Proceedings of the 4th International Conference on Interdisciplinary Humanities and Communication Studies
DOI: 10.54254/2753-7064/2025.ND28720

# *Privileged Access of Self-Knowledge: Its Philosophical Foundation and Challenge from Psychology and Neuroscience*

**Xiangdong Sun**

*International Department of the High School Affiliated to Capital Normal University, Beijing, China*
*zarathustr_2023@163.com*

***Abstract.*** This paper examines the philosophical concept of privileged access to self-knowledge—the notion that individuals possess direct, authoritative, and epistemically superior knowledge of their own mental states (e.g., thoughts, sensations, intentions) compared to knowledge of the external world or others' minds. It analyzes dominant philosophical theories (Acquaintance Theory, Inner Sense Theory, Agency Theory) that seek to explain and justify this privileged status, emphasizing claims of immediacy, reliability, and distinctive epistemic grounding. However, the paper then outlines significant challenges from psychology and cognitive neuroscience, demonstrating systematic limitations and errors in self-knowledge. Key evidence includes interpretivist theories showing self-knowledge as inferential rather than direct, research on implicit attitudes revealing unconscious biases contradicting explicit reports, and neuroscientific models (e.g., forward models, disruptions in agency) indicating that the sense of authorship and awareness is constructed and fallible. In response to these empirical challenges, the paper explores contemporary philosophical strategies aiming to preserve a refined notion of privileged access. The strategies includes redefining privilege as conceptual necessity (Shoemaker), distinguishing foundational pre-reflective self-consciousness (Zahavi), and locating authority within the normative framework of interpretation (Davidson, Sellars), or grounding it in the irreducibility of phenomenal consciousness (Chalmers). Ultimately, the argument concludes that while acknowledging empirical fallibility, philosophical responses defend the epistemic privilege, and argue that self-knowledge retains a distinctive, non-eliminable first-person character essential to subjectivity and rationality, even if its scope and infallibility are significantly constrained.

***Keywords:*** Privileged Access, Self-knowledge, Introspection

## 1. Introduction

A persistence challenge to self-knowledge and introspection which arise to the field of mind of philosophy is it’s special features so called privileged access. It seems to be very direct that we hold authoritative belief about our mental states. We clearly knows what’s inside of us, including our thoughts, sensations, and intentions. This phenomenon, often termed as "privileged access," suggests

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7064/2025.ND28720

a fundamental uniqueness: we have a direct first person path way about the knowledge of our own mind and that self-knowledge has a prior reliability to us, as opposed to the unstable and unsecured external worlds. These reliability, security, and special authority—often bond with the process so called 'introspective reports' — reinforce the intuition that self-knowledge occupies a privileged epistemic position.

Tradition philosophy of mind theory has lists distinct paths to explain and to justify this privileged access. The first is the epistemic security approach, which claims that the self-knowledge has a infallibility in the ideal conditions, such as introspective condition. And the direct awareness in that mental states is claimed to provide an indubitable foundation for knowledge, and through introspective condition can be immune to certain errors that blur the perception.The second is the introspection as a perceptual models. As the so called inner sense theory, introspection is a kind of internal perception that provide a reliable access to the mental states. The last is the rationalist perspective, it holds that our ideas and intentions are a kind of normative commitments, the self-knowledge arise from the rational action of the subject. Consequently, the self-knowledge have a practical authority, the ideas and intentions are not simply a perception state but a commitment to the actions.

However, these philosophical paths now face significant challenges from the findings of psychology and cognitive neuroscience. Psychological research consistently fins the systematic errors and limitations in self-knowledge. The main focus of the psychological perspective is the interpretivist theory of mind. The core of the theory is that it regards the self-knowledge is not a direct internal introspection, but is a interpretative and deductive process. The second is that it shows the existence of many undetectable mental processes beyond the consciousness states. People may hold an implicit attitude that they don't even aware of and may contradict with their explicit self-knowledge. In neural science, the forward models & sensory attenuation distinguish the boundaries of the self-create actions and the external actions. It shows the beliefs which regards a specific actions or feeling arise from the subject himself is noting but a deduction based on the predicted comparison, not a direct and undoubted access of authoritative knowledge. These psychological and scientific perceptive depict a picture of self-knowledge as fallible, limited in scope, and susceptible to significant error, starkly contrasting with claims of robust privileged access.

The responsive strategies of philosophy is redefining the meaning of privilege access to some specific form or specific area. Though the complete reliability can not be ensured, but the privilege access can still be confined into some certain objects. And also, by distinguishing the differentiate types and layers of self-knowledge, we can see that the privileged access can still be found in some of the most fundamental area in our consciousness.

This paper will proceed as follows. In part 1, the paper will provides a detailed analysis of the concept of privileged access and use dominated philosophical theories(Acquaintance, Inner Sense, Rationalist) to distinguish it's core epistemic claims to it's reason of existence. In part 2, the paper will suggests basic psychological and neuron-scientific theories to challenge the basic limitations about the privilege access. In part 3, the paper will presents the basic responsive strategies of the philosophy to the contradiction from the psychological and the neuron-scientific theories.

## 2. Philosophical foundation

In philosophical perspective, we can simply define the concept of privileged access of self-knowledge as a distinctive epistemic position a subject has regarding their current mental states. The paper will propose a few philosophical views to support the standpoint as follows.

DOI: 10.54254/2753-7064/2025.ND28720

According to the Inner Acquaintance Theory (IAT) [1] : the reason why a mental state of a subject has conscious is because of that the subject is being acquainted with mental state. The directness of the acquaintance includes independence of other external representations and no causal mediation. In its developed form (IAT4), consciousness arises when the subject's overall experience at a time enters a state of self-acquaintance. The experience (as the conscious state) is both the termination and the state of acquaintance. This relationship explains a unity where all conscious states at any time belong to experience, subjective character ("for-me-ness").

The Inner Sense Theory, as John Locke [2] and David Armstrong [3] claims, suggests that the self-knowledge arises from a quasi-perceptual mechanism analogous to external sense perception rather than a not a distinctive epistemic and metaphysical approach. Thereby establishing privileged access—where individuals have direct, immediate, and authoritative knowledge of their own mental states. Locke, in his Essay Concerning Human Understanding, laid the groundwork by describing introspection as the "inner sense" by which the mind perceives its own operations, while Armstrong, in A Materialist Theory of the Mind, processes this view to a Central-state Materialist framework. Armstrong argues that mental states are identical to physical brain states and that introspection involves a "self-scanning process" where the brain monitors its own states, like a so-called proprioceptive mechanism. It defines the core mechanism as an internal monitoring or scanning of neural activities that produces introspective awareness without reliance on external evidence. The inner sense process relies on causal relations: mental states, as brain states, causally generate introspective beliefs through reliable processes, such as neural feedback loops. At the same time, it depends on scientific laws rather than logical necessity, thus emphasizing that the mind's access to itself is highly reliable because of its direct, internal nature.

The theory presents such a strong support to the privileged access within the self-knowledge by constructing introspection as an internal scanning process. It accounts for the immediacy and authority of self-knowledge. Individuals can directly detect states like our beliefs without inference, thus reducing error potential compared to external perception, where sensory organs mediate and may distort information. In conclusion, the Inner Sense theory not only reinforces a materialist understanding of the mind but also defends privileged access. It effectively counters the critiques, such as those from Shoemaker [4], who questions the perceptual analogy by highlighting the absence of distinct introspective objects or organs. Consequently, it demonstrates how the theory's empirical and causal foundations provide an effective account of why self-knowledge feels direct and privileged.

Agency Theory presents another perspective that challenges the epistemic view of self-knowledge and asserts that privileged access originates not from particular detection mechanisms but from our active role as rational agents. The roles of ours as rational agents are responsible for our mental states, contrasting with the acquaintance theory that emphasizes direct conscious encounter. As Moran [5] and Burge [6] posits, the self-knowledge arises through normative agency rather than causal relations. Moran's "transparency method" explains this core mechanism by showing in order to answer the question"Do I believe P?". It requires considering reasons for P itself, not like an introspective scanning for belief-states [5]. This process is crucial due to it's constitutively linking to rationality, as Burge's transcendental argument presents:If one lacked entitlement to judgments about one's attitudes, one could not be subject to rational norms governing how one ought to alter those attitudes given that one had reflected on them [6]. This appears to be a version of "ought implies can" reasoning.

The theory's strength lies in explaining why self-knowledge access is uniquely immediate and non-observational: it grounds privileged access in the authorial relation we bear to our commitments

DOI: 10.54254/2753-7064/2025.ND28720

thus avoiding the explanatory gaps of causal-perceptual models. Agency theorists like McGeer [7] extend the framework through "commissive" self-ascriptions that actively shape attitudes, preserving privileged access even for non-reasons-sensitive states by tying knowledge to practical commitment rather than positive observation. Thus, by refining self-knowledge as an exercise of agency that ensuring one's mind is inseparable from authoring it. Consequently, the theory substantiates privileged access as a normative necessity of rational subjectivity and regard us not as spectators but as owners of our mental lives.

## 3. The psychological and neural-scientific challenges

The psychological view advanced by Nisbett and Wilson [8] and elaborated by Wegner [9] challenges the philosophical perspectives of privileged access to self-knowledge. It argues instead that understanding one's own mental states is an interpretive and inferential process functionally identical to how we attribute mental states to others. Nisbett and Wilson's seminal work challenges the assumption of direct introspective access. Trough demonstrating extensive experimental evidence—spanning cognitive dissonance studies, attribution research, and investigations into problem-solving and judgment—, the theory shows that individuals frequently lack accurate awareness of the actual causes of their judgments, choices, and behaviors. When people explaining their cognitive processes, people do not retrieve veridical records of internal operations but they construct reports based on a "priori causal theories" about the relations between stimuli and responses.

Wegner's theory of "apparent mental causation" further refines this model, it specifies that the inferential mechanisms: experiences of conscious will arise not from perceiving actual causation, but from inferences of these conditions, the"priority" (thought precedes action), the "consistency"(thought matches action), and the"exclusivity"(no salient alternative causes). This interpretive framework possesses significant expatiation of the deficit of privileged-access models. It reveals how inferences can diverge from underlying causal mechanisms.

Consequently, the experience of direct self-knowledge is revealed as a illusion generated by the mind's interpretive machinery, not privileged introspection. This mechanism reveals that knowledge of our own minds is constructed, not perceived.

The dual-process framework and implicit attitude research also challenges the philosophical notion of privileged access to self-knowledge by demonstrating that significant psychological processes operate outside conscious awareness. As foundational work by Greenwald and Banaji [10] established, our cognitive processes possess implicit modes of operation – defined as "introspectively unidentified (or inaccurately identified) traces of past experience" that influence judgment without the subject's awareness. This challenges the assumption that introspection provides privileged access to all mental content. Empirical evidence reveals dissociations: people often hold implicit attitudes (e.g., automatic negative associations toward social groups) that contradict their explicitly endorsed beliefs, as measured by tools like the Implicit Association Test (IAT). For instance, individuals explicitly disavowing prejudice may still exhibit robust implicit biases, indicating that self-report captures only a fraction of evaluative cognition.

The Associative-Propositional Evaluation (APE) model provides the core theoretical mechanism explaining this limitation of introspection. It posits two distinct systems: associative processes involving automatic activation of evaluations based on spatiotemporal contiguity or similarity (independent of subjective truth), and propositional processes involving syllogistic reasoning and validation of truth values.

DOI: 10.54254/2753-7064/2025.ND28720

Disorders of agency, such as thought insertion and delusions of control in schizophrenia, also challenges the philosophical concept of privileged access to self-knowledge by demonstrating that the sense of self-agency is not infallible. Frith [11] proposes a neurocognitive model where a comparator mechanism distinguishes between the sense of self-agency (feeling of initiating actions) and self-ownership (experience of undergoing them), with disruptions in this system—often due to prefrontal or supplementary motor area dysfunction—leading to symptoms where patients misattribute self-generated thoughts or movements to external sources.

Consequently, this framework not only explains clinical phenomena but also counters critiques by showing that privileged access is a construct vulnerable to neurological breakdowns, thereby affirming the interdependence of cognitive mechanisms and self-experience.

## 4. The responsive strategies of philosophy

Contemporary philosophical responses to psychological and neuroscientific challenges concerning privileged access to self-knowledge generally acknowledge the empirical evidence demonstrating significant fallibility (e.g., regarding implicit attitudes, emotional predictions, sense of agency). Instead of defending an infallible, transparent Cartesian model, philosophers seek to refine privileged access, recognizing its limitations while preserving a core epistemic approach.

Key strategies include redefining the nature of the privilege itself. Philosophers like Sydney Shoemaker argue for a conceptual necessity: while errors occur, a rational subject possessing the concept of belief cannot be fundamentally "self-blind" to their own occurrent, conscious beliefs in the way one can be blind to another's beliefs, suggesting a foundational difference in kind. Colin McGinn [12] contends the privilege lies primarily in the non-theoretical, non-inferential manner of access to conscious states like pain or basic thoughts, distinct from evidence-based third-person access, even if the resulting judgments can be inaccurate.

Another strategy involves distinguishing different levels of self-knowledge. Dan Zahavi [13], drawing on phenomenology, highlights a basic layer of pre-reflective self-consciousness—a minimal, immediate sense of self accompanying all experience—arguing that while reflective self-attributions are vulnerable to the empirical challenges, this foundational level of self-presence retains a directness that constitutes a form of privilege. Wilfrid Sellars [14] locates a normative authority for first-person reports (e.g., "I am in pain") within the manifest image of persons in ordinary discourse, seeing it as functionally distinct from the scientific image where fallibility is described.

A third approach grounds privileged access in fundamental aspects of mind or knowledge. Donald Davidson’s [15] radical interpretation argument posits that first-person authority is a necessary presupposition for the possibility of interpreting others' language and thought; empirical evidence of error only makes sense within this interpretive framework. David Chalmers [16] invokes the "hard problem" of consciousness, arguing that the irreducible "what-it’s-likeness" of subjective experience, accessible only via first-person acquaintance, demonstrates an epistemic gap that neurocorrelates cannot bridge, securing a unique role for first-person access to phenomenal states. In sum, the contemporary dialogue integrates empirical findings but defends various conceptions of mitigated epistemic privilege—whether rooted in conceptual necessity, distinctive modes of access, foundational layers of selfhood, or presuppositions of understanding—arguing that fallibility itself presupposes and exists against a backdrop of privileged first-person perspective.

DOI: 10.54254/2753-7064/2025.ND28720

## 5. Conclusion

In conclusion, this examination of privileged access to self-knowledge reveals a complex landscape where traditional philosophical justifications—grounded in epistemic security (Acquaintance Theory), quasi-perceptual mechanisms (Inner Sense Theory), and normative agency (Agency Theory)—are compellingly challenged by empirical research demonstrating the inferential nature of self-attribution (Nisbett & Wilson, Wegner), the pervasive influence of implicit cognition (Greenwald & Banaji), and the neurological fragility of self-agency (Frith).

Rather than abandoning the notion of privilege entirely, contemporary philosophical responses strategically refine it: Shoemaker and McGinn recast it as conceptual necessity or distinctive non-inferential access; Zahavi and Sellars locate it in foundational layers of pre-reflective consciousness or normative reporting practices; Davidson and Chalmers anchor it in the indispensable presuppositions of interpretation or the irreducibility of phenomenal experience. Thus, while psychological and neuroscientific evidence decisively undermines claims of infallible, transparent self-knowledge across all domains, it does not eliminate a mitigated form of epistemic privilege—one that persists as a constitutive feature of rational subjectivity, a functional necessity for self-interpretation, and an irreducible dimension of conscious life, ensuring that the first-person perspective remains uniquely authoritative within its circumscribed, fallible scope.

## References

[1] Giustina, A. (2024). Inner acquaintance theories of consciousness. Oxford Studies in Philosophy of Mind, 4. Oxford University Press.
[2] Locke, J. (1689/1975). An Essay Concerning Human Understanding, P.H. Nidditch (ed.), Oxford: Oxford University Press.
[3] Armstrong, D. M. (1968). A materialist theory of the mind. Routledge & Kegan Paul.
[4] Shoemaker, S. (1994). "Self-knowledge and 'Inner Sense' Lecture I: The Object Perception Model." Philosophy and Phenomenological Research, 54(2), 249–269.
[5] Moran, R. (2003). Responses to O'Brien and Shoemaker. European Journal of Philosophy, 11(3), 402–419.
[6] Burge, T., & Peacocke, C. (1995). Our entitlement to self-knowledge. Aristotelian Society Supplementary Volume, 69(1), 255–255.
[7] McGeer, V. (1996). Is "self-knowledge" an empirical problem? Renegotiating the space of philosophical explanation. The Journal of Philosophy, 93(10), 483–515.
[8] Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84(3), 231–259.
[9] Wegner, D. M. (2003). The mind's best trick: How we experience conscious will. Trends in Cognitive Sciences, 7(2), 65–69.
[10] Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. Psychological Review, 102(1), 4–27.
[11] Frith, C. D. (1992). Consciousness, information processing, and the brain. Journal of Psychopharmacology, 6, 436–440.
[12] McGinn, C. (1982). "Anomalous Monism and Kripke's Cartesian Intuitions." Analysis, 42(1), 47-49.
[13] Zahavi, D. (2005). Subjectivity and Selfhood: Investigating the First-Person Perspective. MIT Press.
[14] Sellars, W. (1956). Empiricism and the Philosophy of Mind. Harvard University Press.
[15] Davidson, D. (1984). "First Person Authority." Dialectica, 38(2/3), 101–111.
[16] Chalmers, D. J. (1996). The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press.