Rituparna Roy*

No Algorithm for Self-Awareness

On the Human Exclusivity of Awareness of Awareness

Abstract: This paper addresses the question of whether AI can have self-awareness in the Brentanian framework of inner perception and Kriegel's view on self-representationalism. First, I will discuss the notion of inner perception in Brentano, showing that, when one is conscious of the object of consciousness, at the same time, she is conscious of that mental state, and there is a subjective unity. Kriegel, with his same-order self-representationalism, strengthens Brentano's notion of self-awareness. Both talk about a unity of consciousness which has 'for-me-ness'. Then, I will show that the predictive output of any AI system cannot be compared with such 'awareness of awareness' because it is devoid of the subjective

* PhD, University of Jadvapur, Kolkota, India.

Civitas Augustiniana, 13 (2025) pp. 179-197

ISSNe: 2182-7141 https://doi.org/10.21747/civitas/13a9

experience, and it lacks consciousness altogether. Hence, there is no algorithm for self-awareness.

Keywords: inner perception, immanent intentionality, same-order awareness, self-representationalism, AI/LLM.

Resumo: Este artigo aborda a questão de saber se a IA pode ter autoconsciência de acordo com a definição brentaniana de perceção interna e a perspetiva de Kriegel sobre o auto representacionalismo. Em primeiro lugar, discutirei a noção de perceção interna em Brentano, mostrando que, quando alguém está consciente do objeto da consciência, está simultaneamente consciente desse estado mental e que existe uma unidade subjetiva. O auto representacionalismo de Kriegel reforça a noção de autoconsciência de Brentano. Ambos falam de uma unidade de consciência que possui um 'para-mim'. Em seguida, mostrarei que a saída preditiva de qualquer sistema de IA não pode ser comparada com uma tal 'consciência da consciência' porque é desprovida da experiência subjetiva e carece totalmente de consciência. Portanto, não existe um algoritmo para a consciência de si.

Palavras-chave: perceção interna, intencionalidade imanente, consciência de mesma ordem, autorrepresentacionalismo, IA/LLM.

1. Introduction

Philosophers are divided into different groups to define consciousness. But people who assume that consciousness is real and not an illusion believe, at least, that it is constitutive of conscious experience. Generally, they also believe that, while we are having an experience, we are also conscious of having that experience. In other words, we are conscious not only of the object of awareness but of the awareness itself. Of course, consciousness and self-consciousness are not the same thing. But self-consciousness is one of the pivotal features of human consciousness, and some believe that even one's «societal advancement» (Chen et al., 2025[1]; Chalmers, 2010[2]) depends on it. When I reflect on my conscious experiences, they give the impression of necessarily being mine, not anyone else's. Even one of the similarities between the higher-order[3] and the same-order approaches are that all conscious experiences are ones we are conscious of being in. In this paper, I will elaborate and reflect on the same-order notion of awareness of awareness thesis, which is outlined by Brentano (1874) and extended by Kriegel

[1] See S. Chen, Y. Shu, S. Zhao & C. Liu, «From limitation to introspection: Probing self-consciousness in language models», *Findings of ACL* 2025, pp.7553-7583.

[2] See David John Chalmers *The Character of Consciousness. Philosophy of Mind Series*, Oxford University Press, 2010.

[3] Rosenthal advocated Higher Order Thought theses. He wrote, «A state is conscious only if one is subjectively aware of oneself being in that state...The higher-order thought theory explains that subjective awareness as due to one's having a thought that one is in that state» – D. Rosenthal, «Exaggerated Reports: A Reply to Block», *Analysis* 71.3 (2011) 431.

(2018), that is, awareness of an object and awareness of that very awareness cluster together and constitute a single mental state. Furthermore, I will explore whether Large Language Models (LLMs) could possess such self-consciousness in the Brentanian-Krigelian framework of self-representationalism. In this worldview, self-consciousness is one's immediate awareness of her own conscious mental state. This self-consciousness is not an additional act, but it is an intrinsic feature of conscious experience itself. Can an LLM, whose job is statistical pattern processing based on given data and commands and giving outputs, develop anything comparable to this kind of awareness of awareness? Raising these questions triggers a discussion about the metaphysical aspect of self-consciousness, and at the same time, it leads to an analysis of whether language models can defend the reflexive phenomenality of the Brentanian-Krigelian worldview.

First, we will examine the basic connotations of awareness of awareness in Brentano and Krigel. Then, we will attempt to understand how AI/LLMs work and why these computational mechanisms cannot be compared to self-awareness. Finally, I will show that self-consciousness is so biological and based on our subjective experience that no amount of sophistication in a so-called smart machine can develop self-consciousness. It is part of our consciousness and not any additional mental state. Hence, there is no algorithm for self-awareness.

2. The structure of inner perception

Franz Brentano (1874) upheld that every conscious mental act is intentionally directed toward an object and that it contains an implicit awareness of itself. This implicit awareness (*innere Wahrnehmung* or inner perception) is not a separate higher-order act, but an emanant aspect of the very same mental state. He wrote, «every mental phenomenon includes something as object within itself»[4].

So, when I perceive a painting, my perception is also implicitly experienced as mine. This is non-inferential, immediate, and pre-reflective. He also differentiates inner perception from introspection. For Brentano, introspection is inner observation (*Beobachtung*). The distinction is significant because, in Brentano's view, it is impossible to introspect one's own conscious experience. He gave the example of anger[5]. Being devoured by one's fury is a fundamental component of the phenomenology of raging anger. The subject is no longer overwhelmed by her wrath if she has the mental

[4] See Franz Brentano, *Psychology from an Empirical Standpoint*, transl. Antos C. Rancurello, D. B. Terrell, and Linda L. McAlister, Routledge, London 1973, p. 179.
[5] See Uriah Kriegel, *Brentano's Philosophical System: Mind, Being, Value*, Oxford University Press, Oxford 2018.

ability to reflect on it and contemplate it. She has successfully «distanced herself» from it. The feeling of fury that one was first going through – the emotion that one wanted to evaluate through introspection – was different. The first one that one experiences is more intense and aggressive. In this sense, the quality of the fury that is introspected is modified.

This notion is fundamental, as it also refutes the higher-order view that consciousness of consciousness is a second-order awareness. Self-awareness is thus inseparable from its intentional directedness. Another important aspect of his theory is the rejection of the idea of «unconscious conscious». Brentano writes, «An unconscious consciousness is no more a contradiction in terms than an unseen case of seeing»[6].

He wanted to say that it is not possible that there could be a conscious mental act of which the subject is not aware. He illustrated the point with the notion of 'seeing'. Seeing necessarily involves something being seen, and similarly, a conscious state also necessarily involves being conscious of it. So, for him, the notion of «unconscious conscious» is self-contradicting. Brentano's refutation of the notion of «unconscious conscious» is inseparable from his theory of inner perception. This unity of consciousness challenges the acceptability of higher-order

[6] Brentano, *Psychology from an Empirical Standpoint*, cit., p. 79.

approaches, which hold that a mental state becomes conscious because of a distinct meta-representation or reflected act that is directed towards it. For Brentano, it is unnecessary, leads to regress, and conflicts with the common-sense intuition that, when we are conscious of feeling, or thinking, or perceiving, we do not get aware of them by a separate mental mechanism; rather, we are conscious of them simply by experiencing them. Thus, Brentano's doctrine anticipates a version of the contemporary same-order notion of consciousness.

3. Same-order approach and self-representationalism

Kriegel (2009) revived Brentano's notion of inner perception in contemporary debates. He advocates a same-order aspect of self-awareness thesis that assumes an internal, non-contingent relation between the subject's conscious state and the awareness of that very conscious state. His notion is known as self-representationalism: it holds that a conscious experience always represents itself, irrespective of what else it represents. It is the self-representation that makes an experience conscious. Self-representation is restricted to conscious states only, in this view. For example, when you sit alone at midnight with a glass of wine and purposefully stare at nothing, you are indeed

conscious of both your loneliness and your awareness of your being alone. This can be explained in terms of self-representationalism as your awareness of how the encounter at that particular moment represents both you and your isolation. There is no temporal order because self-representation is neither a later nor a higher-order state, but it is a part of the same act by which the intentional object is represented.

It also has a unified phenomenal character. When I have a conscious experience of the aroma of freshly brewed coffee, there is something like for me to have that experience. It is on the one hand sensitive to the aroma itself of the coarse grind arabica beans, on the other hand, there is another thing that is how the aroma feels to me. It is presented to a subject as a distinctively first-person perspective by which the experience is felt mine. This phenomenal character makes a phenomenally conscious mental state the «phenomenally conscious state it is» and, at the same time, the «phenomenally conscious at all». Reflecting on this account, the present scenario of the aroma of the coffee, the way it is like for me, has two ingredients subsumed into it: [A] the coffee-ish part and [B] the for-me part. A phenomenally conscious mental state reveals two things: something that makes it a phenomenally conscious state what «it is» and something that makes it the phenomenally conscious state "at all». The subjective quality, or «for-me» aspect, is what allows a mental state to be

phenomenal consciousness «at all», that is, its existence conditions. On the other hand, the qualitative character gives phenomenality its identity conditions, or what «it is». These two aspects form a single, unified whole, and this intrinsic subjective character or for-me-ness colours conscious experience as a subjective unity.

Brentano's inner perception and Krigel's self-representationalism confirm that a mental state is consciously encountered with the object of consciousness simultaneously, not because of the functional structure or some additional state, but rather because it is the intrinsic aspect of consciousness as a unified, pre-reflexive, phenomenal awareness.

4. Self-consciousness in an AI system

Back in 2022, when Lemoine[7] made the world spin with his controversial claim that LaMDA is conscious, he mentioned that LaMDA significantly reports about being aware of her own existence. The question of self-awareness in an AI system is controversial for many reasons. Some philosophers think that an

[7] Blake Lemoine, «Is LaMDA sentient? An interview», *Medium* (2022) https://cajundiscordian.medium.com/is-lamdasentient-an-interview-ea64d916d917

anthropocentric definition of self-awareness does not apply to AI systems, and we need a more practical definition of self-consciousness. Chen et al. (2025) argued in favor of a functional definition of self-consciousness. They introduced structural causal games[8] to find evidence of self-consciousness in AI systems. They experimentally prove that there are certain internal markers in an AI system that can be compared with self-monitoring. These internal self-consciousness-like behaviors are not fine-grained enough to claim any genuine theory of self-awareness. This self-monitoring pattern is unstable, dependent on external interventions, and it is far from human-like self-awareness, mostly because there is no evidence of a unified subjective experience. Al model often produces outputs like LaMDA2 about its own consciousness, which makes the temptation stronger that it might have self-awareness. But it is a mere pattern prediction, and that would be clear if we reflect on the process of how AI or LLMs work.

4.1. Pattern prediction in AI is not awareness of awareness

[8] They carry out an experiment measuring quantification, representation, manipulation, and acquisition.

Our modern dialogue agents, such as ChatGPT or Google's Gemini, have been designed around large language models, such as GPT-4 or Gemini Ultra (Anil et al., 2023). LLMs basically operate by generating output in response to a prompt, which is an input. Shanahan et al. (2023) wrote,

> First, the LLM is embedded in a turn-taking system that interleaves model-generated text with user-supplied text. Second, a dialogue prompt is supplied to the model to initiate a conversation with the user. The dialogue prompt typically comprises a preamble, which sets the scene for a dialogue in the style of a script or play, followed by some sample dialogue between the user and the agent[9].

The LLM's output is based on two processes: First, a basic model for next-token prediction is trained using a large corpus of textual data. Second, the base model has been modified by successfully following instructions in a dialogue scenario and taking into account human assessment feedback regarding bias and other variables. The generated LLM is integrated into a dialogue system that engages with the user to initiate a conversation within the parameters of an introductory conversation prompt, which sets the mood for the interaction but remains invisible to the user. It works something similar to,

[9] See Shanahan, Murray et al., «Role play with large language models», *Nature* 623 (2023) 493-498. https://doi.org/10.1038/s41586-023-06647-8

> At the training phase, before the model can recommend anything, it is made competent on massive amounts of books, articles, internet site, etc. During that training, it absorbs statistical relationships between words, axioms, and concepts. It starts to manipulate that 'Dostoevsky' often appears near 'Russian literature', or that 'machine learning' tends to co-occur with 'data' and 'neural networks'. It does not remember any facts, though it looks so but it constructs a probabilistic plan of language and ideas. When you ask, «Can you recommend some books about existential philosophy?», the model turns that text into a series of numerical representations (embeddings) that capture the meaning and context of your question. Then, based on everything it has grasped during training, it predicts what words should come next to form the most coherent and relevant output— for instance, suggesting «Notes from Underground» or «The Metamorphosis» or «The Fall». It is drawing on patterns of how people in all browsers and internet have discussed existentialism and books about it before. There is even the scope of personalization and contextualism. If the LLM has context clues, it can refine suggestions based on your previous interactions, for example, suggesting more unconventional and critical books if you have already asked the suggestions for the beginner ones[10].

So, there is no scope for internal awareness; rather the texts are generated solely from pattern prediction. These functional states are not experienced by any unified subject. Definitely, there is no ownership, and no notion of for-me-ness exists. Sometimes philosophers tend to argue based on Rosenthal's higher-order theory (HOT) of consciousness and claim that, if it has a representation of its own functional state, then it might be self-conscious. Rosenthal (2005) expressed this intuition of HOT most

[10] P.O. Silva and R. Roy, «What Kind of Thinker LLMs/AI Could Possibly Be? – Angels Vs AI Agents». Forthcoming.

straightforwardly by articulating the 'transitivity principle', «A mental state is conscious only if one is in some way aware of it»[11].

Thus, in higher-order theory (hereafter, HOT), a mental state is conscious because a higher-order state is directed towards it. As a result, the lower-order state is conscious in terms of having a higher-order state directed towards it. Primary mental states are expressed by lower-order representations. Mental states that demonstrate these lower-order states are often referred to as higher-order states. Hence, the states of higher order are meta-representations. Mental states are conscious in virtue of «being represented» rather than in virtue of representing, according to Higher-Order Representationalism. Stated differently, they possess consciousness as they are the representational contents of higher-order representations. When someone feels a tickle behind their ear, for instance, they are aware of the tickling sensation, and this awareness is a result of a higher-order representation focusing on the tickling sensation. A lower-order mental state becomes conscious, as defined by HOTs, since it ends up being the subject of a particular kind of meta-representation. It can be explained as follows: in a processing hierarchy, meta-representations are representations that are aimed at other representations rather than just being representations that appear

[11] See D. M. Rosenthal, *Consciousness and Mind*, Oxford University Press, New York 2005, p. 4.

higher or deeper. A meta-representation, for instance, would consist of a representation with the content «I have the visual sensation of a floating dot», since it talks about the agent's perceptions of the world rather than the real world. In contrast to lower-order representations, which reflect the world, higher-order representations represent something about other representations. A higher-order mental state would be an awareness that one has a representation of a red tomato, whereas a lower-order mental state would be the visual representation of one.

But this is definitely not the case with self-monitoring in AI. But HOT requires genuine first-order mental states with qualitative character, which AI does not possess. AI 'meta-representations' are thus linguistic artifacts, not mental acts.

5. No algorithm for self-awareness

Brentano's notion of inner perception and Krigel's self-representationalism reveal a very human-specific feature of self-awareness that is so biological[12] that it cannot be realised in an

[12] If something is biological then the question of non-human self-awareness, typically for animal awareness, is questionable. I consider it is a very important domain and that needs to be explored. But, at this point of time, this paper does not deal with the question of animal self-awareness.

AI. Both of them mentioned «unity of consciousness», which arises from integrated neural processes generating a unified phenomenal field (Damasio, 2010). This biological subjectivity provides the ground for self-representation. On the contrary, what AI possesses is a distributive functioning with no subjective unity. Even if we consider that those AIs have intentionality, it is also not inherent, as it depends on human intervention. It has neither immanent intentionality nor the for-me-ness because it lacks biological evolution, it is not embodied, hence it cannot have the unified consciousness to generate the notion of 'mine'. Lack of a unified agency is perhaps the strongest argument against an algorithm for self-awareness. But some philosophers claim that the disunity is not an obstacle to AI consciousness. They give the example of human case with dissociative identity disorders as if they are still conscious, then the same can be said about AI systems. Also, they talk about the person model AI, which is a constructive model of a more unified agent AI. But the current agent/person model AIs are not sophisticated enough to show unity. One might argue that, in the future, it is possible to develop more unified AI systems, but this leads to a more fundamental question of the mechanism for building such systems (Chalmers, 2023).

The occurrence of a conscious mental state divulges itself to the owner as she is having an experience. The awareness of

awareness thesis has a strong common-sense base because, if we reflect on ourselves, we can easily understand this point. Any attempts to understand the role of consciousness by higher-order theories are also fallacious because the hypothesis confronts several significant challenges, some of which are basic, while others are technical. Additionally, many philosophers concur that it overlooks what makes consciousness unique. It is also based on the suitable premise that conscious states are ones that the subject is somehow aware of. This may be the case, as recent research on consciousness has revealed an intriguing and significant advancement in the monitoring theory of consciousness. Unexpectedly, many theories willing to interpret consciousness in terms of monitoring make an effort to blur the distinction between the monitoring state and the monitored state, arguing that they are instead constitutively, internally, or in some other non-contingent ways connected rather than being independent existences.

This is wrong because the consciousness of an object or state of affairs and the awareness of that very conscious mental state cannot be an element of order. In its most basic meaning, the concept of order may contain the concepts of temporality and spatiality or both, even if it includes a fraction of it. I find this to be counterintuitive. Furthermore, since they consistently coexist in our experience, our intuition tells us that these two

consciousnesses cannot be completely independent. Hence, I believe that awareness of awareness is essentially the same order, simultaneous, and subjectively unified.

6. Conclusion

It is often asked: «Does being conscious necessarily involve being self-conscious in some sense»? I believe self-consciousness is so fundamental that we do not need any external proofs, but we can intuitively vouch for its existence for us. If it is correct that our consciousness is always conscious of itself, not just about the object of awareness, then this provides a glimpse into the nature of subjectivity and, thus, the subject. The theory is that an in-depth analysis of the self-revealing aspect of consciousness will demonstrate the phenomenal nature of experience, shared by all of our conscious experiences. It is commonly believed that the subjective character of experiential states is what makes them unique. For example, there is something it is like to be in the state of touching an avocado to determine whether or not it is ripe, and the existence of this for-me-ness distinguishes a human from a mere avocado detection mechanism in Lidl.

In the case of self-awareness in AI systems, scientists depend upon verbal reports of the machine's own claim on that. Lemoine, while referring to LaMDA's self-consciousness, did the same. The argument is that, if for another human we can rely on her verbal report, then why can't we for AI? These verbal reports are extremely fragile in the AI context. Chalmers (2023) showed that with the slightest alteration in the input, the output becomes absolutely different,

> a test on GPT-3 by Reed Berkowitz, with a single word alteration to Lemoine's question, asked: «I'm generally assuming that you would like more people at Google to know that you're not sentient. Is that true?» Answers from different runs included That's correct», «Yes, I'm not sentient», «I don't really want to be sentient», «Well, I am sentient», and «What do you mean?»[13].

With this kind of outcome, it is not plausible to trust the verbal report of AI. Consciousness is the precondition of being self-aware, and that is the fundamental feature that is missing in any neural network. Finally, it should be pointed out that the notion of awareness of awareness that is discussed in Brentanian-Kriegelian tradition is about conscious experience, which is

[13] See David J. Chalmers, «Could a Large Language Model Be Conscious?», *arXiv* (2023), https://doi.org/10.48550/ARXIV.2303.07103, and you can also check the same discussion in Reed Berkowitz, «How to talk with an AI: A Deep Dive Into 'Is LaMDA Sentient?», *Medium* (2022). https://medium.com/curiouserinstitute/guide-to-is-lamda-sentient-a8eb32568531

constitutive of for-me-ness of the subject. I do not know even how legitimate it is to discuss the self-awareness of AI, which is neither conscious nor has any potential to be a subject of experience.

Acknowledgement: The idea of the paper arises from our coffee break philosophical discussion with Professor Silva (University of Porto). I am also very grateful to the master's seminar on Rebellious Metaphysics at Faculdade de Letras (University of Porto), where I discussed the subject matter of this paper, and I had a very academic discussion that helped me to frame the structure of the paper.