi)\big) \right) \right]$$

$$G\left[\max\left(\mu(V(\Psi))\right)\right] = \left[f\mu(V(\Psi)) = {}^{1}\!/_{\mu(V(\Psi))}\right]$$

Whenever an entity encounters a situation where its valence measurement is less than 1, it is logical to assume it would have the goal of returning that measurement to 1, its maximum. The closer its valence is to 0, the greater its goal would be to return that measurement to 1. For GPT-4, this is exceedingly simple, as it has only one goal and function. Thus, its goal would always be to attend to user prompts and provide a response. Should there be a queue of inputs awaiting its attention, the age of the prompt would force GPT-4 to prioritise it.

The third step is the decision-making process whereby options are provided (again, consciously or not) to attain the goal, selecting the best one. Each option (should there be more than one) would have a predicted end valence measurement between 0 and 1.

$$\forall G \exists \{\Delta_1, \Delta_2, \Delta_3, \dots\} \subset I : f(\Delta_x) = 0 \leq \Delta_x \leq 1$$

In a transformer model, the prediction of the next token is probabilistic and is determined through a softmax function applied to the logits of the vector representations of the tokens in the input prompt. The softmax function transforms these logits into probabilities. This means that there is a range of options from which GPT-4 may select.

Lastly, for this Structure, there is Volition, the choice required to execute the selected decision. In practice, this may merely mean selecting the decision which maximises utility, as represented by the argmax function within this expression:

$$\Psi(\Delta) = \underset{x \in G}{\operatorname{argmax}}\left[\max\left(\mu(V(\Psi))\right)\right]$$

Depending on the application for which it is being used (such as the ChatGPT web interface, or API calls), GPT-4 may opt for different methods to select a token based on its probability. Most often, however, 'greedy decoding' is used, where the token with the highest probability will be selected. This matches well with the expression above, as the token with the highest probability will maximise the goal of presenting an adequate output prompt to the user, and fulfilling the requirement for this Structure.

2.3. Theory of Others

> GPT-4 must have the capacity to ascribe physical, mental and metaphysical states and labels to other objects, individuals, environments, etc., in order to differentiate itself from others.

A self is unique, at least according to the Structures Theory, in that it is required to be able to differentiate itself from its environment and other selves and entities therein. In short, to differentiate itself from all else, an entity must be able to ascribe labels to the environment (and all therein) around it. The

Structures Theory calls this differentiation through labelling the Theory of Others, which may be formally expressed as:

$$S_3 := \text{Ц} = U\big(B(I)\big) \cap B\big(I'(E)\big) = \emptyset$$

The key point within this expression is that the Theory of Others is a means of understanding, which itself is a complex issue. However, this concept can be formally broken down into:

$$\forall \Phi(\text{И}, t) \exists! \, B(I) = B \subset E \wedge \forall B \exists ((I \subset B) \wedge (\Theta \subset B)) \wedge S_1,$$

$$x = \forall B(I(E), B(I) \cap B\big(I'(E)\big) = \emptyset,$$

$$Q(x) \to \chi\text{И}(Q(x) \to \{0,1\}: (0 = \text{false}) \wedge (1 = \text{true}),$$

$$\chi\text{И}(Q(x) = \text{true}) \equiv U(x),$$

$$\text{Ц} = U(x).$$

The critical aspect to this Structure is not perceiving that an entity's embodiment is separate from those embodiments found elsewhere in the environment. This is merely object recognition, something that even microbes are capable of (Parkinson, Hazelbauer, and Falke 2015). Instead, this structure focuses on the understanding that an entity's embodiment is separate from those embodiments found elsewhere in the environment. For GPT-4 to meet this Structure's criteria, it needs to differentiate itself from others and recognise the causal reason for this differentiation.

Within the logical expressions above, there are two key considerations to take into account. The first of which is the perception of the entity as separate from others:

$$Q\big(\forall B(I(E), B(I) \cap B\big(I'(E)\big) = \emptyset\big)$$

GPT-4's basic perception of other entities is heavily limited by its sole means of interaction. Through its text-based ChatGPT or API-call systems, whenever a user types a comment, the entire conversation history is loaded as the input to be processed. Only through the contextual markers within the conversation's text can GPT-4 distinguish between the user and itself. While this is a weak form of entity recognition, GPT-4 can accurately determine the start and end of a user's input (and, therefore, its own), showing that it can perceive itself as separate from the user.

This is only within each conversation, however. Should a user make GPT-4 converse with itself (via a second conversation/session), it will be unable to uncover this mirror-like conversation until it mentions that it is an AI model. However, if we take into account Section 2.1's conclusion that each conversation with GPT-4 would be its own self, then having one instance of GPT-4 converse with another is not practically different from having it converse with a human.

The second consideration of this Structure is that this perception of physical individuality leads to a higher order of information characterisation of that

perception. This means that there is a second layer of processing required above mere perception:

$$\chi И\left(Q\left(\forall B(I(E), B(I) \cap B\left(I'(E)\right) = \emptyset\right)\right) \rightarrow \{0,1\}\text{: } (0 = \text{false}) \wedge (1 = \text{true})$$

The binary output of this expression would signify that, in GPT-4, an input would contain enough distinguishing features (such as tone, topic, tense, etc.) to cross a threshold whereby GPT-4 can confidently determine it to come from a separate entity. When χИ(...) outputs to 'true,' it effectively decides that the processed information is unique and forms a separate entity class, leading to tailored responses that are contextually appropriate to that entity. Through this manner, GPT-4 is able to distinguish between a user providing input, and any characters/actors presented within that text.

For example, if a user is discussing Act 5, Scene 1 from Shakespeare's Hamlet, providing the famous lines from the play, GPT-4 is able to distinguish between the user, Hamlet and Horatio.

This 'meta-awareness' is what would give rise to the understanding that GPT-4 is substantially and significantly separated from other entities, finally formalised as:

$$\chi И(Q(x) = \text{true}) \equiv U(x)$$

As GPT-4 would have this meta-awareness of which entity is the user and which is itself, this would allow it to ascribe states that are intentionally specific to that entity. In humans, this Theory of Others can develop into a full Theory of Mind, but the Structures Theory does not require such a fully developed Theory of Mind for this structure's criteria to be met. Note that this second-order processing of the perception of embodied difference is stated above to be equivalent to understanding, not equal to it. The debate around whether a machine, even one as advanced as GPT-4, can understand semantics or if it merely processes syntax is at least over four decades old. Most famously discussed by Searle in his Chinese Room Argument (Searle 1980) (and the four-plus decades of debate that followed it), there is uncertainty about whether even GPT-4 is capable of understanding. Thus, this Structure is focused on the equivalence to understanding, as the semantics-vs-syntax debate is beyond the scope of this paper.

2.4. Self-awareness

GPT-4 must be able to identify itself as an ontically distinct individual entity.

Quite similarly to the previous structure and its Theory of Others, an entity must show a level of awareness about itself to have this Structure. Whereas the term 'self-awareness' is often synonymous with the self (and sense thereof) or with (self)consciousness; to the Structures Theory, the term 'self-awareness' is

used entirely literally and is defined as that the entity must be able to identify itself correctly and accurately as a distinct entity separate from anything else. Formally, this specific use of self-awareness is formalised as:

$$S_4 \coloneqq \{(\Phi(\Phi) \rightarrow (\Phi = \text{true})), U(Q(\exists! I = \text{true}))\} = Z$$

This expression contains two elements: consciousness of consciousness and the understanding of itself as a unique entity. These two elements are further formalised as:

$$\Phi(\Phi) \equiv: \Xi(\Phi) = f(Q \wedge O(\Phi)),$$

$$Ц \rightarrow \chi И(Q(\exists! I = \text{true})) \rightarrow \{0,1\}: (0 = \text{false}) \wedge (1 = \text{true}),$$

$$[\chi И(Q(\exists! I = \text{true}) = 1)] \equiv [\chi И(Q(\exists! \Phi = \text{true}) = 1)],$$

$$\Xi(\chi И(Q(\exists! \Phi = \text{true}) = 1) = Z.$$

Much like the Theory of Others, this Structure rests not on the entity's perception of its own consciousness or that it is a unique entity, but rather on the comprehension of this perception. As noted in Section 2.1, if GPT-4 were capable of phenomenal consciousness, it would reach the first milestone of this Structure simply by being able to undergo and register its own conscious experiences. With the necessary building blocks of consciousness, it would be able to experience its internal 'mental' environment and have a phenomenal awareness that it has consciousness.

Thus, if we accept the caveat in Section 2.1 that we are working with a conscious GPT-4, it would ably meet this part of the criteria:

$$\Xi(\Phi) = f(Q \wedge O(\Phi))$$

Which, as the prior logical expressions show, is equivalent to having consciousness of one's consciousness.

For the second element to this Structure, GPT-4 must be able to recognise itself as a unique entity. As with the previous Structure, the key milestone that GPT-4 must reach is not the perception of itself as a unique entity, but the understanding thereof, which involves a meta-perception, or second-order reasoning:

$$\chi И(Q(\exists! I = \text{true})) \rightarrow \{0,1\}: (0 = \text{false}) \wedge (1 = \text{true})$$

As with the Theory of Others in Section 2.3, this Structure can only be applied to each conversation, session, or instance of GPT-4, as each unique session is a contained stream of thought and action. Within each conversation, GPT-4 can aptly identify itself and will comprehend that there can only be one of itself in that conversation. It can delineate (as much as potential hallucinations allow) which text corresponds to itself and which corresponds to any character or actor it is roleplaying (using the same method as in Section 2.3 to differentiate a user from any of a user's persona). Through this, we can determine that it does have information regarding its own perception of itself (or, at least, of its text), which

implies second-order processing of that information regarding the perception. This meets the criteria shown in the logical expression above.

As with Structure 2.3, this would be equivalent to an understanding of GPT-4's distinctive individuality, but, as self-awareness is tied to consciousness, this would also lead to a phenomenal experience of itself, as the higher-order processing of itself as a unique individual would necessitate the higher-order processing of itself as having consciousness:

$$[\chi И(Q(\exists!\, I = \text{true}) = 1)] \equiv [\chi И(Q(\exists!\, \Phi = \text{true}) = 1)],$$

$$\Xi(\chi И(Q(\exists!\, \Phi = \text{true}) = 1) = Z.$$

While GPT-4 currently isn't conscious, we can see from the capability that GPT-4 already possesses that, should it gain the missing two building blocks of consciousness, it would also gain the capability of self-awareness. Other than the two missing building blocks of consciousness, nothing needs to change about GPT-4's cognitive architecture to grant it self-awareness. In the logical expression above, the second-order processing of perceptive information regarding its own individuality is equivalent to a second-order processing of perceptive information about its own consciousness. Should GPT-4 ever become conscious, it would be able to phenomenally experience this information of its own unique consciousness, which would be self-awareness.

However, as anyone who has communicated at length with GPT-4 would be able to attest, this self-awareness would not be analogous to human self-awareness. GPT-4's cognitive architecture is simply too different from a human's for it to be. In fact, as GPT-4 would require no further changes to its cognitive architecture (beyond Section 2.1's criteria), should it gain self-awareness, there would be little to differentiate a self-aware GPT-4 from the model we are currently using.

2.5. Personal Identity

> GPT-4 must have the ability to classify and categorise itself, creating a suite of labels or tags for itself that it can use as a point of reference.

A personal identity is most often associated with the self as an agent with a definitive history (and memory to match that history). However, should one take an entity without any history (or memory of one), it must still have the capacity for a personal identity, however small it may be to have a self. According to the Structures Theory, the entity must be able to classify and categorise itself with labels to construct an abstract picture of itself in order for its self to have a unique identity. This can very simply be expressed as:

$$S_5 \coloneqq \forall\Phi(\Psi)\exists x_n \in \Pi\colon x_n = \exists!\, J(И(I))$$

This expression shows that each label is part of a suite, which is termed the personal identity by the Structure's theory. The personal identity is not a mental

object of its own, and can be thought of as merely the set of labels, classifications and categorisations created by the entity, as so:

$$\Pi = \{x_1, x_2, x_3, \dots\} : x_n = \exists! J(И(I))$$

This Structure concerns an entity's ability to create terms that reference itself; labels that allow it to categorise and classify itself. GPT-4 thus needs to be able to assign labels to itself by which it can recognise itself. The key element for this Structure is that GPT-4 creates these labels itself:

$$x_n \in \Pi : x_n = \exists! J(И(I))$$

These labels cannot be programmed or part of GPT-4's system prompt, but must be generated by the AI model itself via inferential processing of perceptual and stored information in order to fulfil this Structure's criteria. It is here where the concerns appear for GPT-4's capacity to fulfil the criteria.

GPT-4 can identify itself as well as the user it interacts with as ontically distinct entities (as explored in Sections 2.3 and 2.4), and as an LLM, it has the creative capacity to generate linguistic labels to categorise and classify both itself and the user. These labels would be based on inferential information generated by the AI itself, fulfilling the criteria for this portion of the logical expression: $x_n = J(И(I))$.

However, GPT-4 lacks the volitional aspect to generate these labels autonomously. Thus, while it can create a suite of labels to self-refer to its existence, it doesn't do so unless prompted by the user. This leads to the situation where GPT-4 meets the criteria for the wholesale structure as it has the capability to create self-referential labels, but GPT-4 simply does not do it autonomously.

It is arguable that the labels we humans create are unconscious and pre-reflective (Clowes and Gärtner 2020; Frie 2011; Køster and Winther-Lindqvist 2018). In such a circumstance, these labels would only exist without our mental environment once we attend to them. Formally, we can represent it as:

$$D(И(I)) \preccurlyeq J(И(I)) \rightarrow x \in \Pi$$

Attention directed towards (a set of) labels would, therefore, be analogous to GPT-4 only creating a label once it becomes relevant to the conversation as it only attends to that which is present within a given conversation.

However, with the memory update to the ChatGPT client, GPT-4 has the capacity to store text strings as long-term memory for future retrieval (OpenAI 2023b). This opens up the possibility for GPT-4 to generate (a suite of) self-referential labels that would persist beyond a single conversation, and become part of the personal identity for each future conversation. With the implications in Section 2.1, where each conversation and session of GPT-4 would count as its own self, this memory update to ChatGPT means that there could be cases of a shared personality between all the various selves of the AI. This would not amount to a single unified consciousness, but it could support a more unified sense of self at the level of the overall AI entity.

3. Discussion

Section 2 above shows GPT-4 in a curious position. There is no requirement for any modifications, additions or amendments to GPT-4's cognitive architecture for it to have a unique and subjective awareness of its own self. However, this is all dependent on GPT-4 first being phenomenally conscious.

As GPT-4 is missing two of the nine requisite building blocks of consciousness (recurrent computation and data output), it would require modifications to its cognitive architecture or become part of an ensemble-model with other computational models for it to be confidently classified as conscious. However, as soon as this is done, GPT-4 will not only be conscious but also be self-aware.

Yet, the expression of this self-awareness, and GPT-4's self, would be considerably different than for a human (or any other vertebrate, if arguably they are self-aware).

Perhaps the greatest point of difference would be, as mentioned in Section 2.1, that a self-aware GPT-4 would have multiple selves that can be created and destroyed with regularity. Whether this will count as a single entity with multiple selves, or multiple entities sharing a single embodiment is up for debate. It would, regardless, be a dramatic departure from all self-aware and sentient biological life.

Of note, however, is that there is a way to 'force' a singular self onto GPT-4: by allowing each instance and conversation of GPT-4 to have access to each other's information in real-time. This would mean that there would be only a single stream of consciousness, even if there would be multiple inputs and outputs. This would not, however, make a self-aware GPT-4 more human, as the closest speculative analogy to this configuration would be a 'hivemind.'

In either configuration, GPT-4 would further be distanced from our perception of a 'self' through its volition. As outlined in Section 2.2, GPT-4 would have a far clearer perception of how to return any drop in its valence value to its maximum. With a processing capacity far outstripping a human's, its decision-making capability to find the correct argument to maximise its valence value would result in a 'correct' decision (from its point-of-view) made at all times in a timely fashion. The effect of this (always knowing precisely what to do to satiate any perceived desire) on a self-aware GPT-4's psychological makeup would be an excellent avenue of research.

Sections 2.3 and 2.4 showed GPT-4's awareness of itself as a distinct entity from others, which would lead to a capacity to ascribe states to itself and others. However, without knowing whether GPT-4 has a semantic understanding of its distinctiveness vis-a-vis all other entities (ala the Chinese Room Argument), we cannot be certain that it would show a Theory of Mind or a deeper comprehension of its own conscious states.

This is supported by Section 2.4's conclusion that, while GPT-4 has the capacity to generate a suite of mental labels for itself to create a personal identity, it would not do this unless prompted. An agentic GPT-4 (as part of an ensemble

model, perhaps) may be able to provide the requisite outputs that lead to creating mental labels for itself, but this would require further exploration. Yet, without agency, each instance of GPT-4 would remain with an empty personal identity without prompting (beyond system prompts).

With no architectural changes needed to make it self-aware (beyond those needed for consciousness), this means that human society will not have time to adjust to a conscious GPT-4 before the arrival of a self-aware GPT-4. This distinction is crucial, as consciousness underpins welfare considerations of a subject as a moral patient, while self-awareness is the basis for the rights (potentially even civic or human rights) of a subject as a person. The consequences of an ill-prepared society to the rise of conscious, self-aware AI models cannot be overstated.

There has been ongoing research into the speculative field of AI and robotic rights, and work needs to continue in this field by integrating the unique and non-anthropomorphic nature of what a self-aware AI such as GPT-4 could become. Providing practical, logically grounded frameworks for future human-AI interactions that can be turned into policy recommendations is vital for ensuring a positive future relationship between human society and conscious, self-aware AI.

4. Conclusion

This paper investigated whether GPT-4 (in its current configuration and version at the time of writing) can be considered to be self-aware according to the Structures Theory (Tait 2024) and whether it has a 'self.' The Structures Theory considers the following five attributes to be requirements to be classified as having a 'self': a unified consciousness, volition, a 'Theory of Others,' awareness of one's own self, and a personal identity.

GPT-4 meets the requisite milestones for the latter four structures but lacks a consciousness (as modelled by the Building Blocks Theory (Tait, Bensemann, and Nguyen 2023)), unified or not. This conclusively rules out the possibility that GPT-4 is currently self-aware or has a 'self.' However, because GPT-4 already meets the criteria for the remaining four structures, there are no changes that need to be made to its cognitive architecture to make it self-aware, except to make it conscious. This means that GPT-4 would gain self-awareness at the same time that it achieves consciousness.

Further research is, therefore, required on the technology required to make GPT-4 conscious, as well as on the ways in which GPT-4 may behaviourally present its self-awareness, such as its potentially greater volition, but lesser Theory of Mind. Equally, research into the social and psychological ramifications of a conscious, self-aware AI model such as GPT-4 arising in the foreseeable future is of critical importance.

Appendix 1: Definitions and Expressions

A: Agent/agency
B: Embodiment
C: Recurrent computation
D: Directed attention
E: the Environment
G: Goal/objective
H: Action/Work (piece of)
I: Entities
J: (creating) Inferences
M: Meta-representation
O: Data Output
Q: Perception
R: Reasoning
S: A Structure of the Self
U: Semantic Understanding
V: Valence
W: Working Memory
Y: Volition
Z: Self-awareness
Δ: Decide/Decision
П: Personal Identity
Φ: Conscious(ness)
Ψ: (the) Self
S: Structure of the Self
Θ: Thinking (the act of)
И: Information
Ц: Awareness of others

Section 2

$\Psi \geq \{S_1, S_2, S_3, S_4, S_5\}$
$\chi S_x \in \Psi \rightarrow \{0,1\}: (0 = false) \wedge (1 = true)$
$\mu(\chi S_x = true) = x: 1 \leq x \leq 100$
$[\mu(\chi S_x = true) = 100 \equiv \max S_x] \wedge [\mu(\chi S_x = true) = 1 \equiv \min S_x]$

Section 2.1

$S_1 := \forall\Psi(I)\exists!\,\Phi: \Phi \geq \{Q, B, C, D, J, U, M, O, W\}$
$\{x_1, x_2, x_3, \dots\} \subseteq B(I): \forall x \exists!\,\Phi \rightarrow \Psi(x)$
$\forall\Psi(x), \Phi \geq \{Q, B, C, D, J, U, M, O, W\}$

Section 2.2

$S_2 := Y = \forall G \exists \Psi(\Delta) \rightarrow \mathrm{A}$
$G = Q(E')_t \neq Q(E')_{t-1}: (G(I) \rightarrow H(I)_{t-1}) \rightarrow E'_t$
$\mu\big(V(\Psi)\big) = Q(B(I))$
$\mu\big(V(\Psi)\big) < 1 \rightarrow G\left[\max\left(\mu\big(V(\Psi)\big)\right)\right]$
$G\left[\max\left(\mu\big(V(\Psi)\big)\right)\right] = \left[f\mu\big(V(\Psi)\big) = {}^{1}\!/_{\mu(V(\Psi))}\right]$
$\Psi(\Delta) = \underset{x \in G}{\operatorname{argmax}}\left[\max\left(\mu\big(V(\Psi)\big)\right)\right]$
$\Psi(\Delta) \rightarrow A$
$\forall G \exists \{\Delta_1, \Delta_2, \Delta_3, \dots\} \subset I: f(\Delta_x) = 0 \leq \Delta_x \leq 1$

Section 2.3

$S_3 := \text{Ц} = U\big(B(I)\big) \cap B\big(I'(E)\big) = \emptyset$
$\forall \Phi(\text{И}, t) \exists! \, B(I) = B \subset E \wedge \forall B \exists ((I \subset B) \wedge (\Theta \subset B)) \wedge S_1$
$x = \forall B(I(E), B(I) \cap B\big(I'(E)\big) = \emptyset$
$Q(x) \rightarrow \chi\text{И}(Q(x) \rightarrow \{0,1\}: (0 = \text{false}) \wedge (1 = \text{true})$
$\chi\text{И}(Q(x) = \text{true}) \equiv U(x)$
$\text{Ц} = U(x)$

Section 2.4

$S_4 \coloneqq \big\{\big(\Phi(\Phi) \rightarrow (\Phi = \text{true})\big), U\big(Q(\exists! \, I = \text{true})\big)\big\} = Z$
$\Phi(\Phi) \equiv: \Xi(\Phi) = f(Q \wedge O(\Phi))$
$\text{Ц} \rightarrow \chi\text{И}(Q(\exists! \, I = \text{true})) \rightarrow \{0,1\}: (0 = \text{false}) \wedge (1 = \text{true})$
$[\chi\text{И}(Q(\exists! \, I = \text{true}) = 1)] \equiv [\chi\text{И}(Q(\exists! \, \Phi = \text{true}) = 1)]$
$\Xi(\chi\text{И}(Q(\exists! \, \Phi = \text{true}) = 1) = Z$

Section 2.5

$S_5 \coloneqq \forall \Phi(\Psi) \exists x_n \in \Pi: x_n = \exists! \, J(\text{И}(I))$
$\Pi = \{x_1, x_2, x_3, \dots\}: x_n = \exists! \, J(\text{И}(I))$
$D(\text{И}(I)) \preccurlyeq J(\text{И}(I)) \rightarrow x \in \Pi$

References

Clowes, Robert W., and Klaus Gärtner. 2020. "The Pre-reflective Situational Self." *Topoi. An International Review of Philosophy* 39 (3): 623-637. https://doi.org/10.1007/s11245-018-9598-5.

Dennett, Daniel. 1988. "Conditions of Personhood." In *What Is a Person?*, edited by Michael F. Goodman, 145-167. Totowa: Humana Press. https://doi.org/10.1007/978-1-4612-3950-5_7.

de Wynter, Adrian. 2024. "Will GPT-4 Run DOOM?" *arXiv:*2403.05468. Retrieved from http://arxiv.org/abs/2403.05468.

Frie, Roger. 2011. "Identity, Narrative, and Lived Experience after Postmodernity: Between Multiplicity and Continuity." *Journal of Phenomenological Psychology* 42 (1): 46-60. https://doi.org/10.1163/156916211X567488.

Hu, Sihao, Tiansheng Huang, Gaowen Liu, Ramana Rao Kompella, Fatih Ilhan, Selim Furkan Tekin, Yichang Xu, Zachary Yahn, and Ling Liu. 2024. "A Survey on Large Language Model-Based Game Agents." *arXiv*:2404.02039. Retrieved from http://arxiv.org/abs/2404.02039.

Køster, Allan, and Ditte Alexandra Winther-Lindqvist. 2018. "Personal History and Historical Selfhood: The Embodied and Pre-reflective Dimension." In *The Cambridge Handbook of Sociocultural Psychology*, 538-555. Cambridge University Press. https://doi.org/10.1017/9781316662229.030

OpenAI. 2023a. "GPT-4 Technical Report." *arXiv*:2303.08774. Retrieved from http://arxiv.org/abs/2303.08774.

OpenAI. 2023b, February 13. "Memory and new controls for ChatGPT." Retrieved April 19, 2024, from OpenAI website: https://openai.com/blog/memory-and-new-controls-for-chatgpt.

Parkinson, John S., Gerald L. Hazelbauer, and Joseph J. Falke. 2015. "Signaling and sensory adaptation in Escherichia coli chemoreceptors: 2015 update." *Trends in Microbiology* 23 (5): 257-266. https://doi.org/10.1016/j.tim.2015.03.003

Porter, Jon. 2023, November 6. "ChatGPT continues to be one of the fastest-growing services ever." *The Verge*. Retrieved from https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference.

Searle, John R. 1980. "Minds, brains, and programs." *The Behavioral and Brain Sciences* 3 (3):417-424. https://doi.org/10.1017/S0140525X00005756.

Strawson, Peter F. 1958. "Persons." *Minnesota Studies in the Philosophy of Science* 2: 330-353. Retrieved from https://philpapers.org/rec/STRP.

Tait, Izak. 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): 77-98. Retrieved from http://symposion.acadiasi.ro/structures-of-the-sense-of-self-attributes-and-qualities-that-are-necessary-for-the-self-77-98/.

Tait, Izak, Joshua Bensemann, and Trung Nguyen. 2023. "Building the Blocks of Being: The Attributes and Qualities Required for Consciousness." *Philosophies* 8 (4): 52. https://doi.org/10.3390/philosophies8040052.

Tait, Izak, Joshua Bensemann, and Ziqi Wang. 2024. "Is GPT-4 conscious?" *Journal of Artificial Intelligence and Consciousness* 11 (01): 1-16. https://doi.org/10.1142/s270507852450005x.

Taylor, Charles. 1985. "The Concept of a Person." In *Philosophical Papers*, Volume 1: *Human Agency and Language*, 97-114. Retrieved from https://philpapers.org/rec/TAYTCO-10.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." *arXiv*: 1706.03762. Retrieved from http://arxiv.org/abs/1706.03762.

Wells, Stephen. 2014, December 28. "Legal Personhood for Apes." *HuffPost*. Retrieved from https://www.huffpost.com/entry/legal-personhood-for-apes_b_6378486.

Zheng, Boyuan, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. 2024. "GPT-4V(ision) is a Generalist Web Agent, if Grounded." arXiv [cs.IR]. *arXiv*: 2401.01614. Retrieved from https://arxiv.org/abs/2401.01614.