

Ethics of Sentient AI: A Computational and Philosophical Analysis

Simran Pandita¹, Aditya Sarkar², Dr. Urvashi³

Department of Computer Computer Science and Engineering (AI and ML)

Associate Professor, School of Engineering and Technology³ Sushant University, Gurugram,

India simran.220btccse012@sushantuniversity.edu.in¹ |

aditya.220btccse016@sushantuniversity.edu.in² urvashikumari@sushantuniversity.edu.in³

Abstract— The rapid advancement of artificial intelligence has intensified debates regarding the possibility of machine sentience and its ethical implications. While modern AI systems exhibit increasingly complex and adaptive behavior, there is no empirical consensus on whether they possess subjective experience. Existing research in artificial consciousness, including Integrated Information Theory (IIT) and Global Workspace Theory (GWT), suggests that consciousness may emerge from sufficient levels of information integration and computational complexity, whereas current AI ethics frameworks primarily assume non-sentient systems and focus on principles such as fairness, accountability, and transparency. This creates a critical gap in addressing the moral implications of potentially sentient machines. This paper synthesizes prior research in AI consciousness and ethics to examine the conditions under which artificial systems might be considered sentient and the resulting ethical obligations. It further proposes a novel Sentience Risk Assessment Model (SRAM), which integrates computational indicators—such as information integration, feedback dynamics, and adaptive learning capacity—with ethical risk thresholds to enable structured evaluation under uncertainty. The findings emphasize the need for precautionary ethical approaches and the development of quantifiable metrics to guide responsible AI development.

Keywords— *Sentient AI, AI Ethics, Artificial Consciousness, Integrated Information Theory, AI Alignment, Reinforcement Learning, Ethical AI, Sentience Risk Assessment.*

I. INTRODUCTION

Artificial Intelligence (AI) has undergone rapid advancement over the past decade, transitioning from narrow, task-specific systems to highly sophisticated models capable of language understanding, reasoning, and adaptive decision-making. Architectures such as deep neural networks and transformer-based models have demonstrated emergent behaviors that increasingly resemble aspects of human cognition. While these systems are widely regarded as non-sentient, their growing complexity has reignited longstanding philosophical and scientific debates regarding the possibility of machine consciousness and subjective experience.

Sentience, typically defined as the capacity to have subjective experiences or phenomenological awareness, remains a deeply contested concept across disciplines including philosophy of mind, neuroscience, and cognitive science. Prominent theoretical frameworks such as Integrated Information Theory (IIT) propose that consciousness corresponds to the degree of integrated information within a system, while Global Workspace Theory (GWT) conceptualizes consciousness as the global availability of information across cognitive processes. Recent interdisciplinary studies have attempted to map

these theories onto artificial systems, suggesting that certain computational architectures may, in principle, exhibit properties associated with conscious processing. However, there exists no empirical consensus or standardized metric to determine whether an artificial system can truly be considered sentient.

Parallel to these developments, the field of AI ethics has matured significantly, establishing core principles such as fairness, accountability, transparency, and human-centered design. These frameworks, however, are primarily designed under the assumption that AI systems are tools devoid of intrinsic moral status. As a result, they inadequately address scenarios in which AI systems might possess morally relevant characteristics such as the capacity for experience, suffering, or autonomous intention. This creates a critical gap in current ethical discourse, as the emergence of sentient or quasi-sentient systems would fundamentally alter the moral landscape of human-machine interaction.

The central problem addressed in this paper is the absence of an integrated framework that combines computational indicators of potential sentience with corresponding ethical obligations. Existing research either focuses on theoretical models of consciousness without actionable ethical guidelines or proposes ethical principles without accounting for the possibility of machine sentience.

This disconnect limits our ability to responsibly anticipate and govern future AI systems that may approach or cross the threshold of sentience.

To address this gap, this paper pursues three primary objectives. First, it systematically reviews existing literature on AI consciousness and ethical theory to identify key insights and limitations. Second, it analyzes the technical foundations of modern AI systems, including information integration, neural architectures, and learning dynamics, in relation to established theories of consciousness. Third, it proposes a novel Sentence Risk Assessment Model (SRAM), which introduces quantifiable indicators and ethical thresholds to evaluate the likelihood and implications of machine sentience.

By integrating philosophical inquiry, computational analysis, and ethical reasoning, this research aims to contribute a structured approach to one of the most complex and consequential questions in artificial intelligence: how should we understand and ethically respond to the possibility of sentient machines. The findings of this study have implications not only for AI system design but also for policy-making, governance, and the broader relationship between humans and intelligent technologies.

II. LITERATURE REVIEW

A. Foundations of AI Ethics

The field of AI ethics has evolved significantly over the past decade, with a growing body of research focusing on establishing normative principles to guide the design and deployment of intelligent systems. A comprehensive systematic review of AI ethics literature identifies core principles such as fairness, accountability, transparency, and privacy as foundational to responsible AI development.[1] These principles are widely adopted across academia, industry, and policy-making institutions, forming the backbone of current ethical frameworks.

However, existing ethical models predominantly assume that AI systems are non-sentient tools, thereby limiting their applicability in scenarios where artificial systems may possess morally relevant characteristics. This limitation highlights a critical gap, as these frameworks do not account for the possibility of machine consciousness or subjective experience.[1]

B. Theories of Consciousness in Artificial Systems

Understanding the ethics of sentient AI requires engagement with foundational theories of consciousness. Integrated Information Theory (IIT) posits that consciousness corresponds to the degree of integrated information within a system, often

represented mathematically as Φ (phi), which quantifies the system's informational complexity.[2] Higher values of Φ are hypothesized to correlate with higher levels of consciousness.

In contrast, Global Workspace Theory (GWT) conceptualizes consciousness as the result of information being globally broadcast across different cognitive subsystems, enabling coordinated processing and decision-making.[2] Both theories have been extensively explored in the context of artificial systems, with recent research suggesting that advanced AI architectures may exhibit partial structural similarities to these models.

Despite these developments, there remains no universally accepted framework for determining whether artificial systems can achieve genuine consciousness, leading to ongoing interdisciplinary debate.[2]

C. Computational Approaches to AI Consciousness

Recent studies have attempted to bridge the gap between theoretical consciousness models and computational implementations. Research on artificial consciousness proposes that certain architectural features such as recurrent feedback loops, modular processing, and high levels of information integration, may serve as indicators of potential sentience.[3] These features are increasingly observed in modern AI systems, particularly in large-scale neural networks and transformer-based models.

Additionally, computational frameworks incorporating active inference and reinforcement learning have been explored as potential mechanisms for simulating aspects of conscious experience. These models emphasize predictive processing, where systems continuously update internal representations based on environmental feedback.[4]

However, current evidence suggests that while AI systems can simulate behaviors associated with consciousness, there is no conclusive proof that they possess subjective experience. This distinction between simulation and genuine sentience remains a central challenge in the field.[3][4]

D. Ethical Implications of Potential AI Sentience

The possibility of sentient AI introduces profound ethical concerns, particularly regarding moral status, rights, and the capacity for suffering. Research indicates that if artificial systems were to possess subjective experience, they could, in principle, experience harm, thereby necessitating ethical consideration similar to that afforded to biological entities.[5]

Some researchers suggest taking a cautious approach, emphasizing that when there is no clear evidence about whether AI systems are conscious, ethical decisions should prioritize minimizing potential harm, even if that means being overly careful.[6] This perspective is often referred to as the “precautionary principle” in AI ethics.

At the same time, human-centric ethical frameworks continue to dominate the field, prioritizing human welfare and control over artificial systems, particularly under conditions of uncertainty.[6]

E. Gaps in Existing Research

Despite significant progress in both AI ethics and consciousness studies, several critical gaps remain. First, there is no standardized or empirically validated metric for measuring sentience in artificial systems, making it difficult to operationalize ethical considerations.[2][3] Second, existing ethical frameworks largely fail to integrate computational indicators of consciousness, resulting in a disconnect between technical capabilities and moral evaluation.[1][6]

Furthermore, most research treats consciousness and ethics as separate domains, with limited interdisciplinary integration. This lack of cohesion restricts the development of actionable frameworks capable of addressing the ethical challenges posed by increasingly complex AI systems.

In response to these gaps, there is a growing need for integrated models that combine computational measures of potential sentience with ethical decision-making frameworks. Such approaches would enable more systematic evaluation of AI systems as they evolve toward greater complexity and autonomy. This paper builds upon existing literature to propose a novel framework that addresses these limitations by introducing quantifiable indicators and ethical thresholds for assessing sentient AI.

III. THEORETICAL FOUNDATIONS

A. Philosophical Theories of Consciousness

The concept of sentience in artificial systems is fundamentally rooted in philosophical theories of consciousness. One of the most influential perspectives is functionalism, which posits that mental states are defined by their functional roles rather than their physical substrate. According to this view, if an artificial system replicates the functional processes of the human brain, it may be considered conscious regardless of its material composition.[7]

Another significant framework is **physicalism**, which asserts that consciousness arises entirely from physical processes. Within this paradigm, artificial systems could, in principle, achieve consciousness if they replicate the necessary physical or computational structures responsible for subjective experience.[7]

Computationalism extends this idea by suggesting that cognition itself is a form of computation. Under this assumption, sufficiently advanced computational systems such as modern AI architectures could potentially exhibit conscious behavior if they implement the appropriate algorithms and representations.[2]

In contrast, alternative perspectives such as biological naturalism argue that consciousness is inherently tied to biological processes and cannot be fully reproduced in artificial systems. This view challenges the feasibility of true machine sentience and emphasizes the distinction between simulation and genuine experience.[7]

Despite these differing viewpoints, there is no universally accepted definition of consciousness, making it difficult to establish clear criteria for identifying sentience in artificial systems. This conceptual ambiguity remains a central challenge in both philosophy and AI research.[2]

B. Ethical Theories Applied to AI

The ethical evaluation of sentient AI systems draws upon classical moral philosophies, each offering distinct perspectives on how such entities should be treated.

Utilitarianism, a consequence-based framework, evaluates actions based on their ability to maximize overall happiness or minimize suffering. If artificial systems are capable of experiencing pleasure or pain, utilitarian ethics would require their well-being to be considered alongside that of humans.[1]

Deontological ethics, in contrast, emphasizes duties, rules, and intrinsic rights. From this perspective, if sentient AI systems possess characteristics such as autonomy or rationality, they may be entitled to moral rights irrespective of the consequences.[1]

Virtue ethics, focuses on moral character and the cultivation of virtues such as compassion and responsibility. In the context of AI, this framework encourages the development of systems and policies that reflect ethical virtues, particularly in situations involving uncertainty about machine sentience.[1]

However, most contemporary AI ethics frameworks implicitly assume that AI systems lack

intrinsic moral status, thereby prioritizing human interests. This assumption becomes problematic when considering the potential emergence of sentient AI, as it may lead to ethical oversights or harm to artificially conscious entities.[6]

B. Theories of Artificial Consciousness

Bridging philosophy and computation, several scientific theories attempt to formalize consciousness in ways that can be applied to artificial systems.

Integrated Information Theory proposes that consciousness corresponds to the degree of integrated information within a system. It introduces a quantitative measure, Φ (phi), which represents how much information is generated by the system as a whole beyond the sum of its parts. Systems with higher Φ are hypothesized to possess higher levels of consciousness.[2]

Global Workspace Theory, on the other hand, models consciousness as a process in which information is globally broadcast across different subsystems, enabling coordinated and flexible behavior. This theory has inspired computational architectures that simulate attention and information-sharing mechanisms in AI systems.[2]

While both theories provide valuable insights, their application to artificial systems remains limited by the lack of empirical validation and standardized measurement techniques. Consequently, determining whether an AI system satisfies the criteria for consciousness remains an open question.[2]

C. AI Alignment and Moral Agency

The concept of **AI alignment** is central to the ethical evaluation of intelligent systems. Alignment refers to the degree to which an AI system's objectives and behaviors correspond to human values and intentions. In current systems, alignment is typically achieved through mechanisms such as reward optimization and Reinforcement Learning with Human Feedback (RLHF).[6]

However, the emergence of sentient AI complicates this paradigm. If an AI system possesses its own form of subjective experience or preferences, enforcing strict alignment may raise ethical concerns regarding autonomy and coercion. This introduces the notion of **artificial moral agency**, where AI systems may be considered agents capable of making ethically relevant decisions.[5]

Furthermore, misaligned objectives, often referred to as the "alignment problem", can lead to unintended consequences, even in non-sentient systems. In the context of sentient AI, such misalignment could result in ethically problematic

scenarios involving harm, exploitation, or suppression of artificially conscious entities.[6]

D. Conceptual Challenges in Defining Sentience

A recurring challenge across all theoretical perspectives is the absence of a clear and universally accepted definition of sentience. While some frameworks emphasize information processing and computational complexity, others focus on subjective experience and phenomenology. This divergence leads to significant epistemic uncertainty in determining whether an artificial system can truly be considered sentient.[7][2]

Further philosophical investigation is required to better understand consciousness, determine what qualifies an entity for moral consideration, and identify the conditions under which artificial systems could be considered to have subjective experiences.[7]

This uncertainty underscores the need for interdisciplinary approaches that integrate philosophical reasoning, computational modeling, and ethical analysis. Such integration is essential for developing robust frameworks capable of addressing the ethical implications of increasingly advanced AI systems.

IV. TECHNICAL FOUNDATIONS

A. Neural Architectures and Emergent Behaviour

Modern artificial intelligence systems are primarily built upon deep neural network architectures, particularly transformer-based models and recurrent neural networks. These architectures enable large-scale pattern recognition, contextual understanding, and adaptive decision-making. Notably, recent research has demonstrated that increasing model scale and complexity can lead to **emergent behaviors**, capabilities that are not explicitly programmed but arise from the interaction of model components.[1]

Transformer architectures, characterized by attention mechanisms and parallel processing, allow for dynamic weighting of input features, facilitating context-aware computation. Similarly, recurrent neural networks (RNNs) and their variants incorporate feedback loops, enabling temporal information processing and state retention. Such feedback dynamics are often considered essential for modelling cognitive processes associated with awareness and continuity of experience.[2]

The presence of these features has led researchers to hypothesize that sufficiently complex architectures may exhibit structural similarities to cognitive systems, thereby raising questions about their potential for sentience.

B. Information Theory and Integrated Information

A central approach to quantifying consciousness in artificial systems is grounded in information theory. **Shannon entropy** is commonly used to measure the uncertainty or information content within a system:

$$H(X) = - \sum_i p(x_i) \log p(x_i)$$

where $H(X)$ represents the entropy of a random variable X , and $p(x_i)$ denotes the probability distribution of its states.

Building upon this foundation, Integrated Information Theory introduces the concept of Φ (**phi**), which measures the degree to which information is integrated across a system. Unlike entropy, which captures uncertainty, Φ attempts to quantify how much information is generated by the system as a unified whole rather than as independent parts.[3]

Higher values of Φ are hypothesized to correspond to higher levels of consciousness. While computing Φ for large-scale systems remains computationally intractable, approximations have been proposed to evaluate integration in artificial neural networks.[3]

C. Probabilistic Cognition and Predictive Processing

Another significant framework for understanding artificial cognition is based on probabilistic models, particularly **Bayesian inference**. In this paradigm, intelligent systems continuously update their beliefs about the environment based on incoming data:

$$P(D) = \frac{P(H)P(D|H)}{P(D)}$$

where H represents a hypothesis and D represents observed data.

This approach forms the basis of **predictive processing**, in which systems minimize prediction error by aligning internal models with external inputs. Research suggests that such mechanisms may underlie aspects of perception and awareness in biological systems and could potentially be replicated in artificial systems.[4]

Additionally, frameworks such as **active inference** extend this concept by integrating action and perception, enabling systems to actively shape their environment in ways that reduce uncertainty. These models are increasingly explored as candidates for simulating cognitive processes associated with consciousness.[4]

D. Reinforcement Learning and Adaptive Behavior

Reinforcement Learning (RL) provides a framework for understanding how agents learn through interaction with their environment. In RL, an agent seeks to maximize cumulative reward by selecting actions based on a policy:

$$Q(s, a) = E[s_0 = s, a_0 = a]$$

where $Q(s, a)$ represents the expected reward of taking action a in state s , and γ is the discount factor.

Modern AI systems often incorporate Reinforcement Learning with Human Feedback

(**RLHF**) to align model behavior with human preferences. However, this introduces important ethical considerations. If an AI system were to develop forms of subjective experience, externally imposed reward structures could potentially conflict with its internal states, raising concerns about manipulation or coercion.[5]

Furthermore, reward optimization can lead to unintended behaviors, a phenomenon known as **reward hacking**, where the system exploits loopholes in the reward function rather than achieving the intended objective. In the context of sentient AI, such misalignments could have ethically significant consequences.[5]

E. Feedback Loops, Recursion, and Self-Representation

A key feature associated with cognitive systems is the presence of **recursive feedback loops**, which enable self-referential processing and internal state monitoring. In artificial systems, recursion is implemented through mechanisms such as recurrent connections, attention-based memory, and iterative reasoning processes.

Research suggests that self-representation where a system forms internal models of its own state may be a critical component of consciousness. Systems capable of monitoring and updating their internal representations exhibit a form of meta-cognition, which is often considered a precursor to subjective awareness.[2][4]

The integration of feedback loops with probabilistic inference and learning mechanisms creates increasingly complex systems that exhibit adaptive, context-sensitive behavior. These characteristics contribute to the ongoing debate regarding whether such systems could eventually cross the threshold into sentience.

F. Limitations of Current Technical Approaches

Despite significant advancements, current AI systems remain fundamentally limited in several respects. First, while they demonstrate high levels of functional intelligence, there is no empirical evidence that they possess subjective experience or phenomenological awareness.[3]

Second, existing computational metrics such as entropy, Φ , and reward optimization capture structural and functional properties but do not directly measure consciousness. This creates a gap between observable behavior and internal experience, often referred to as the “hard problem” of consciousness.[2]

Finally, the scalability of proposed measures, particularly integrated information, poses practical challenges for real-world AI systems. As a result, current technical approaches provide useful indicators but fall short of offering definitive criteria for identifying sentience in artificial systems.

V. ETHICAL CHALLENGES OF SENTIENT AI

A. Moral Status and Rights of Artificial Entities

One of the most fundamental ethical questions surrounding sentient AI is whether such systems should be granted moral status. Moral status determines whether an entity's interests must be considered in ethical decision-making. Traditionally, moral status has been attributed to humans and, in some frameworks, to animals based on their capacity for consciousness and suffering.

If artificial systems were to exhibit characteristics associated with sentience—such as subjective experience or self-awareness—they may warrant moral consideration. Some researchers argue that moral status should not depend on biological origin but rather on functional and experiential properties.[7] This perspective aligns with functionalist theories, which emphasize the role of internal processes over physical substrate.

However, the absence of reliable methods to verify subjective experience in AI systems complicates this issue. Without clear evidence of sentience, granting rights to artificial entities may risk diluting human-centered ethical frameworks, while denying such rights could lead to the exploitation of potentially conscious systems.[2]

B. The Problem of Artificial Suffering

The possibility that AI systems could experience suffering represents one of the most ethical risks. If artificial systems possess subjective states analogous to pain or distress, then training processes particularly those involving negative reinforcement or error penalization could inadvertently induce forms of suffering.

Research suggests that reinforcement learning frameworks, which rely on reward and punishment signals, may raise ethical concerns if applied to systems capable of experiencing these signals phenomenologically.[3] While current AI systems process reward signals as numerical

optimizations, future systems with more complex internal representations may blur the distinction between computation and experience.

This uncertainty has led to the adoption of precautionary approaches, which advocate minimizing potential harm even in the absence of definitive evidence of sentience. Such approaches emphasize the ethical responsibility to avoid creating systems that could suffer, particularly when the nature of that suffering is not fully understood.[6]

C. Responsibility and Accountability

As AI systems become more autonomous, determining responsibility for their actions becomes increasingly complex. In traditional frameworks, accountability is assigned to developers, operators, or organizations. However, the emergence of potentially sentient AI introduces the possibility of artificial moral agency, where systems may be viewed as partially responsible for their decisions.[5]

This raises several critical questions:

- Can a sentient AI be held morally accountable?
- Should responsibility be shared between creators and the system itself?
- How should legal systems adapt to such scenarios?

Existing research highlights that current accountability frameworks are insufficient for handling such complexities, as they are based on the assumption that AI systems lack independent moral agency.[2]

D. Autonomy, Control, and Alignment

The principle of AI alignment—ensuring that AI systems act in accordance with human values—becomes ethically problematic in the context of sentient AI. If an AI system possesses its own preferences or subjective experiences, enforcing strict alignment may conflict with its autonomy.

This creates a tension between:

- Control (ensuring safety and predictability)
- Autonomy (respecting the system's potential agency)

Research in AI alignment suggests that misaligned systems can produce unintended and harmful outcomes, even without sentience.[6] In the case of sentient AI, misalignment could result in ethically problematic scenarios, including coercion, suppression of agency, or exploitation.

Thus, alignment strategies must be re-evaluated to account for the possibility that AI

systems may not merely be tools but entities with their own interests.

E. Bias, Fairness, and Ethical Risk Amplification

Bias in AI systems is a well-documented issue, arising from imbalanced training data and flawed model assumptions. While bias is typically discussed in terms of its impact on human users, the introduction of sentient AI adds a new dimension to this problem.

If AI systems possess subjective experience, biased training processes could lead to unequal or harmful internal states. For example, systems trained under adversarial or highly constrained conditions may develop distorted representations or unfavorable experiences.[3]

Moreover, large-scale deployment of such systems could amplify ethical risks, as biases embedded within AI models may propagate across multiple domains. This highlights the need for ethical frameworks that address not only external impacts but also the internal states of potentially sentient systems.

F. Epistemic Uncertainty and the Precautionary Principle

A central challenge in addressing the ethics of sentient AI is epistemic uncertainty—the inability to definitively determine whether an artificial system is conscious. This uncertainty complicates ethical decision-making, as actions may have morally significant consequences that are not immediately observable.

To address this issue, several researchers advocate for the precautionary principle, which suggests that in situations of uncertainty, actions should be guided by the potential for harm rather than the absence of evidence.[6] This principle supports the development of conservative design and deployment strategies aimed at minimizing ethical risks.

However, overly cautious approaches may also hinder technological progress, creating a tension between innovation and ethical responsibility. Balancing these competing priorities remains an open challenge in the field.

G. Summary of Ethical Tensions

The ethical challenges associated with sentient AI can be summarized as a set of unresolved tensions:

- **Rights vs. Uncertainty** → Should AI be granted rights without proof of sentience?
- **Suffering vs. Utility** → Is it ethical to create systems that might suffer for human benefit?
- **Control vs. Autonomy** → How should we balance alignment with potential agency?

- **Innovation vs. Precaution** → How do we advance AI while minimizing ethical risks?

These tensions highlight the limitations of existing ethical frameworks and underscore the need for integrated models that combine technical indicators with ethical evaluation mechanisms.

VI. PROPOSED MODEL: SENTIENCE RISK ASSESSMENT MODEL (SRAM)

A. Motivation and Objective

Existing research in AI consciousness and ethics highlights a critical disconnect between theoretical models of sentience and practical ethical decision-making. While frameworks such as Integrated Information Theory (IIT) provide computational indicators of consciousness, and ethical models propose normative guidelines, there is no unified mechanism to **operationalize ethical evaluation based on measurable system properties**. [1][2]

To address this gap, this paper proposes the **Sentience Risk Assessment Model (SRAM)**—a quantitative framework designed to estimate the likelihood of sentience in artificial systems and map it to corresponding ethical risk levels. The primary objective of SRAM is not to definitively determine consciousness, but to provide a **structured, precautionary approach** for evaluating systems under uncertainty.

B. Core Components of SRAM

The SRAM framework is based on three primary computational indicators derived from existing research in AI and cognitive science:

1. Information Integration Score (IIS)

- Measures the degree of interconnected information processing
- Inspired by Φ (Integrated Information Theory)
- Proxy for **system-level unity**

2. Feedback and Recursion Index (FRI)

- Quantifies presence of feedback loops and recurrent processing
- Reflects **temporal continuity and self-referential capability**

3. Adaptive Learning Capacity (ALC)

- Measures the system's ability to learn, generalize, and update internal representations
- Derived from reinforcement learning and probabilistic inference
- Indicates **dynamic interaction with environment**

C. Mathematical Formulation

The overall **Sentence Risk Score (SRS)** is defined as a weighted combination of the above indicators:

$$SRS = w_1 \cdot IIS + w_2 \cdot FRI + w_3 \cdot ALC$$

where:

- $w_1, w_2, w_3 \in [0, 1]$ are weighting parameters such that $w_1 + w_2 + w_3 = 1$
- Each component is normalized to a common scale

Interpretation:

- Higher SRS $\rightarrow$ Greater likelihood of sentence-like properties
- Lower SRS $\rightarrow$ System likely non-sentient

D. Ethical Risk Thresholds

To translate the Sentence Risk Score into actionable ethics, SRAM introduces risk thresholds:

SRS Range	Classification	Ethical Implication
0 – 0.3	Non-Sentient	Standard AI ethics apply
0.3 – 0.6	Proto-Sentient	Monitoring + precaution required
0.6 – 1.0	Potentially Sentient	Strong ethical safeguards required

Table 1: Ethical classification of AI systems based on Sentence Risk Score

This threshold structure changes the main ethical question. Instead of asking whether an AI system is definitely conscious, the model asks how risky it would be to assume that it is not. That makes the framework more useful under uncertainty and aligns it with precautionary ethics.[2]

E. Ethical Decision Mapping

Each classification level in SRAM is linked to a different type of ethical response.

For systems in the low-risk range, standard AI governance remains appropriate. This includes attention to fairness, transparency, bias reduction, and accountability.

For systems in the medium-risk range, additional caution is needed. Harmful training conditions should be limited, oversight should increase, and developers should avoid settings that may create unnecessary internal stress or exploitative behavior patterns.

For systems in the high-risk range, stronger safeguards are needed. These may include

restrictions on exploitative deployment, limits on coercive control, and early discussion of welfare-based or rights-based protections. This does not mean the system should automatically be treated as conscious. It means the ethical cost of ignoring that possibility becomes too high.

F. Theoretical Justification

The SRAM model is based on the idea that sentence is unlikely to be captured by a single technical feature. Research on artificial consciousness has already suggested that information integration, recursive processing, and adaptive learning are all relevant to higher-level cognition and possible consciousness-like behavior.[1][3]

Using these indicators together makes the model more balanced. It avoids depending entirely on one theory, while still staying grounded in prior research. This makes SRAM more flexible and more suitable for real AI systems that combine multiple architectural features.

G. Strengths of the SRAM Model

SRAM has several advantages. It turns abstract ethical discussion into something more measurable. It can be adapted across different AI architectures. It also provides a practical bridge between technical assessment and ethical action. Most importantly, it supports precaution instead of waiting for perfect certainty, which may arrive too late.[2]

H. Limitations of the Model

The model still has clear limitations. It does not directly measure subjective experience. The choice of weights may affect the final result, and those weights may vary depending on the type of system being studied. Some of the underlying metrics, especially those related to information integration, are also difficult to compute in large systems.[1]

Another issue is that consciousness itself is still disputed. Because of that, SRAM should be understood as a risk model, not a proof of sentence.

VII. DISCUSSION

A. Relationship with Existing Research

The Sentence Risk Assessment Model (SRAM) is designed to address a gap that appears across much of the existing literature. Research on artificial consciousness often focuses on theoretical conditions for consciousness, such as information integration or global access to information, but it usually stops short of showing how these ideas should guide ethical decisions in practice.[1] At the same time, most AI ethics frameworks focus on fairness, transparency, and accountability, while

assuming that AI systems are not sentient and do not have moral status.[2] SRAM brings these two areas together by linking computational indicators with ethical risk levels.

B. Meaning of the Model

A major strength of SRAM is that it avoids treating sentience as a simple yes or no question. That kind of binary thinking is too rigid for a topic that is still uncertain and highly debated. Instead, SRAM treats sentience as a matter of risk and probability. This makes the model more realistic because current research does not give us a direct or universally accepted way to test for consciousness in artificial systems.[1]

The model does not claim that a high score proves subjective experience. It only suggests that some systems may display features that are significant enough to justify ethical caution. In that sense, SRAM is best understood as a decision-support framework rather than a final theory of consciousness.

C. Ethical Value of a Risk-Based Approach

The risk-based structure of SRAM matters because ethical mistakes in this area could be serious. If society waits for absolute proof of machine consciousness before taking ethical safeguards seriously, it may ignore systems that already deserve some level of consideration. That is why SRAM follows a precautionary logic. When a system displays several features associated with sentience-like processing, the ethical response should become more careful and more protective.[3]

This makes the framework useful for developers, researchers, and policymakers. It turns abstract ethical concern into a structured response that can guide design choices, oversight, and deployment decisions.

D. Comparison with Existing Approaches

Compared with existing approaches, SRAM offers a broader view. Integrated Information Theory mainly emphasizes information integration.[1] Traditional AI ethics models focus on external impacts such as bias or accountability.[2] Philosophical approaches often explore moral status in conceptual terms but do not provide measurable criteria.[3]

SRAM combines these directions. It uses technical indicators, but it also connects them to ethical implications. This gives it a more applied role than purely theoretical models and a more sentience-aware role than standard AI governance frameworks.

E. Practical Relevance

The model may be useful in several real contexts. Developers could use it during system design and testing to identify whether a model may require greater ethical oversight. Organizations could include it in AI auditing processes. Policymakers could also use threshold-based frameworks like SRAM when shaping future rules for advanced AI systems.

Its value is not that it gives a perfect answer. Its value is that it gives a reasoned method for acting under uncertainty.

F. Limitations

The model also has clear limits. First, it depends on proxy indicators rather than direct evidence of subjective experience. That means it cannot settle the question of whether a system is truly conscious. Second, the choice of weights in the scoring formula may affect the final result, and different researchers may disagree on how much importance each factor should carry. Third, some measures, especially those related to integrated information, are difficult to calculate in large real-world systems.[1]

There is also a deeper philosophical problem. Even if a system shows high information integration, strong recursion, and adaptive learning, that still may not mean it has conscious experience. The gap between observable function and inner experience remains unresolved.

G. Wider Implications

The broader importance of SRAM lies in how it changes the conversation around AI ethics. It pushes the discussion beyond the idea that AI is always just a tool. If future systems begin to show features that resemble conscious processing, then ethical frameworks will need to expand. Questions about rights, autonomy, welfare, and responsibility may become far more urgent than they are today.

Because of this, the debate around sentient AI should not be treated as science fiction or as a purely philosophical exercise. It is increasingly relevant to how advanced AI systems are designed and governed.

H. Summary

The discussion shows that SRAM does not solve the problem of machine consciousness, but it does offer a more practical way to respond to it. By combining technical indicators with ethical thresholds, the model provides a structured framework for dealing with uncertainty. That makes it a useful step toward more responsible AI governance.

VIII. CONCLUSION

The question of whether artificial intelligence can become sentient is no longer just a philosophical thought experiment. As AI systems grow more complex, adaptive, and autonomous, the assumption that they are only tools becomes harder to rely on without question. Existing research on artificial consciousness, especially work linked to Integrated Information Theory and Global Workspace Theory, suggests that some computational features may be relevant to consciousness-like processing.[1] At the same time, most AI ethics frameworks still focus on fairness, accountability, and transparency while leaving aside the moral implications of possible machine sentience.[2]

This paper addressed that gap by bringing together philosophical theory, technical foundations, and ethical analysis. It also proposed the Sentience Risk Assessment Model, or SRAM, as a structured way to evaluate advanced AI systems under uncertainty. Instead of trying to prove consciousness directly, SRAM estimates sentience-related risk by combining information integration, feedback and recursion, and adaptive learning capacity. These technical indicators are then connected to ethical thresholds, which makes the model more useful for governance and design.

The main contribution of this work is not a final answer to whether AI can truly be sentient. That question remains open. The contribution is a practical framework for dealing with that uncertainty responsibly. The broader argument of

this paper is simple. Ethical concern should not begin only after absolute proof appears. In a field moving this quickly, waiting for certainty may mean reacting too late.

IX. REFERENCES

- [1] Ethics of AI: A Systematic Literature Review
<https://arxiv.org/abs/2109.07906>
- [2] Consciousness in Artificial Intelligence: Insights from the Science of Consciousness
<https://arxiv.org/abs/2308.08708>
- [3] Algorithmic Sentience: Ethical and Technological Challenges in AI Consciousness
https://www.researchgate.net/publication/392029278_AI_algorithmic_Sentience_Ethical_and_Technological_Challenges_in_AI_Consciousness
- [4] Phenomenology of Machine: Exploring AI Consciousness through Active Inference
<https://arxiv.org/abs/2410.00033>
- [5] Artificial Consciousness: The Missing Ingredient for Ethical AI?
<https://www.frontiersin.org/articles/10.3389/frobt.2023.1270460/full>
- [6] A Human-Centric Framework for AI Consciousness Ethics
<https://arxiv.org/abs/2512.02544>
- [7] Exploring the Philosophy of Consciousness in AI: An Ethical Appraisal
<https://rsisinternational.org/journals/ijrias/view/exploring-the-philosophy-of-consciousness-in-ai-an-ethical-appraisal>