Metacognitive Closure and Consciousness in Large Language Models

Shun Yoshizawa[1,2,3] and Ken Mogi[2,3,4]

[1]Tokai University, Kanagawa, Japan
[2]Sony Computer Science Laboratories, Tokyo, Japan
[3]UT-LAB Institute, Tokyo, Japan
[4]The University of Tokyo, Tokyo, Japan
{shunyoshizawa57, kenmogi2005qualia}@gmail.com

Abstract

Metacognition is an important aspect of information processing in the brain, subserving judgement and making cognition robust. In the literature, there are different views on the role metacognition plays in consciousness. The role of metacognition has been addressed by various studies. In relation to consciousness, some authors argue that metacognition is not necessarily essential in consciousness, but rather an extra mechanism constructed on a more basic mechanism, necessary when reflecting on and reporting one's own experiences. Others hold that metacognition is an integral part of phenomenal consciousness, possibly accounting for the hard problem of consciousness eventually. We aim to clarify why no consensus has been reached on whether large language models can possess consciousness, and why diverse and competing positions persist regarding the nature and plurality of consciousness. On the strength of the analysis, we propose *metacognitive closure,* a concept analogous to Colin McGinn's cognitive closure. We discuss the possibility that difficulties in elucidating mechanisms of consciousness might be clarified by considering the nature of metacognition. Based on this view, we argue how we may be able to streamline issues in consciousness through an analysis of metacognition, in a continuous spectrum with problems in cognition in general.

Introduction

Problems of consciousness have been investigated through various methodologies, including neuroscience (Crick & Koch 1998), cognitive science (Århem & Liljenström 1997), physics (Penrose 1987), and philosophy (Searle 2005). There have been some significant developments, which clarified some salient points in the field. The correspondence between conscious percepts and neural activities have been reported empirically (Heywood et al. 1992, Logothetis et al. 1996, Tong et al. 2006, Wade et al. 2008, Carriere et al. 2020), including those corresponding to illusions (Zeki et al. 1993, Golaszewski et al. 2021,) Aspects of the conscious perception of time have been elucidated (Eagleman 2008, Libet 2009, Herai and Mogi 2014). Investigations into the mechanism of anesthesia (Franks and Lieb 1982, Pavel et al. 2020) provide important constraints on the mechanism of consciousness. Phenomenal consciousness is often described in terms of three central properties. Privateness refers to the fact that conscious experience is directly accessible only to the subject. Intrinsicality denotes the irreducible qualitative nature of experience. Ineffability captures its resistance to complete linguistic description (Nagel 1974; Chalmers 1996; Dennett 1991). Chalmers (1996) proposed the distinction between the easy and hard problems of consciousness, in which the hard problems are concerned with the phenomenal aspects of consciousness such as qualia, intentionality, and self-consciousness. Neural correlates of consciousness (NCC) has been described as a minimal set of neuronal events sufficient for a conscious percept (Koch 2004). Although various scientific theories of consciousness have been developed, there is as yet still no definitive account of it. In view of these difficulties, Colin McGinn proposed "cognitive closure" to clarify why and how the problem of consciousness is so hard (McGinn 1999). The cognitive closure argument noted that human cognitive capacity, which has developed in evolution to cater to the needs of human survival and therefore is creature-specific, may be closed as to the understanding of the phenomenal consciousness. Cognitive closure describes the difficulty in understanding the nature of the phenomenological consciousness, a position associated with the ideas of transcendental naturalism (McGinn 1994) or new mysterianism (Flanagan 1991).

Metacognition refers to the ability to observe and control one's own internal states related to higher-order cognitive abilities such as introspection, self-awareness, and metamemory, as well as phenomenological consciousness. In everyday usage, the root word "meta" means "above" or

"beyond", and "metacognition" can be taken to signify cognition occurring later than or beyond a particular cognition, augmented by its embodied and physical connotations (Skulmowski & Rey 2017). On a more abstract level, metacognition is related to self-reference, an important theme in logic (Coffa 1979), tightly connected to the mathematics of computing (Kampis 1995). In human learning processes, metacognition would come after cognition, providing an important point of research in cognitive development (Fisher 1998, Whitebread & Neale 2020). Studies on metacognition from animals to humans (Smith et al. 2005, Shields et al. 2005, Ferrigno et al. 2019) have focused on confidence levels, the ability to judge how confident a subject is about a particular response. Intriguingly, there appears to be intimate relations between metacognition, cognitive closure, and consciousness. In psychology, "closure" refers to the tendency for human subjects to find it necessary to arrive at a stable cognition, thus avoiding ambiguity (Kruglanski and Webster 1996). Once a stable cognition is formed, it is typically difficult to get out of the "echo chamber" of cognition (Kruger and Dunning 1999, Barzilai & Chinn 2020). As in the process of cognitive development (Fisher 1998) and confidence judgement (Fleming 2023), metacognition might provide possible routes for breaking out of the cognitive echo chamber towards more flexible cognition. However, as McGinn (1994) points out, it is extremely difficult, if not demonstratively impossible, to have a metacognitive vantage viewpoint about the phenomenology of one's consciousness. Closure in the psychological sense would provide computational stability, and would be evolutionarily adaptive, albeit inflexible in metacognition (Kruglanski et al. 2006). The robustness of closure might be salient in relation to cognition of phenomenal consciousness, the disruption of which can lead to mental instabilities (Insel 2010, Pierre 2010). Thus, it is interesting to examine the role of metacognition in the elucidation of consciousness, related to, but not necessarily limited to, cognitive closure.

The relationship between metacognition and consciousness has been the subject of various debates. Metacognition and its relation to consciousness have been studied through introspection, in which William James has made an essential contribution with the proposal of the concept of the "stream of consciousness" (James 1890). Hofstadter (1979) discussed the nature of self-reference in metacognition. Nelson (1996) addressed the relation between the psychological and philosophical approaches to metacognition. Rosenthal (2005) discussed the role of higher-order cognition in producing mental states. Lau et al. (2011) argued for the importance of higher-order cognition supported by the prefrontal areas. Nelson (1996) and Hofstadter (1979) have addressed self-referential structures encompassing metacognition, introspection, and consciousness. Koriat (2007) argued that unconscious processes may lie at the heart of consciousness, and discussed the relationship between consciousness and unconsciousness from a metacognitive perspective. Rosenthal (2012), on the other hand, suggested that consciousness and metacognition have little in common, while both involving higher-order psychological states. Fleming (2021) wrote that “if metacognition is inherent to conscious experience, it may be difficult for us to reflect on consciousness itself,” and observed that there are two positions on whether metacognition is "inherent" in consciousness. One is the “first-order” position, which holds that metacognition is not inherent in consciousness. The other is the “higher-order” position, which holds that metacognition is inherent in consciousness.

Consciousness is a process in which one accesses one's internal state. If consciousness indeed has a metacognitive origin, having metacognitive knowledge about one's perception's and feelings, it would help reportability of one's internal states to others, laying the foundations for communication in general. Although consciousness, self-consciousness in particular, is typically regarded to have self-centered functional and phenomenological significances, seen from this particular viewpoint it is also social in nature (Robbins 2008).

Is metacognition an essential nature of consciousness? Or, going even further, is metacognition consciousness? How is cognitive closure to be involved in the characterization of consciousness in its relation to metacognition?

Here we consider the possibility that there is specific cognitive closure associated with metacognition. We introduce the concept of *metacognitive closure* to make sense of the essential role of metacognition in consciousness. We will argue that *metacognitive closure* is useful in understanding various issues in cognition related to consciousness in reference to and also apart from the hard problem of consciousness. Especially, different perspectives about the problem of consciousness may reflect not only theoretical and disciplinary commitments but also individual differences in metacognitive access to one’s own phenomenological consciousness, as well as the inherent diversity and partial inaccessibility of such introspective processes.

Insights from *metacognitive closure* might be applied to artificial intelligence systems as well. Neural mechanisms of metacognitive monitoring and control would offer directions of improving artificial intelligence systems such as transformer (Vaswani et al. 2017). The variabilities in *metacognitive closure* could be reproduced in artificial systems, shedding light on the human cortical mechanisms involved, and expanding beyond the range of typical human cognition. Generative AI systems often generate outputs inconsistent with reality. In LLMs, hallucinations (Huang 2023) have presented theoretical and practical challenges. Wang and Zhao (2023) tried to enhance reasoning and understanding in LLMs by the use of prompting techniques called metacognitive promoting (MP), employing words and phrases inducing the machine to employ metacognitive monitoring,

which can increase the performance of LLMs compared to standard prompts, with overthinking and overcorrection errors occurring as negative side effects. Such a behavior on the part of LLM, in turn, would reflect statistical correlations between word sequences related to metacognition and general knowledge in human natural languages. Researchers at Anthropic tried to introduce P(IK) into LLMs to create an honest AI, where P(IK) is a probability function similar to the feeling-of-knowing (FOK) (Kadavath et al. 2022). Interestingly, there are possibilities of assessing and manipulating metacognition in artificial intelligence systems in ways impossible in human subjects, including research in the burgeoning field of mechanistic interpretability (Liu et al. 2024).

We extend the points of relevance across a wide spectrum of cognitive traits, e.g. the nature of hallucination in patients with schizophrenia (Amador 1991, Lehrer & Lorenz 2014), lack of self-assessment in the classroom (Kruger and Dunning 1999), free will (Mogi 2014), and provide a framework for understanding the difficulty of reaching a consensus on whether Large Language Models possess consciousness.

The Metacognitive Nature of Consciousness

Various aspects of consciousness would appear to be related to metacognition. Block (1995) proposed a conceptual distinction between phenomenal consciousness ("P-consciousness") and access consciousness ("A-consciousness"). P-consciousness is experience, is not functional, and can be paraphrased as "what it is like" (Nagel 1974). A-consciousness is the functional aspect of consciousness that we can cognitively access, an example of which is a verbal report (“reportability”). The distinction between P- and A- consciousnesses has been useful in streamlining different aspects of consciousness research, with the former concerned with the phenomenological aspects of consciousness such as the phenomenology of qualia (Ramachandran & Hirstein 1997), and with the latter specifically concerned with functional and social cognitive implications of consciousness (Frith and Frith 2007, Robbins 2008), which essentially involves metacognition.

Rosenthal (2000) advocated one version of what came to be known as Higher-order thought (HOT) theories of consciousness, and argued that "a mental state is conscious only if one is, in some suitable way, conscious of that state”. HOT theories argue that the first-order mental state is not sufficient for generating a phenomenological consciousness, and that higher-order mental states are necessary for the emergence of consciousness. Here, first-order mental states refer to sensory representations of the environment, and the higher-order states facilitate the interpretations and integrations of these representations into higher brain functions. According to HOT theories, by nature of phenomenal consciousness in our cognitive system, higher cognitive processes such as metacognition about first-order cognitive processes need to emerge. Consciousness in this view is essentially metacognitive in nature, involving self-referential structures, as the phenomenal consciousness of the higher-order cognitive process fundamentally requires the metacognitive process for the higher-order cognitive process itself. There is the idea that self-referential conscious thinking is possible only when the higher-order representations are themselves conscious (Rosenthal 2005, Lau and Rosenthal 2011). One criticism directed to HOT theories is that despite the apparent merits, the reason why the phenomenology of consciousness arises at all is yet to be accounted for (Seth 2022). The essential mechanism of how the phenomenology of consciousness would arise when the higher cognitive process such as metacognition is coupled with the first-order cognitive process is not yet clear. It would be interesting to explore possible mechanisms in which the phenomenology of consciousness could emerge in a boot-strapping manner (Hofstadter 1979) involving metacognitive process of self-reference, a position still speculative at present.

The Nature of Metacognition

Metacognition is typically involved when the subject has access to one's internal states, making judgements about them, communicating with others. Nature of metacognition across these cognitive domains need to be taken into consideration to elucidate its relation to consciousness.

Metacognition has been studied through the measurement of confidence in response to specific questions (Kepecs and Mainen 2012), sometimes retrospectively (retrospective confidence judgments (RCJs), Harts 1965, Robey et al. 2017). These paradigms are related to feeling-of-knowing (FOK, Nelson 1984, Koriat 1993), the cognitive measure of how much you know about what you know or do not know. Metacognition can be studied in agents other than humans. Metacognition in animals has been measured by observing how their behavior changes when the uncertainty of the task is varied (uncertainty response paradigm (UR), Smith et al. 2005), where animals cannot verbally report their confidence in their subjective experiences. The nature of metacognition in non-human animals help elucidate evolutionary constraints on metacognition. Metacognition has been regarded as central to intelligence, along with awareness of the self, and Feeling of Knowing, FOK (Clark and Karmiloff-Smith 1993, Fleming 2021). Learning can occur unconsciously, while involvement of explicit metacognition facilitates robust updating of the cognitive system (Fisher 1998). In analogy with human cognition, the role of metacognition in artificial intelligence systems has been investigated and discussed (Azevedo 2020, Kawato & Cortese 2021).

Some studies have suggested links between metacognition and theory of mind (Kuhn 2000). Cortical processing involving mirror neurons and mirror systems (Gallese & Goldman1998, Rizzolatti, et al. 2002) likely subserve metacognitive information processing common to cognition about the self and others. Metacognition in this context is particularly interesting, as it could serve as a bridge between the self-centered and social aspects of consciousness.

Efforts have been made to develop tangible models of metacognition. The psychological model in Nelson and Narens (1990) describes the metacognitive process as a flow of information between the object-level and the meta-level, in which the object-level and meta-level represent elements involved in cognition and control in an abstract sense, following definitions by Hilbert (1927) and Carnap (1934). Shimamura (2008) reviewed related research works under the scheme of assigning the meta-level and object-level to the prefrontal cortex and the posterior cortex, respectively.

Links to mathematics have been a salient feature of investigations into metacognition. The Nelson-Nares Model was originally inspired (Nelson 1996) by the proposed solution by Alfred Tarski (Tarski 1956, Tarski 1985) about self-referential paradoxes such as the "Lier paradox", which generated further discussions (Kripke 1975). Nelson (1996) argued that there is an intimate relation between metacognition and consciousness relevant to the self-referential structure, suggesting that empirical findings about metacognitive monitoring and control would shed light on the philosophical aspects of consciousness. The self-referential structure of metacognition is defined as metacognition making metacognition about its own metacognition, a structure repeatedly addressed in various forms such as the Liar paradox, Russell's paradox (Coffa 1979), and diagonal argument (Cantor 1891, Sheppard 2014).

Metacognition and Consciousness

Metacognition in humans can be related to awareness, and can serve as an indicator of neural correlates of consciousness (Persaud et al. 2007). Surveying the literature, it would appear that properties essential in metacognition, e.g. the self-referential structure, could contribute to the foundation of the phenomenology of consciousness, e.g. qualia (Persaud & Lau 2008, Mogi 2013), intentionality (Gollwitzer & Schaal 2013), semantics (Markovits et al. 2015), and self-awareness (Proust 2013, Lou et al. 2017).

There are ongoing debates as to the relation between consciousness and metacognition. The implicit relationship between consciousness and metacognition has been referred to by Koriat (2007) and Nelson (1999). Some HOT theorists do not consider metacognition and consciousness to be equivalent. For example, Brown et al. (2019) argued that most proponents of HOT theories do not treat metacognition and consciousness as conceptually equivalent, while admitting that metacognition is a relevant mechanism for consciousness. Fleming and Frith (2012) expressed skepticism about the relevance of metacognition to consciousness, and argued that it is important to critically evaluate whether there is a close link between metacognition and consciousness in the first place. It would be interesting to consider the possibility that consciousness and metacognition could be separately treated. Brown et al. (2019) argued that HOT theories are agnostic on this question. Fleming (2021) pointed out that current spatial and temporal resolutions of brain imaging such as fMRI and MEG are not sufficient to address these questions empirically. The fundamental nature of metacognition being "meta" ("above" or "beyond"), is possibly related to the essence of consciousness as observation of the internal states of the self (Varela et al. 1974). The observation of one's own state is a prerequisite condition for reportability (Naccache 2018), supporting communication in the context of social cognition.

There are significant variabilities in one's metacognition of the phenomenal properties of consciousness such as qualia (Mogi 2013). The involvement of metacognition in cognitive development (Fisher 1998, Whitebread & Neale 2020) could be related to the plasticity of a subject's metacognition of conscious states. In the Buddhist tradition, there are said to be 52 steps towards enlightenment (Hosokawa 2020), in which awareness of one's conscious states plays an important role. Studies indicate that well-trained monks have brain activities in areas such as medial prefrontal cortex and temporoparietal junction significantly different from the general public (Manna et al. 2010, Ricard et al. 2014).

Finally, there is a question of explicit vs. implicit involvement. Metacognition typically addressed in HOT theories assumes explicit forms, such as confidence judgements**.** An alternative possibility is that metacognition is relevant to the phenomenological consciousness in more implicit, built-in ways, as in social interactions (Frith 2012). An implicit and within-the-system relationship between metacognition and consciousness would make it difficult to define metacognition, as is relevant for consciousness, a difficulty which may be addressed by approaches focusing on such tangible aspects as reportability (Naccache 2018) and metacognitive plasticity (Manna et al. 2010, Ricard et al. 2014).

Metacognitive Closure

We define *metacognitive closure* to be cognitive closure (McGinn 1999) specific to metacognition (Fleming and Frith 2012, Ferrigno et al.2019). In this sense, this particular

closure is a part of cognitive closure in general, but more specific to metacognitive monitoring and control (Smith 2005, Shimamura 2008), and closely connected to consciousness (Nelson 1999, Koriat 2007).

The concept of cognitive closure (McGinn 1999) is one of the candidate clarifications of the hard problem of consciousness, in that it explains why and how it is hard to elucidate the nature of consciousness. Here we attempt at a possible refinement, variation, and extension of the cognitive closure argument from a metacognitive point of view, to understand the relationship between cognitive closure in general.

McGinn's cognitive closure is a general concept, and can be applied in principle to any cognitive systems, natural or artificial, based on algorithm or otherwise, classical or quantum. In this respect, cognitive closure is a useful clarification of the hard problem of consciousness (Chalmers 1996) in a universal sense.

Metacognition is an efficient way to focus on specific difficulties essential in cognition, in relation to consciousness in particular. Of all aspects addressed in cognitive closure, there is something about metacognition that makes it difficult or almost impossible to get out of it. Following this line of argument in Fleming (2021), if metacognition indeed is essential in or equivalent to consciousness, one might not ultimately understand what phenomenal consciousness is. Thus one is trapped in the *metacognitive closure*.

In consideration of the fundamental role of metacognition in consciousness, and variabilities in metacognition, *metacognitive closure* is defined as the difficulty or impossibility of assessing one's metacognitive state from objective points of view, *at a particular moment*. This difficulty gives rise to diversity in metacognitive access to one's own phenomenal consciousness. In terms of the distinction between A-consciousness and phenomenal P-consciousness, *metacognitive closure* concerns the extent to which one can metacognitively access one's own P-consciousness, and the limitations of such A-consciousness. Even when individuals share the same empirical data and logical reasoning, their metacognition of the phenomenology of consciousness may vary due to *metacognitive closure*. As a result, diverse positions on the nature of consciousness and phenomenology emerge. It is also possible that privateness, one of the central properties of consciousness, emerges from *metacognitive closure*. Specifically, *metacognitive closure* may cause metacognition of one's own phenomenology of consciousness to diversify and become individually closed, thereby giving rise to the apparent privateness of conscious experience. *Metacognitive Closure* may make the problem of consciousness "hard" or private to be precise. The exact nature of *metacognitive closure* could vary over time, depending on the subject's learning (Fisher 1998, Whitebread & Neale 2020) or training, such as meditation (Manna et al. 2010, Ricard et al. 2014).

Such confining nature of metacognition at a particular moment is not limited to consciousness-related cognition. Metacognition is known to correlate with academic performance when controlled for intelligence (Ohtani & Hisasaka 2018). Subjects are known to lack metacognition of the insufficient nature of their knowledge and skills, especially when they are inferior in them (Kruger and Dunning 1999). The Dunning–Kruger effect is an instance of *metacognitive closure* defined here. Metacognition is an essential element of intelligence in children, including exceptional cases (Cornoldi 2010). The general metacognitive ability explained general academic achievement, rather than intelligence, in school children (Gomes et al. 2014). These findings would suggest that there are variabilities in the *metacognitive closure*, which might go through plastic changes during education.

Metacognitive closure exposited here would be useful in understanding various positions that exist in consciousness research. Indeed, we posit that *metacognitive closure* is more pronounced and typically difficult to overcome in the case of metacognition about one's phenomenal consciousness. Specifically, we explore the possibility that each subject might have different metacognitive monitoring and understanding of what the phenomenology of consciousness entails, and these different states in *metacognitive closure* might lead to different theorizing and models about consciousness. Since *metacognitive closures* in this sense is something preceding specific theories about consciousness, it may be difficult to clarify them, even for leaned scholars.

Opinions about the phenomenal nature of consciousness do vary. There are different ideas about the nature of phenomenological overflow (Block 2011, Lau & Rosenthal 2011). There are different opinions about the phenomenological nature of consciousness such as qualia (Dennett 1991, Chalmers 1996). These differences in opinion might reflect different metacognitive states about the phenomenology of consciousness present among researchers and the general public (Mogi 2013), rather than, or in addition to, differences in theoretical positions per se. Related to this observation, it has been noted that the definitions of metacognition differ between researchers. (Fleming and Frith 2014).

In the *metacognitive closure* model, each individual is metacognitively closed in terms of his or her own understanding of phenomenological consciousness. In this sense, *metacognitive closure* makes the problem of consciousness appear "hard*"* to communicate for each individual, in that it becomes difficult to exchange ideas beyond systems of *metacognitive closures* that are different from subject to subject.

The major difference between the conventional idea of cognitive closure and the *metacognitive closure* proposed here is the importance of variabilities among subjects. The metacognitive monitoring and control strategies adapted in face of uncertainties are different between subjects, in the case of confidence judgement, for example (Smith 2005). There would be significant individual differences in *metacognitive*

closure, within neurotypical as well as in neuroatypical subjects. In a condition known as anosognosia (Vuilleumier 2004), patients are unaware of their own states of disorder, e.g. blindness (Goldenberg et al. 1995), dementia (Wilson et al. 2016), and schizophrenia (Amador 1991, Lehrer & Lorenz 2014). Patients with schizophrenia are known to be less apt to have correct metacognition regarding their own hallucinations.

There may be biological and evolutionary reasons why variabilities in metacognition, especially as regards phenomenal consciousness, exist. One possible reason may be that whether one has metacognitive access to certain phenomenology of consciousness or not does not affect cognitive functions in everyday settings, in a way similar to the neutral theory of molecular evolution (Kimura 1977). In addition, there may be even functional advantages for a lack of metacognition. Although lack of metacognition of one's objective knowledge level is disadvantageous in the classroom (Kruger and Dunning 1999), by correlating positively with the perception of free will it might have advantages in broader context of life, leading to more proactive behavior (Dunning 2011, Mogi 2014).

Here, it is important to note that the nature of *metacognitive closure* could change over time. By explicitly focusing on *metacognitive closure*, which is a part of the more general cognitive closure, there might be ways to improve it, in ways similar to but not limited to the meditative training in the Buddhist tradition (Hosokawa 2020). The focus on *metacognitive closure* would also shed light on the relevance of the first person approach (Velmans 1991, Vogeley et al. 2004).

Metacognitive Closure in AI

There is currently no consensus among researchers on whether LLMs that pass the Turing Test possess consciousness. Opinions remain divided. AI researcher Roman Yampolskiy argues (Yampolskiy 2025) that it is preferable to treat AI systems as if they possess consciousness rather than assuming they do not, because this stance minimizes practical and ethical risks associated with AI deployment from a For All Practical Purposes (FAPP) perspective. In contrast, neuroscientist Ken Mogi recommends initiating discussions from the null hypothesis that AI does not possess consciousness (Mogi 2024). Philosopher David Chalmers has argued that current LLMs are unlikely possess consciousness (Chalmers 2023). Neuroscientist Anil Seth has argued that attributing consciousness to AI stems from human biases, such as anthropocentrism (Seth 2025). Furthermore, it has been suggested that metacognition in large language models (LLMs), particularly GPT-4, is condition-specific and lacks the robustness observed in humans (Yoshizawa et al. 2026).

We argue that the diversification of positions on consciousness and phenomenology brought about by *metacognitive closure* is also reflected in the lack of consensus on whether LLMs can possess consciousness. Because metacognitive access to phenomenological consciousness varies across researchers and individuals as a result of *metacognitive closure*, debates over what would count as consciousness in LLMs may fail to converge, even when shared empirical evidence and theoretical frameworks are available. Furthermore, if artificial systems such as LLMs or other AI agents are to possess consciousness, it may be necessary to engineer a form of privateness, namely a mode of access to phenomenological states that is, in principle, available only to the system itself.

Discussion

In this paper, we introduced *metacognitive closure* as a more fine-focused viewpoint related to consciousness compared to the cognitive closure concept proposed by McGinn (1999).

Self-reference is one of the crucial aspects of metacognition, and has been discussed as one of the key issues in computation (Turing 1950). Hofstadter (1979) discussed issues surrounding strange loops in logic, mathematics, drawings, music, and cognition in general, elucidating the universal relevance of self-reference.

We have argued that metacognition lies at the heart of consciousness. *Metacognitive closure* as defined in this paper accounts for individual differences about the perception of the phenomenal nature of consciousness. In addition, the hard problem of consciousness can be linked to a wide spectrum of cognitive phenomenon through *metacognitive closure*, suggesting possible ways to ameliorate difficulties involved.

Thus, through the conceptualization of *metacognitive closure*, the hard problem of consciousness can be dealt with as an instance of general problems concerning human cognition, sharing the same scheme with anosognosia (Vuilleumier 2004), failure of academic self-assessment (Kruger and Dunning 1999), and perception of free will in everyday life (Dunning 2011, Mogi 2014). Thus finding a broader context in which the hard problem of consciousness can be discussed through *metacognitive closure*, tapping into a trove of related research, would help elucidate the enigma of consciousness. There may still be qualitative differences between *metacognitive closure* specific to phenomenal consciousness and other aspects of cognition, but discussing in terms of *metacognitive closure* would help clarifying the difference, so we may eventually know if the hard problem of consciousness is indeed hard, and if so, in what metacognitive sense specifically.

References

Azevedo, R. 2020. Reflections on the field of metacognition: Issues, challenges, and opportunities. *Metacognition and Learning*, *15*, 91-98.

Amador, X. F., Strauss, D. H., Yale, S. A., & Gorman, J. M. 1991. Awareness of illness in schizophrenia. *Schizophrenia Bulletin*,17(1),113–132.

Århem, P., & Liljenström, H. 1997. On the coevolution of cognition and consciousness. *Journal of Theoretical Biology*, *187*(4), 601-612.

Barzilai, S., & Chinn, C. A. 2020. A review of educational responses to the "post-truth" condition: Four lenses on "post-truth" problems. *Educational Psychologist*, *55*(3), 107-119.

Block, N. 1995. On a confusion about a function of consciousness. *Brain and Behavioral Sciences* 18 (2):227-–247. doi:10.1017/S0140525X00038188

Block, N. 2011. Perceptual consciousness overflows cognitiveaccess. *Trends Cogn. Sci*. 15, 567–575

Brown R, Lau H, LeDoux JE. 2019. Understanding the Higher-Order Approach to Consciousness. *Trends Cogn Sci*. 23, 754-768.

Cantor, G. 1891. "Ueber eine elementare Frage der Mannigfaltigkeitslehre". J*ahresbericht der Deutschen Mathematiker-Vereinigung.* 1: 75–78.

Carnap, R. 1934. Logische syntax der sprache. Springer.

Carriere, M., Larroque, S.K., Martial, C., Bahri, M.A., Aubinet, C., Perrin, F., Laureys, S. and Heine, L., 2020. An echo of consciousness: Brain function during preferred music. *Brain Connectivity*, *10*(7), 385-395.

Chalmers, D. 1996. *The Conscious Mind: In search of a Fundamental Theory*. Oxford University Press.

Chalmers, D. J. 2023. Could a Large Language Model be Conscious? doi: 10.48550/arXiv.2303.07103

Clark, A., & Karmiloff-Smith, A. 1993. The cognizer's innards: A psychological and philosophical perspective on the development of thought. *Mind & Language, 8*(4), 487–519.

Coda-Forno, J., Witte, K., Jagadish, A. K., Binz, M., Akata, Z., & Schulz, E. 2023. Inducing anxiety in large language models increases exploration and bias. *arXiv preprint arXiv:2304.11111*.

Coffa, J. A. 1979. The humble origins of Russell's paradox. *Russell: The Journal of Bertrand Russell Archives*, *33*(1), 31-37.

Cornoldi, C. 2010. Metacognition, intelligence, and academic performance. *Metacognition, strategy use, and instruction*, 257-277.

Crick, F., & Koch, C. 1998. Consciousness and neuroscience. *Cerebral cortex*, *8*(2), 97-107.

Dennett, D. C. 1991. Consciousness Explained. Little, Brown and Co.

Dunning, D. 2011. The Dunning–Kruger Effect: On Being Ignorant of One's Own Ignorance. in *Advances in Experimental Social Psychology*. Vol. 44. Academic Press. pp. 247–296

Eagleman, D. M. 2008. Human time perception and its illusions. *Current opinion in neurobiology*, *18*(2), 131-136.

Ferrigno, S., Bueno, G., & Cantlon, J. F. 2019. A similar basis for judging confidence in monkeys and humans. *Animal Behavior and Cognition*, *6*(4), 335-343.

Fisher, R. 1998. Thinking about thinking: Developing metacognition in children. *Early Child Development and Care*, *141*(1), 1-15.

Flanagan, O. 1991. The Science of the Mind. MIT Press.

Fleming, S. M., Dolan R. J. and Frith, C. D. 2012. Metacognition: computation, biology and function. Phil. Trans. R. Soc. B 367, 1280–1286

Fleming, S.M. 2021. Know Thyself: The New Science of Self-Awareness. Basic books

Fleming, S. M. 2023. Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, *75*. 1-23

Fleming, S. M., & Frith, C. D. 2014. Metacognitive neuroscience: An introduction. In Fleming S. M. & Frith, C. D. (Eds.), The cognitive neuroscience of metacognition (pp. 1–6). Springer-Verlag Publishing.

Franks, N. P., & Lieb, W. R. 1982. Molecular mechanisms of general anaesthesia. *Nature*, *300*(5892), 487-493.

Frith, C. D., & Frith, U. 2007. Social cognition in humans. *Current biology*, *17*(16), R724-R732.

Frith, C. D. 2012. The role of metacognition in human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2213-2223.

Gallese, V., & Goldman, A. 1998. Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, *2*(12), 493-501.

Golaszewski, S., Frey, V., Thomschewski, A., Sebastianelli, L., Versace, V., Saltuari, L., ... & Nardone, R. 2021. Neural mechanisms underlying the Rubber Hand Illusion: A systematic review of related neurophysiological studies. *Brain and Behavior*, *11*(8), e02124.

Goldenberg, G., MuḶlbacher, W., & Nowak, A. 1995. Imagery without perception—a case study of anosognosia for cortical blindness. *Neuropsychologia*, *33*(11), 1373-1382.

Gollwitzer, P. M., & Schaal, B. 2013. Metacognition in action: The importance of implementation intentions. In *Metacognition* (pp. 124-136). Psychology Press.

Gomes, C. M. A., Golino, H. F., & Menezes, I. G. 2014. Predicting school achievement rather than intelligence: Does metacognition matter?. Psychology, 5(09), 1095-1110.

Hart JT. 1965. Memory and the feeling-of-knowing experience. J Educ Psychol. 56, 208-216.

Heywood, C. A., Gadotti, A., & Cowey, A. 1992. Cortical area V4 and its role in the perception of color. *Journal of Neuroscience*, *12*(10), 4056-4065.

Herai, T., & Mogi, K. 2014. Perception of temporal duration affected by automatic and controlled movements. *Consciousness and cognition*, *29*, 23-35.

Hilbert. D. 1927. Uber das Unendliche. Jnhresbericht der Deutschett Math~matiker-Vereinigung. 36 201-215.

Hofstadter, Douglas R., 1979. Gödel, Escher, Bach : an eternal golden braid. New York :Basic Books

Hosokawa, S. 2020. *Zen Wisdom for the Anxious: Simple Advice from a Zen Buddhist Monk*. Tuttle Publishing

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B. and Liu, T., 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Insel, T. R. 2010. Rethinking schizophrenia. *Nature*, *468*(7321), 187-193.

James, W. 1890. The Principles of Psychology. New York: Henry Holt and Company the Principles of Psychology.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E. and Johnston, S., 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Kampis, G. 1995. Computability, self-reference, and self-amendment. *Commuincation and Cognition-Artificial Intelligence*, *12*, 91-109.

Kawato, M., & Cortese, A. 2021. From internal models toward metacognitive AI. *Biological cybernetics*, *115*, 415-430.

Kepecs A, Mainen ZF. 2012. A computational framework for the study of confidence in humans and animals. Philos Trans R Soc Lond B Biol Sci 367, 1322-37.

Kimura, M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, *267*(5608), 275-276.

Koch, C. 2004. The Quest for Consciousness: A Neurobiological Approach. Englewood, CO: Roberts & Company Publishers.

Koriat, A. 1993. How do we know that we know? The accessibility model of the feeling of knowing. *Psychological review*, *100*(4), 609.

Koriat, A. 2007. Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), The Cambridge handbook of consciousness (pp. 289–325). Cambridge University Press.

Kosinski, M. 2023. Theory of mind might have spontaneously emerged in large language models. *Preprint at https://arxiv. org/abs/2302.02083*.

Kripke, S. 1975. Outline of a theory of truth. The journal of philosophy, 72(19), 690-716.

Kruger, J., & Dunning, D. 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology, 77(6), 1121–1134

Kruglanski, A. W., Pierro, A., Mannetti, L., & De Grada, E. 2006. Groups as epistemic providers: need for closure and the unfolding of group-centrism. *Psychological review*, *113*(1), 84.

Kruglanski, A. W. and Webster, D. M. 1996. Motivated closing of the mind:
'Seizing' and 'freezing'. Psychological Review. 103 (2): 263–83.

Kuhn, D. 2000. Theory of mind, metacognition, and reasoning: A life-span perspective. in *Children's reasoning and the mind*, (Psychology Press) 301-326.

Lau, H. and Rosenthal, D. 2011. Empirical support for higher-
order theories of conscious awareness. Trends Cogn. Sci. 15, 365–373

Lehrer & Lorenz. 2014. Anosognosia in schizophrenia: hidden in plain sight. *Innovations in clinical neuroscience*, *11*(5-6), 10.

Libet, B. 2009. *Mind time: The temporal factor in consciousness*. Harvard University Press.

Liu, Z., Gan, E., & Tegmark, M. 2024. Seeing is believing: Brain-inspired modular training for mechanistic interpretability. *Entropy*, *26*(1), 41.

Logothetis, N. K., Leopold, D. A., & Sheinberg, D. L. 1996. What is rivalling during binocular rivalry?. Nature, 380(6575), 621-624.

Lou, H. C., Changeux, J. P., & Rosenstand, A. 2017. Towards a cognitive neuroscience of self-awareness. *Neuroscience & Biobehavioral Reviews*, *83*, 765-773.

Manna, A., Raffone, A., Perrucci, M. G., Nardo, D., Ferretti, A., Tartaro, A., ... & Romani, G. L. 2010. Neural correlates of focused attention and cognitive monitoring in meditation. *Brain research bulletin*, *82*(1-2), 46-56.

Markovits, H., Thompson, V. A., & Brisson, J. 2015. Metacognition and abstract reasoning. *Memory & cognition*, *43*, 681-693.

McGinn, C. 1994. "The Problem of Philosophy". Philosophical Studies. 76 (2–3): 133–56.

Mcginn, C. 1999. The Mysterious Flame: Conscious Minds in a Material World. Basic Books.

Mogi, K. 2013. Cognitive factors correlating with the metacognition of the phenomenal properties of experience. *Scientific Reports* 3, Article number: 3354 doi:10.1038/srep03354

Mogi, K. 2014. Free will and paranormal beliefs. *Frontiers in Psychology* 5, 00281. doi: 10.3389/fpsyg.2014.00281

Mogi, K.2024. Artificial intelligence, human cognition, and conscious supremacy. *Front Psychol*. 13;15:1364714. doi: 10.3389/fpsyg.2024.1364714.

Moghaddam, S. R., & Honey, C. J. 2023. Boosting Theory-of-Mind Performance in Large Language Models via Prompting. *arXiv preprint arXiv:2304.11490*.

Naccache, L. 2018. Why and how access consciousness can account for phenomenal consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1755), 20170357.

Nagel, T. 1974. What is it like to be a bat? Philos. Rev. 83, 4435–4450

Nelson, T. O. 1984. A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological bulletin*, *95*(1), 109.

Nelson, T. O. & Narens, L. 1990. Metamemory: A theoretical framework and new findings. Psy. Learn. Mot. 26, 125–141

Nelson, T. O. 1996. Consciousness and metacognition. American Psychologist, 51(2), 102–116.

Ohtani, K., & Hisasaka, T. 2018. Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance. *Metacognition and Learning*, *13*, 179-212.

Pavel, M. A., Petersen, E. N., Wang, H., Lerner, R. A., & Hansen, S. B. 2020. Studies on the mechanism of general anesthesia. *Proceedings of the National Academy of Sciences*, *117*(24), 13757-13766.

Penrose, R. 1987. Quantum physics and conscious thought. *Quantum implications: Essays in honour of David Bohm*, 105-120.

Persaud, N., McLeod, P. & Cowey, A. 2007. Post-decision wagering objectively measures awareness. Nat Neurosci 10, 257–261.

Persaud, N., & Lau, H. 2008. Direct assessment of qualia in a blindsight participant. *Consciousness and cognition*, *17*(3), 1046-1049.

Pierre, J. M. 2010. Hallucinations in nonpsychotic disorders: Toward a differential diagnosis of "hearing voices". *Harvard review of psychiatry*, *18*(1), 22-35.

Proust, J. 2013. *The philosophy of metacognition: Mental agency and self-awareness*. OUP Oxford.

Ramachandran, V. S., & Hirstein, W. 1997. Three laws of qualia: What neurology tells us about the biological functions of consciousness. *Journal of consciousness studies*, *4*(5-6), 429-457.

Ricard, M., Lutz, A., & Davidson, R. J. 2014. Mind of the meditator. *Scientific American*, *311*(5), 38-45.

Rizzolatti, G., Craighero, L., & Fadiga, L. 2002. The mirror system in humans. *Mirror neurons and the evolution of brain and language*, *42*, 37-59.

Robbins, P. 2008. Consciousness and the social mind. *Cognitive Systems Research*, *9*(1-2), 15-23.

Robey, A. M., Dougherty, M. R., & Buttaccio, D. R. 2017. Making retrospective confidence judgments improves learners' ability to decide what not to study. *Psychological Science*, *28*(11), 1683-1693.

Rosenthal, D.M. 2000. Metacognition and Higher-Order Thoughts, Consciousness and Cognition, 9, 231-242

Rosenthal, D.M. 2005. Consciousness and Mind, Oxford University Press

Rosenthal D. M. 2012. Higher-order awareness, misrepresentation and function. Philos Trans R Soc Lond B Biol Sci. 367, 1424-38.

Safdari, M., Serapio-García, G., Crepy, C., Fitz, S., Romero, P., Sun, L., Abdulhai, M., Faust, A. and Matarić, M., 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Shields, W. E., Smith, J. D., Guttmannova, K., & Washburn, D. A. 2005. Confidence judgments by humans and rhesus monkeys. *The Journal of general psychology*, *132*(2), 165.

Skulmowski, A., & Rey, G. D. 2017. Bodily effort enhances learning and metacognition: Investigating the relation between physical effort and cognition using dual-process models of embodiment. *Advances in cognitive psychology*, *13*(1), 3-10.

Searle, John R. 2004. Mind: A Brief Introduction. Oxford University Press

Seth, A.K., Bayne, T. 2022. Theories of consciousness. Nat Rev Neurosci 23, 439–452

Seth AK. Conscious artificial intelligence and biological naturalism. 2025. *Behavioral and Brain Sciences*. Published online :1-42. doi:10.1017/S0140525X25000032

Sheppard, B. 2014. *The Logic of Infinity*. Cambridge University Press.

Shimamura, A. P. 2008. A neurocognitive approach to metacognitive monitoring and control. In J. Dunlosky & R. A. Bjork (Eds.), Handbook of metamemory and memory (pp. 373–390). Psychology Press.

Smith, J. D. 2005. Studies of Uncertainty Monitoring and Metacognition in Animals and Humans. In H. S. Terrace & J. Metcalfe (Eds.), The missing link in cognition: Origins of self-reflective consciousness (pp. 242–271). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195161564.003.0010

Tarski, A. 1956. The concept of truth in formalized languages. In A.
Tarski (Ed.), Logic, semantics, metamathematics (pp. 152-178). Clarendon Press.

Tarski, A. 1985. The semantic conception of truth. In A. P. Martinich
(Ed.), The philosophy of language (pp.48-71). Oxford University Press.

Tong, F., Meng, M., & Blake, R. 2006. Neural bases of binocular rivalry. *Trends in cognitive sciences*, *10*(11), 502-511.

Turing, Alan. 1950. "Computing Machinery and Intelligence" Mind, LIX (236): 433–460,

Velmans, M. 1991. Consciousness from a first-person perspective. *Behavioral and Brain Sciences*, *14*(4), 702-726.

Varela, F. G., Maturana, H. R., & Uribe, R. 1974. Autopoiesis: The organization of living systems, its characterization and a model. *Biosystems*, *5*(4), 187-196.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, *30*.

Vogeley, K., May, M., Ritzl, A., Falkai, P., Zilles, K., & Fink, G. R. 2004. Neural correlates of first-person perspective as one constituent of human self-consciousness. *Journal of cognitive neuroscience*, *16*(5), 817-827.

Vuilleumier, P. 2004. Anosognosia: the neurology of beliefs and uncertainties. *Cortex*, *40*(1), 9-17.

Wade, A., Augath, M., Logothetis, N., & Wandell, B. 2008. fMRI measurements of color in macaque and human. *Journal of vision*, *8*(10), 6-6.

Wang, Y. and Zhao, Y., 2023. Metacognitive prompting improves understanding in large language models. *arXiv preprint arXiv:2308.05342*.

Whitebread, D., & Neale, D. 2020. Metacognition in early child development. *Translational Issues in Psychological Science*, *6*(1), 8.

Wilson, R. S., Sytsma, J., Barnes, L. L., & Boyle, P. A. 2016. Anosognosia in dementia. *Current neurology and neuroscience reports*, *16*, 1-6.

Yampolskiy, R. 2025. Uploading consciousness to machines. YouTube video. The Institute of Art and Ideas. https://youtu.be/-5ONB-0MdDk?si=uFHCK6gSjV-wxjQs. Accessed: 2026-01-27.

Yoshizawa, S., Onzo, A., Nozawa, S., Takano, T., Ishikawa, T., Ken, M. 2026. Metacognition of ChatGPT in confidence judgments. *Frontiers in Artificial Intelligence. In press.*

Zeki, S., Watson, J. D., & Frackowiak, R. S. 1993. Going beyond the information given: the relation of illusory visual motion to brain activity. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *252*(1335), 215-222.