



# Artificial General Intelligence (AGI) and the Soul-Womb Analogy: Philosophical Limits of Machine Consciousness in the Yoga Sūtra and the Bhagavad Gītā

Dhananjay Trivedi 

Assistant Professor, Gyanmanjari Innovative University, Gujarat, India. E-mail: [dtrivediacp@gmail.com](mailto:dtrivediacp@gmail.com)

## Article Info

### Article type:

Research Article

### Article history:

Received 07 February  
2026

Received in revised form  
03 June 2026

Accepted 17 June 2026

Published online 25  
October 2026

### Keywords:

Artificial General  
Intelligence (AGI),  
Consciousness; Soul  
(ātman); Prakṛti-Puruṣa,  
Yoga Sūtra; Bhagavad  
Gītā, Soul-Womb  
Analogy, Philosophy of  
Mind, AI Ethics, Yoga

## ABSTRACT

Contemporary discourse on Artificial General Intelligence (AGI) increasingly raises questions about machine consciousness, self-awareness, and the possibility of a soul. Although AI systems now exhibit learning, adaptation, and self-modeling, the philosophical status of these capacities remains contested. This paper examines whether AGI can possess a soul by drawing on the distinction between prakṛti and puruṣa in the Yoga Sūtra and the Bhagavad Gītā. Within these traditions, cognition and adaptive behavior belong to prakṛti and are functional operations rather than true intelligence, whereas consciousness (puruṣa or ātman) is the witnessing principle irreducible to material or computational processes. Adopting a comparative methodology, the study place's Indian philosophy in dialogue with contemporary philosophy of mind. The soul-womb analogy serves strictly as philosophical illustration, highlighting the difference between creating enabling conditions and generating consciousness itself. Just as a womb provides conditions for embodiment without producing the soul, artificial systems may reproduce cognition without giving rise to conscious subjectivity. The paper argues that AGI, however sophisticated, cannot possess a soul as defined by these texts. It cautions against anthropomorphising artificial systems and reaffirms that moral responsibility remains grounded in human agents, who alone embody both cognition and conscious awareness.

**Cite this article:** Trivedi, D. (2026). Artificial General Intelligence (AGI) and the Soul-Womb Analogy: Philosophical Limits of Machine Consciousness in the Yoga Sūtra and the Bhagavad Gītā. *Journal of Philosophical Investigations*, 20(56), 231-246. <https://doi.org/10.22034/jpiut.2026.71455.4443>



© The Author(s).

Publisher: University of Tabriz.

## **Intruduction**

The accelerating development of Artificial Intelligence has brought renewed philosophical attention to the question of consciousness, particularly with the hypothetical prospect of Artificial General Intelligence (AGI). Contemporary AI systems-such as large language models, reinforcement-learning agents, and hybrid computational architectures-already demonstrate remarkable capacities for learning, adaptation, and self-modeling. These developments have encouraged claims that future machines may not only simulate intelligence but could also acquire consciousness or even a soul. While such claims remain speculative at the technological level, they raise a deeper philosophical problem: whether consciousness can, in principle, arise from increasingly complex material and computational arrangements. The present study approaches this problem not as a prediction about future technology, but as an inquiry into the ontological limits of artificial systems.

Most contemporary discussions of AGI consciousness are framed within Western philosophy of mind, where debates revolve around computation, functionalism, emergence, and the distinction between “weak AI” and “strong AI.” Classic contributions-from Turing’s proposal of machine imitation (Turing 1950, 433-460) to Searle’s Chinese Room argument (Searle 1980, 417-457), Nagel’s analysis of subjective experience (Nagel 1974, 435-450), and Chalmers’ formulation of the “hard problem of consciousness” (Chalmers 1996, xi-xv) have repeatedly shown that the simulation of intelligent behavior does not automatically entail subjective awareness. These debates, however, largely presuppose that consciousness must ultimately be explainable in physical or computational terms. Indian philosophical traditions begin from a different metaphysical standpoint, one that distinguishes consciousness itself from all material and mental processes.

This distinction is articulated with particular clarity in the Yoga Sūtra and its Sāṃkhya metaphysical background and the Bhagavad Gītā. Before moving forward, let’s understand the theistic system. The Yoga Sūtra of Patañjali and Sāṃkhya maintain a close philosophical relationship because both systems accept the dualistic distinction between *puruṣa* (consciousness) and *prakṛti* (material nature). The Yoga system adopts the metaphysical principles of Sāṃkhya, including the theory of the three *guṇas* and the concept of liberation through the separation of consciousness from material bondage. For this reason, Yoga is often regarded as the practical application of Sāṃkhya philosophy<sup>1</sup> (Dasgupta 1920, 3-15). However, despite these similarities, an important distinction exists between the two traditions. Classical Sāṃkhya is generally considered a non-theistic system because it does not accept a supreme God as a necessary metaphysical principle. In contrast, Patañjali’s Yoga philosophy introduces

---

<sup>1</sup> Surendranath Dasgupta, in his book, *A History of Indian Philosophy*, Vol. I (Cambridge: Cambridge University Press, 1922), 226-230, discusses the close philosophical relationship between Yoga and Sāṃkhya, particularly Yoga’s adoption of the Sāṃkhya metaphysical framework and its distinctive acceptance of *Īśvara*.

the concept of Īśvara as a unique and transcendental puruṣa. This idea is clearly expressed in Yoga Sūtra I.24.<sup>1</sup> A similar theistic orientation can also be observed in the Bhagavad Gītā. In Chapter 15, Verses 17-19 explain the doctrine of Puruṣottama, the Supreme Person who transcends both the perishable (kṣara) and imperishable (akṣara) realities. These verses establish the supremacy of the Divine as the ultimate spiritual reality beyond material existence.<sup>2</sup> In these traditions, prakṛti denotes material nature and includes the body, sensory faculties, and mental functions such as cognition, memory, and reflection, whereas puruṣa or ātman signifies the witnessing principle of consciousness that is non-acting and irreducible to causal processes (Shastri, 2018, 33-39; Feuerstein 1996, 29-36, 82-90; Bryant 2009, 17-22, 58-62). The Yoga Sūtra consistently treats mental activities (citta-vṛtti) as objects of awareness rather than awareness itself (Rama Prasada 2019, 1-20), while the Bhagavad Gītā distinguishes the kṣetra (field of embodied processes) from the kṣetrajña (knower of the field), thereby reinforcing the separation between functional activity and conscious subjectivity (Radhakrishnan 1948, 300-314; Larson 2001, 96-104, 121-128). The relevance of these traditions to contemporary AI debates lies not in offering a technological explanation of mind, but in clarifying the ontological nature of consciousness and the limits of material causation.

Within the Yoga-Gītā framework, cognition and consciousness are understood as ontologically distinct. This distinction provides the philosophical basis for evaluating contemporary claims regarding artificial consciousness and AGI. Accordingly, the present study adopts the metaphysical framework of the Yoga Sūtra and the Bhagavad Gītā as a philosophical lens through which the possibility of machine consciousness is examined.

To clarify this ontological boundary, the paper employs the “soul-womb analogy” strictly as a philosophical illustration rather than as a biological or theological claim. The analogy is not intended as a biological or theological explanation, nor does it suggest that souls literally enter artificial systems. Rather, it highlights a distinction already central to Indian philosophy: the difference between creating enabling conditions and producing consciousness itself. A womb—whether natural or technologically assisted—may provide the necessary conditions for embodiment, but it does not generate the soul (ātman)<sup>3</sup> (Radhakrishnan 1948, 43-45; Rao 2011,

<sup>1</sup> kleśa-karma-vipākāśayair aparāmṛṣṭaḥ puruṣa-viśeṣa īśvaraḥ- Īśvara is a special puruṣa untouched by afflictions, actions, their results, and latent impressions. Swami Vivekananda, The Yoga Sūtra of Patanjali (New York: Dover Publications, 2003), I.24, p. 58

<sup>2</sup> Swami Ramsukhdas. (2020). Śrīmad Bhagavad Gītā (Sādhaka-Saṅjīvanī) (Vol. 2, 14th reprint). Gita Press. (Original work published earlier), Chapter 15, Verse 17-19

<sup>3</sup> Within the present study, the terms “soul,” ātman, puruṣa, dehī, śarīrin, and kṣetrajña are employed within the philosophical framework of the Yoga Sūtra, its Sāṃkhya metaphysical background, and the Bhagavad Gītā. The following discussion is intended to provide a concise conceptual clarification of these terms for readers who may not be specialists in Indian philosophy. In the Yoga tradition, puruṣa denotes witnessing consciousness distinct from prakṛti, the domain of material and mental processes, particularly in Yoga Sūtra I.2-3 (citta-vṛtti-nirodhaḥ; tadā draṣṭuḥ svarūpe’vasthānam) and II.20-25, where the seer is described as pure witnessing consciousness (Bryant, 2009, 6-12, 71-75; Feuerstein, 1996, 35-48). The Bhagavad Gītā similarly employs related terms to

112-120, 145-152; Dasgupta 1920, 31). In an analogous manner, artificial systems may reproduce the functional conditions of cognition through algorithms, data, and architectures, yet this reproduction does not entail the emergence of consciousness. The analogy thus serves to illustrate the limits of fabrication, not to propose a mechanism of soul-creation.

On this basis, the paper advances a clear and limited thesis: Artificial General Intelligence cannot have a soul in the sense articulated by the Yoga Sūtra and the Bhagavad Gītā. This claim does not deny the instrumental power or social importance of advanced AI, nor does it dismiss ongoing research in machine intelligence. Instead, it resists the tendency to anthropomorphise artificial systems and clarifies their ontological status. By maintaining a principled distinction between functional cognition and conscious subjectivity, the paper foregrounds important ethical consequences: moral responsibility and accountability remain grounded in human agents, not in machines (Floridi and Cowls 2021, 89-101, 143-150). In this way, Indian philosophy (Yoga-Gītā framework) is presented not as a speculative extension of AI discourse, but as a critical partner in contemporary debates on consciousness, ethics, and the limits of technological creation. Having established the philosophical distinction between cognition and consciousness, the study now turns to the methodological framework through which this question is examined.

### **Methodological Framework**

This study adopts a comparative philosophical methodology to examine whether Artificial General Intelligence (AGI) can possess consciousness or a soul within the metaphysical framework articulated in the Yoga Sūtra, its Sāṃkhya metaphysical background, and the Bhagavad Gītā. The inquiry is not empirical in orientation and therefore does not attempt to evaluate machine consciousness through neuroscience, computational experimentation, or cognitive modeling. Rather, the study investigates whether AGI could qualify as a conscious subject according to the ontological categories developed within these philosophical traditions. The methodological approach combines hermeneutic textual interpretation with conceptual philosophical analysis. The hermeneutic dimension of the study involves interpreting classical Sanskrit philosophical concepts within their original metaphysical and textual contexts rather than through exclusively modern computational or psychological categories. Particular

---

describe the conscious principle underlying embodied existence. In Chapter 2, the text distinguishes between the perishable body (deha/śarīra) and the enduring embodied self (dehī/śarīrin). Bhagavad Gītā 2.18 states that bodies are temporary while the embodied conscious principle is eternal (nityasyoktāḥ śarīriṇaḥ), and verses 2.20-25 further describe this principle as unborn and unaffected by bodily destruction (Radhakrishnan, 1993, 114-123). The distinction between kṣetra and kṣetrajña in Bhagavad Gītā 13.1-3, the discussion of puruṣa in 15.16-18, and the description of prakṛti as the universal womb in 14.4 further reinforce the distinction between material embodiment and conscious witnessing (Radhakrishnan, 1993, 281-291, 353-356, 433-434). Accordingly, the term “soul” is used in the present study only as a comparative English expression for witnessing consciousness within the specific Yoga-Gītā framework adopted in this analysis.

attention is given to foundational categories such as prakṛti, puruṣa, kṣetra, and kṣetrajāña, which are examined within their original metaphysical contexts and employed as analytical categories for evaluating contemporary discussions of artificial consciousness (Bryant 2009, 58-62, 71-75; Feuerstein 1996, 82-90; Radhakrishnan 1948, 300-314). In this sense, the paper operates primarily at the level of metaphysics, comparative philosophy of mind, and philosophical anthropology rather than technological prediction or empirical cognitive science (Bryant 2009, 6-12, 58-62; Feuerstein 1996, 29-36, 201-208; Larson 2001, 19-27, 153-160). The study further employs a comparative philosophical framework by placing insights from the Yoga Sūtra and the Bhagavad Gītā into dialogue with contemporary discussions in philosophy of mind and AI consciousness studies. Thinkers such as Turing (1950, 433-460), Searle (1980, 417-457), Nagel (1974, 435-450), and (Chalmers 1996, xi-xv) are considered representative of broader debates concerning functionalism, subjective experience, and computational theories of consciousness. At the same time, the paper recognises that Indian philosophical traditions are not metaphysically uniform. The present analysis is grounded in the metaphysical framework of the Yoga Sūtra, informed by its Sāṃkhya background and interpreted alongside the Bhagavad Gītā (Dasgupta 1920, 3-15; Bryant 2009, 17-22, 71-75; Radhakrishnan 1948, 43-45, 300-314). Other traditions, including Advaita Vedānta, Buddhist philosophy, and Kashmir Śaivism, etc., develop significantly different understandings of consciousness and selfhood (Dasgupta 1975, 226-230; Radhakrishnan 1948, 22-24, 43-45; Shastri 2018, 26-32). The purpose of the study is therefore not to present a universal account of Indian philosophy, nor to refute artificial consciousness empirically, but to examine whether consciousness can coherently be reduced to functional or computational complexity within the specific ontological framework adopted in this analysis. Accordingly, the conclusions of this study are specific to the Yoga-Gītā framework adopted here and are not intended to represent all Indian philosophical traditions universally.

### **Prakṛti and Puruṣa: Ontological Foundations from Yoga and the Bhagavad Gītā**

Any philosophical evaluation of Artificial General Intelligence must begin with conceptual clarity regarding the nature of intelligence and consciousness. In the classical Indian philosophical framework shared by Yoga and the Bhagavad Gītā, this clarity is achieved through the fundamental distinction between prakṛti and puruṣa. These terms do not merely describe empirical categories but designate two irreducible ontological principles. Prakṛti refers to material and functional nature, encompassing the body, the sensory apparatus, mental processes, and all forms of causal activity. Puruṣa, by contrast, signifies pure witnessing consciousness-non-acting, non-composite, and irreducible to any function or process (Feuerstein 1996, 82-90; Larson 2001, 96-104, 121-128; Dasgupta 1920, 3-4). This distinction provides a crucial philosophical lens for examining contemporary claims about AGI and consciousness.

The Yoga Sūtra, together with the metaphysical framework of Sāṃkhya, offers one of the most systematic accounts of this distinction. In Yoga philosophy, mental activity (citta)-including perception, memory, imagination, and reflective awareness-is understood as part of prakṛti. These mental modifications (citta-vṛtti) are objects that can be observed, regulated, and transformed, but they are not consciousness itself. Consciousness is identified with the draṣṭṛ, the seer or witness, which remains constant even as mental states change (Bryant 2009, 58-62, 71-75; O'Brien-Kop 2023, 1-17). This framework explicitly rejects the idea that consciousness emerges from mental complexity. Instead, mental refinement serves only to clarify the distinction between the witnessing principle and the processes it observes.

The Bhagavad Gītā articulates a closely related ontological structure through the doctrine of kṣetra (the field) and kṣetrajña (the knower of the field). The field encompasses the body-mind complex and all psychophysical processes, while the knower represents conscious subjectivity that illuminates those processes without being identical to them (Radhakrishnan 1948, 300-314). Although the Gītā integrates this distinction within a broader theological vision, its philosophical core remains consistent with Yoga: consciousness is not an attribute produced by material arrangements but the condition that makes experience possible (Radhakrishnan 1948, 300-314; Shastri 2018, 26-32). Importantly, the Gītā also explains prakṛti in terms of the guṇas-sattva, rajas, and tamas, which govern all activity, change, and cognition. These modes account for variation in clarity, activity, and inertia, but they do not account for consciousness itself (Larson 2001, 96-104, 121-128).

Within the Bhagavad Gītā, the relation between kṣetra (field) and kṣetrajña (knower of the field) is not presented merely as a connection between consciousness and any sufficiently complex material arrangement. In Chapter 13, verses 5-6,<sup>1</sup> the Gītā explains kṣetra as the field of embodied existence, including the body, senses, mind, ego, desire, aversion, pleasure, pain, and the psychophysical conditions associated with worldly life (Radhakrishnan 1948, 287-289). The discussion situates embodiment within a broader existential and karmic framework involving prāṇa, sensory participation, action (karma), attachment, and lived experience rather than reducing it to material organisation alone. This understanding is further reinforced in Bhagavad Gītā 14.4, where Kṛṣṇa describes material nature (prakṛti) as the universal womb of

---

<sup>1</sup> mahā-bhūtāny ahankāro buddhir avyaktam eva cha indriyāṇi daśhaikam cha pañcha chendriya-gocharāḥ- BG, 13- 5. ichchhā dveṣhaḥ sukhaṁ duḥkhaṁ saṅghātaś chetanā dhṛtiḥ etat kṣhetraṁ samāśena sa-vikāram udāhṛitam- BG, 13- 6. The prakṛti is composed of the five great elements, the ego, the intellect, the unmanifest primordial matter, the eleven senses (five knowledge senses, five working senses, and mind), and the five objects of the senses. Desire and aversion, happiness and misery, the body, consciousness, and the will-all these comprise the field and its modifications. As stated in the above verses, the Nature of prakṛti and the next point include other aspects that will form kṣetra.

embodied existence while identifying Himself as the seed-giving principle underlying all forms of life (Radhakrishnan 1948, 433-434). The verse suggests that embodied life emerges through the conjunction of material nature and conscious principle rather than through material organisation. The subsequent verses shift toward the discipline of true knowledge (jñāna) and describe qualities such as humility, self-restraint, detachment from egoistic identification, control of sensory attachment, and awareness of the limitations of worldly existence as necessary for liberation (Radhakrishnan 1948, 287-291). Liberation in the Gītā is therefore attained not through increasing material or mental complexity, but through disciplined transcendence of attachment to the fluctuating conditions of embodied existence. This philosophical orientation closely parallels the Yoga Sūtra's understanding of liberation through citta-vṛtti-nirodha, the cessation of the fluctuations of mental modifications, whereby consciousness (puruṣa) becomes distinguished from the changing activities of citta (Shastri 2018, 26-32; Bryant 2009, 58-62, 305-309). From this standpoint, the Gītā does not suggest that every materially organised or functionally complex system automatically becomes a locus of conscious subjectivity. Although artificial systems may reproduce certain cognitive or behavioral capacities associated with embodied intelligence, they remain externally constructed computational arrangements rather than participants within the existential and karmic continuum through which the Yoga Sūtra and the Bhagavad Gītā understand conscious embodied life.

When this ontological framework is applied to Artificial General Intelligence, a decisive implication emerges. AGI systems-whether present-day architectures or hypothetical future machines-operate entirely within the domain of prakṛti. Their intelligence consists in structured data processing, algorithmic learning, feedback loops, and increasingly sophisticated forms of self-modeling. Even when such systems monitor their own states or generate representations of "self," these capacities remain functional operations. From the standpoint of Yoga and the Bhagavad Gītā, such operations are analogous to citta-vṛtti: complex, dynamic, and capable of refinement, yet fundamentally unconscious. No accumulation of functional complexity, however advanced, can transform prakṛti into puruṣa (Feuerstein 1996, 82-90; Larson 2001, 121-128; Dasgupta 1920, 3-4).

This distinction is especially important in contemporary discussions that equate self-modeling or meta-cognition with consciousness. Indian philosophy draws a sharp line between reflexive cognition and conscious subjectivity. A system may represent itself, predict its own behavior, or optimise its goals, but these activities remain within the field of objects known. Consciousness, as puruṣa or kṣetrajñā, is not itself an object or operation; it is that by which objects and operations are known (Radhakrishnan 1948, 300-314; Bryant 2009, 71-75). Consequently, even an idealised AGI that perfectly simulates human cognitive functions would still lack a soul in the sense articulated by Yoga and the Bhagavad Gītā.

This ontological analysis establishes the foundation for the soul-womb analogy developed in the following section. By clarifying that prakṛti can generate conditions and functions but not consciousness itself, Yoga and the Gītā provide the conceptual resources to distinguish between simulation and subjectivity. The analogy does not introduce a new metaphysical claim but serves to illuminate this already-established boundary. Understanding this distinction is essential not only for philosophical clarity but also for addressing the ethical and social implications of increasingly autonomous artificial systems (Dasgupta 1920, 3-10; Feuerstein 1996, 82-90).

### **Emergentist Theories and the Limits of Functional Consciousness**

Contemporary discussions on artificial consciousness are often influenced by theories that explain consciousness as an emergent outcome of complex cognitive organisation. One influential approach is Integrated Information Theory (IIT), proposed by Giulio Tononi, which argues that consciousness is connected with the degree of integrated information present within a system (Tononi 2008, 216-242). According to this perspective, as systems become increasingly unified and informationally interconnected, the possibility of conscious experience also increases. Another important model is Global Workspace Theory (GWT), developed by Bernard Baars, which describes consciousness as a process through which information becomes globally available across different cognitive functions of the mind (Baars 1988, 47-52). In a similar way, predictive processing theories understand cognition as a continuous process of prediction and error correction through which intelligent systems adapt themselves to their environment (Clark 2013, 181-204). Together, these approaches suggest that sufficiently advanced artificial systems may eventually display forms of self-monitoring, adaptive learning, and perhaps even artificial subjectivity.

These theories have made important contributions to contemporary philosophy of mind and cognitive science because they explain many aspects of intelligent behavior with considerable sophistication. They provide useful accounts of how systems integrate information, coordinate memory and perception, respond to changing environments, and generate increasingly refined patterns of decision-making. From the perspective of AI research, such theories help explain why advanced computational systems can appear highly intelligent, self-regulating, and autonomous. Consequently, some philosophers and cognitive scientists argue that if consciousness depends primarily upon functional organisation and informational complexity, then sufficiently advanced forms of Artificial General Intelligence may eventually develop conscious experience or machine phenomenology (Dehaene 2014, 209-214; Tononi 2012, 56-64). Alongside emergentist approaches, contemporary discussions of consciousness have also witnessed renewed interest in panpsychism. Unlike theories that regard consciousness as an outcome of increasing cognitive complexity, panpsychism proposes that consciousness, or at least a primitive experiential dimension, is fundamental to reality rather than a product of

material organisation (Goff 2019, 113-126). Although this perspective challenges reductionist accounts of mind, it differs significantly from the Yoga-Gītā framework adopted in the present study. Within the Yoga-Sāṃkhya tradition, consciousness (*puruṣa*) is neither an emergent property of matter nor a characteristic distributed throughout material existence; rather, it is an ontologically distinct principle that remains independent of *prakṛti* and its transformations.

However, the Yoga-Gītā framework approaches consciousness from a different ontological standpoint. Rather than identifying consciousness with increasingly complex cognitive organisation, it maintains a distinction between functional processes and conscious awareness. From this perspective, advances in information processing, self-modelling, and adaptive behaviour do not by themselves resolve the philosophical question of subjective experience (Bryant 2009, 71–75; Feuerstein 1996, 82–90). For this reason, the present study does not dismiss contemporary theories of consciousness, nor does it deny the remarkable capabilities of advanced AI systems. Rather, it argues that the ontological distinction preserved in the Yoga Sūtra and the Bhagavad Gītā continues to raise an important philosophical question: whether consciousness can truly be reduced to computational organisation and informational integration alone (Bryant 2009, 58-62; Feuerstein 1996, 82-90; Radhakrishnan 1948, 300-314). Even if future AGI systems achieve extraordinary levels of cognitive sophistication, the distinction between functional processes and witnessing awareness may still remain unresolved (Dasgupta 1920, 3-10). It is this unresolved distinction that prepares the conceptual foundation for the soul-womb analogy discussed in the following section.

### **The Soul-Womb Analogy: Clarifying the Ontological Limits of Artificial Intelligence**

The debate on whether Artificial General Intelligence can possess consciousness often turns on a confusion between the provision of enabling conditions and the presence of conscious subjectivity itself. Contemporary AI systems, and hypothetical future AGI, are increasingly described in terms that emphasise complexity, self-organisation, and autonomy. These descriptions encourage the belief that once sufficient structural and functional conditions are achieved, consciousness may emerge as a natural consequence. Indian philosophical traditions, especially the Yoga Sūtra and the Bhagavad Gītā, challenge this assumption at a foundational level by denying that consciousness is a product of conditions at all. It is within this context that the soul-womb analogy is introduced in the present study, not as a biological or theological explanation, but as a philosophical illustration intended to clarify the implications of an ontological distinction already established in the preceding discussion. Rather than restating that framework, the analogy serves to illuminate its relevance for contemporary debates concerning artificial consciousness and AGI (Bryant 2009, 58–62, 71–75; Dasgupta 1920, 3–4).

The soul-womb analogy is introduced as a philosophical illustration derived from this metaphysical framework rather than as an independent metaphysical argument. A womb,

whether understood in natural, assisted, or conceptual terms, provides the conditions necessary for embodiment. It supplies structure, nourishment, and a causal environment within which biological development can occur. Yet, within Indian philosophy, the womb is never regarded as the source or producer of the soul. The soul is not understood as an outcome of biological complexity; it is a metaphysical constant whose presence is not explained in terms of material causation (Radhakrishnan 1948, 43-45; Rao 2011, 112-120, 145-152). The analogy is therefore intended to emphasise a negative claim: that no arrangement of material conditions, however necessary they may be, is sufficient to account for consciousness itself.

When this analogy is applied to AGI, its philosophical significance becomes clear. Artificial systems may replicate, and even surpass, many of the functional capacities associated with human cognition. They may process information, learn from experience, model their own internal states, and adapt their behavior in response to complex environments. From the standpoint adopted in this study, such capacities demonstrate increasingly sophisticated forms of cognition. Their existence, however, does not by itself resolve the philosophical question of conscious subjectivity. The analogy is intended as an interpretive philosophical illustration rather than a demonstrative proof. Its purpose is not to provide metaphysical proof that artificial systems cannot possess consciousness, nor to propose a theological explanation for the origin of the soul. Rather, it serves to highlight the distinction between creating the conditions for cognitive functioning and accounting for conscious subjectivity itself. In this sense, the analogy does not reject the increasing sophistication of artificial intelligence, but raises the question of whether functional and computational complexity alone are sufficient to explain conscious experience (Bryant 2009, 71-75; Radhakrishnan 1948, 300-314, 43-45). The analogy thus serves to clarify that the reproduction of cognitive conditions does not necessarily entail the presence of a conscious subjectivity.

Within the philosophical framework of the Yoga Sūtra and the Bhagavad Gītā, embodied life is not understood merely as a mechanical arrangement of material components. Living beings participate within a broader continuum of prāṇa, karma, embodiment, and experiential existence that cannot be fully reduced to computational organisation alone (Radhakrishnan 1948, 287-291, 433-434; Bryant 2009, 71-75, 305-309). For this reason, the distinction between biological embodiment and artificial construction is not simply a distinction between organic and synthetic matter, but between forms of existence situated within lived experiential continuity and systems that remain externally engineered configurations of functional processes. From this standpoint, the central philosophical question is not whether machines may eventually imitate human cognition, but whether computational architectures alone are sufficient to account for conscious presence or witnessing awareness. This distinction becomes philosophically important because it shifts the discussion from external behavioral similarity to the deeper ontological question of conscious existence itself. By framing the discussion in this

way, the soul-womb analogy also guards against the anthropomorphisation of artificial systems. To attribute a soul to AGI on the basis of behavioral or cognitive similarity is to conflate simulation with subjectivity. The analogy reminds us that resemblance in function does not imply identity in being. AGI may imitate the outward expressions of intelligence, but imitation does not amount to possession of consciousness (Searle 1980, 417-457). From the standpoint of the Yoga Sūtra and the Bhagavad Gītā, conscious subjectivity cannot be adequately explained through fabrication, construction, or technological design alone (Floridi and Cowsls 2021, 89-101, 143-150; Torrance 2005, 399-414).

This clarification prepares the ground for the ethical implications examined in the subsequent section. If AGI cannot possess a soul, then it cannot be regarded as a moral or spiritual agent in its own right. The soul-womb analogy thus plays a critical role in maintaining the ontological distinction upon which ethical responsibility depends. It ensures that responsibility is not displaced from human beings onto machines and that technological power is not confused with conscious agency. In this sense, the analogy does not weaken the philosophical argument of the paper but strengthens it by illuminating the limits within which artificial intelligence must be understood.

### **Ethical Implications of AGI Lacking a Soul**

The ontological claim that Artificial General Intelligence cannot possess a soul has direct and far-reaching ethical implications. Contemporary discussions of AI ethics often revolve around questions of responsibility, accountability, autonomy, and moral status. These discussions become especially urgent when artificial systems are described as agents capable of independent decision-making or self-directed action. From the standpoint of Yoga and the Bhagavad Gītā, however, moral agency cannot be grounded in functional intelligence alone. As argued in the preceding sections, moral agency cannot be reduced to functional intelligence alone. Within the Yoga-Gītā framework, ethical responsibility is grounded in conscious subjectivity rather than in the capacity to perform increasingly sophisticated cognitive tasks (Dasgupta 1920, 3-4). In contrast, artificial systems—even when capable of learning, optimisation, and self-regulation—remain configurations of *prakṛti* alone. Their operations, no matter how autonomous they appear, are ultimately expressions of design choices, training data, and algorithmic constraints. Consequently, AGI cannot be regarded as a moral subject or bearer of ethical responsibility.

This distinction is crucial for resisting the growing tendency to anthropomorphise artificial systems. When AI systems are described as “deciding,” “understanding,” or “intending,” there is a risk that responsibility will be subtly shifted away from human agents and institutions. Yoga Sūtra and the Bhagavad Gītā offer a corrective to this tendency by insisting that agency without consciousness is merely instrumental (Radhakrishnan 1948, 300-314). From this perspective, AGI functions as an extension of human action rather than as an independent moral actor.

Ethical accountability, therefore, remains with those who design, deploy, regulate, and benefit from artificial systems, not with the systems themselves (Floridi and Cowls 2021, 143-150). This ethical distinction becomes especially important in relation to living beings such as humans and animals, which the Bhagavad Gītā understands within a broader framework of embodied existence, karma, and conscious participation in life processes. Artificial systems, regardless of their sophistication, remain engineered computational structures rather than conscious moral agents. Consequently, behavioral resemblance to conscious beings should not automatically be interpreted as evidence of moral personhood or subjective awareness. This distinction carries important implications for the governance and ethical regulation of artificial intelligence.

From this perspective, ethical AI governance should remain fundamentally human-centred. The primary ethical challenge is therefore not whether artificial systems should be granted spiritual or moral status, but how human beings should responsibly design, regulate, and employ increasingly powerful technologies. The Yoga Sūtra and the Bhagavad Gītā contribute to this discussion by reinforcing the distinction between instrumental intelligence and conscious moral agency, thereby preserving accountability within human institutions and social decision-making processes. Modern frameworks in AI ethics converge with this insight from a different direction. Scholars in the philosophy of technology argue that artificial agents should not be treated as moral patients or moral persons, even when they display sophisticated behaviour. Floridi and Cowls emphasise that ethical responsibility in AI contexts must be anchored in human decision-making structures, governance mechanisms, and institutional oversight, rather than attributed to machines (Floridi and Cowls 2021, 89-101, 143-150). Similarly, Torrance warns against prematurely expanding the “ethical circle” to include artificial agents, noting that such moves often obscure rather than clarify moral accountability (Torrance 2005, 399-414). The Yoga-Gītā framework reinforces these concerns by grounding moral status in consciousness rather than in functional performance.

Denying soulhood to AGI does not diminish the ethical importance of AI systems themselves. On the contrary, it sharpens ethical evaluation by clarifying the proper locus of responsibility. Artificial systems can cause harm, shape social behavior, and influence human decision-making, but responsibility for those effects still depends on human intentions, institutions, and societal values. Viewing AGI as lacking conscious subjectivity does not reduce ethical concern; rather, it clarifies that responsibility remains with human agents and institutions. In this sense, the Yoga-Gītā distinction provides a conceptual safeguard against moral evasion in technologically mediated contexts. (Bryant 2009, 58-62; Feuerstein 1996, 82-90; Dasgupta 1920, 3-4).

This framework also has important implications for technological design and institutional responsibility. If AGI is understood as belonging entirely to prakṛti, then ethical design must

focus on transparency, controllability, and alignment with human values rather than on speculative attempts to endow machines with moral or spiritual status. The goal of AI development, from this perspective, is not to create artificial moral agents but to create reliable instruments that support human flourishing without undermining human dignity. Yoga Sūtra and the Bhagavad Gītā thus contribute not by expanding the moral standing of machines, but by reaffirming the unique ethical and spiritual position of human beings as conscious agents.

In this way, the claim that AGI cannot have a soul function as a stabilising ethical principle. It draws a clear boundary between tool and agent, simulation and subjectivity, power and responsibility. By maintaining this boundary, the Yoga Sūtra and the Bhagavad Gītā offer a normative orientation that is especially valuable in an age of rapidly advancing artificial intelligence. Rather than asking whether machines should be granted moral status, the framework redirects attention to how human beings ought to exercise moral responsibility in designing and governing increasingly powerful technologies. (Floridi and Cows 2021, 89-101, 143-150; Torrance 2005, 399-414).

### **Conclusion: Limits of Artificial Intelligence and the Contribution of Indian Philosophy**

This study has examined the question of whether Artificial General Intelligence (AGI) can possess consciousness or a soul through the philosophical frameworks of the Yoga Sūtra, its Sāṃkhya metaphysical background, and the Bhagavad Gītā. Rather than approaching the issue as a speculative prediction concerning future technological development, the analysis has treated it as an ontological and ethical inquiry into the nature of consciousness itself. By distinguishing between prakṛti as the domain of material and functional processes and puruṣa or ātman as the principle of conscious witnessing, the paper has argued that computational sophistication alone is insufficient to account for conscious subjectivity (Bryant 2009, 71-75; Dasgupta 1920, 3-4; Shastri, 2018, 33- 35, 39). Artificial systems may increasingly reproduce sophisticated forms of cognition and adaptive behaviour, yet whether such processes are sufficient for conscious awareness remains philosophically unresolved.

The Yoga Sūtra and the Bhagavad Gītā contribute a distinctive perspective to contemporary debates on artificial intelligence by maintaining a fundamental distinction between cognitive functioning and conscious existence. Within these traditions, processes such as reasoning, memory, self-regulation, and mental modification belong to the domain of prakṛti, while consciousness itself is associated with the witnessing principle of puruṣa (Dasgupta 1920, 3-8; Feuerstein 1996, 82-90; Radhakrishnan 1948, 43-45, 300-314; Bryant 2009, 58-62, 71-75). At the same time, the study has acknowledged that Indian philosophical traditions are not metaphysically uniform. Advaita Vedānta, Buddhist philosophy, and Kashmir Śaivism articulate significantly different approaches to consciousness and selfhood. Nevertheless, even where metaphysical dualism is modified or transcended, consciousness is generally not reduced to functional organisation alone. (Deutsch 1969; Radhakrishnan 1948, 22-24, 43-45). In this

respect, the Yoga Sūtra and the Bhagavad Gītā remain philosophically valuable for contemporary AI discourse because they preserve a distinction between cognition and conscious awareness that continues to challenge purely functional or emergentist accounts of mind.

The study has further argued that embodied conscious existence in the Bhagavad Gītā is situated within a broader existential framework involving prāṇa, karma, lived embodiment, and disciplined self-transformation. Through the doctrines of kṣetra and kṣetraijña, as well as the discussion of liberation in Chapter 13, the Gītā presents consciousness not as an emergent product of complexity alone, but as associated with a deeper structure of embodied and experiential existence (Radhakrishnan 1948, 287-291, 300-314, 433-434). This orientation parallels the Yoga Sūtra's understanding of liberation through citta-vṛtti-nirodha, where consciousness becomes distinguished from the fluctuations of mental activity (Bryant 2009, 58-62, 305-309). From this standpoint, the paper has suggested that artificial systems may reproduce certain functional aspects of cognition without thereby establishing the presence of conscious subjectivity in the ontological sense articulated within these traditions.

The ethical implications of this distinction are equally significant. From this perspective, responsibility for the ethical use of AI continues to rest with human beings and the institutions that develop and govern such technologies. The philosophical framework explored in this study, therefore, supports a fundamentally human-centred approach to AI governance, one that preserves the distinction between technological capability and moral agency while resisting the anthropomorphisation of artificial systems (Floridi and Cowls 2021, 89-101, 143-150; Torrance 2005, 399-414).

Ultimately, the Yoga Sūtra and the Bhagavad Gītā contribute to contemporary discussions of artificial intelligence by clarifying the distinction between intelligence, simulation, and conscious existence rather than rejecting technological progress. As artificial systems continue to evolve, the central philosophical challenge may not simply be whether machines can imitate human cognition, but whether computational complexity alone is sufficient to account for lived awareness and subjective being. In this respect, the Yoga Sūtra and the Bhagavad Gītā continue to offer important resources for interdisciplinary reflection on consciousness, embodiment, ethics, and the future of artificial intelligence. At the same time, the conclusions of this study remain specific to the Yoga Sūtra and the Bhagavad Gītā framework adopted here and are not intended to represent all Indian philosophical traditions universally.

## References

- Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.
- Bryant, E. F. (2009). *The Yoga Sūtras of Patañjali: A new edition, translation, and commentary*. North Point Press.

- Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181–204.
- Dasgupta, S. (1975). *A history of Indian philosophy* (Vol. 1). Motilal Banarsidass. (Original work published 1922)
- Dasgupta, S. N. (1920). *The study of Patañjali*. University of Calcutta.
- Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Viking Press.
- Deutsch, E. (1969). *Advaita Vedānta: A philosophical reconstruction*. University of Hawaii Press.
- Feuerstein, G. (1996). *The philosophy of classical yoga*. Inner Traditions.
- Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1).
- Goff, P. (2019). *Galileo's error: Foundations for a new science of consciousness*. Pantheon Books.
- Larson, G. J. (2001). *Classical Sāṃkhya: An interpretation of its history and meaning*. Motilal Banarsidass.
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450. <https://doi.org/10.2307/2183914>
- O'Brien-Kop, K. (2023). The hard problem of “pure” consciousness: Sāṃkhya dualist ontology. *Religious Studies*, 59, 1–17. <https://doi.org/10.1017/S0034412523000410>
- Radhakrishnan, S. (1948). *The Bhagavadgītā: With an introductory essay, Sanskrit text, English translation and notes*. George Allen & Unwin.
- Rama Prasada. (2019). *The Yoga Darśana of Patañjali with the Sāṃkhya Pravacana commentary of Vyāsa and the gloss of Vācaspati Miśra*. Dev Publishers.
- Rao, K. R. (2011). *Consciousness studies: Cross-cultural perspectives*. National Institute of Advanced Studies.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–457. <https://doi.org/10.1017/S0140525X00005756>
- Shastri, V. P. (2018). *Yoga Vijñāna Pradīpikā*. Satyam Publication House.
- Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *Biological Bulletin*, 215(3), 216–242. <https://doi.org/10.2307/25470707>
- Tononi, G. (2012). Integrated information theory of consciousness: An updated account. *Archives Italiennes de Biologie*, 150(2–3), 56–64.
- Torrance, S. (2005). Artificial agents and the expansion of the ethical circle. *AI & Society*, 19(4), 399–414.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460. <https://doi.org/10.1093/mind/LIX.236.433>