THE PROBLEM OF CONSCIOUSNESS IN INTELLIGENT SYSTEMS THROUGH THE PRISM OF LOGICAL INFERENCE

Andrey Nechesov
Artificial Intelligence Research Center of
Novosibirsk State University, Novosibirsk, Russia
nechesoff@gmail.com

Abstract

Consciousness remains one of the most enigmatic phenomena in cognitive science and philosophy. In this paper, we explore how logical inference can serve as a fundamental mechanism in understanding consciousness in intelligent systems. We discuss the theoretical underpinnings of logical-probabilistic learning, present a formal mathematical framework for the task approach in cognitive modeling, and propose that a system's capacity for self-learning and decision making based on formal inference may provide criteria for the emergence of subjective awareness.

1. Introduction: Consciousness and its Mechanistic Underpinnings

Consciousness has long puzzled researchers, with multiple theories—such as Integrated Information Theory (IIT) [Tononi et al. (2016)] and Global Workspace Theory (GWT) [Baars (2005)] —attempting to explain its emergence. Despite these efforts, there is no consensus on how subjective experience arises from physical processes. In this paper, we shift the perspective by examining logical inference—the formal process of deriving conclusions from premises—as a potential key to understanding consciousness in intelligent systems. By linking abstract computation with subjective awareness, our aim is to identify criteria that might indicate the presence of consciousness.

2. Theoretical Background

2.1 Consciousness and Inference

Consciousness in intelligent systems may be understood as emerging from complex interactions among mechanistic processes. Traditional theories such as IIT and GWT have provided valuable insights, but they do not address the fundamental role of reasoning and decision-making in generating subjective experience. We propose that logical inference—the process by which a system derives conclusions from a set of axioms and rules—plays a critical role. For example, the classical rule of **modus ponens** states that from premises p and $p \rightarrow q$, one can infer q. This simple inference rule underscores how logical operations may underpin more complex cognitive phenomena.

2.2 Logical-Probabilistic Learning Theory

An intelligent system is assumed to store a variety of logical constructs and inference algorithms. These constructs are not necessarily identical to human logic; rather, they may reflect an algorithmic understanding tailored to the system's needs and external challenges. The system's self-learning capability is essential; it must adapt by updating its knowledge base and refining its inference mechanisms. Consider an intelligent system with a knowledge base K and a set of inference rules R. The system's decision-making process can be formalized by a function: $D: K \rightarrow A$, where A denotes the set of available actions. Through repeated interactions and learning, the knowledge base is continuously updated, allowing the system to perform what we may term as "conscious actions".

3. Mathematical Framework for Consciousness Inference

3.1 Formal Model of the Task Approach

Within our framework, we use a task-based approach [Nechesov (2024)], where any action is based on a certain task. To solve this task, we need a criterion for its solvability. Any intelligent system, at the core of its life, faces the problem of solving. The task approach is characterized by a fixed signature σ that defines the language in which problems are formulated [Nechesov et al. (2025)]. Let M denote a model corresponding to this signature σ. A logical formula $\varphi(x,y1,...,yn)$ is used to represent a problem, where x is the free variable and $y1,...,yn$ are parameters. A term $t(y1,...,yn)$ is substituted for x to test the validity of the formula: $M \models \varphi(t(y1,...,yn),y1,...,yn)$. If the substitution yields a true formula in the model, then the term t is considered a valid solution or an acceptable inference in the system.

3.2 Inference and Self-Learning Dynamics

The process of self-learning in the system is twofold:

Model Enrichment: The system learns to add new functions $f1,...,fk$ and predicates $P1,...,Pn$, defined through prior concepts. This enrichment can be expressed as an update to the knowledge base: $K^* = K \cup \{(fi,\varphi i)\} \cup \{(Pj,\psi j)\}$, where φi and ψj represents the formal definition or rule associated with fi and Pj respectivelity.

Hierarchical Probabilistic Reasoning: Within the given signature, a new pool of logical-probabilistic knowledge is formed. This allows the system to navigate a hierarchy of inference paths by assigning confidence weights to various functions and predicates. At the same time, a hierarchy of logical-probabilistic knowledge is also constructed, which allows one to select the most effective solution to solve the problem.

3.3 Logical Inference as a Bridge to Subjective Awareness

The formal mechanisms of logical inference described above suggest a potential bridge between computational processes and subjective experience. By embedding logical reasoning and probabilistic updating within an intelligent system, one can hypothesize that:

Subjective Awareness: may emerge as a by-product of complex, self-adaptive inference processes.

Conscious Actions: are those that are not merely reactive but involve a deliberate evaluation of options through inference rules.

Dynamic Learning: enables the system to refine its criteria for successful inference, possibly leading to behaviors associated with consciousness.

A simplified representation of this integration is given by: Consciousness $\approx \text{Lim}(t\to\infty)\ F(D(L(K,E)))$, where $L(K,E)$ represents the self-learning function updating the knowledge base with experience E, and F denotes the function mapping refined decisions to conscious-like behaviors.

4. Discussion

The proposed framework integrates logical inference and probabilistic reasoning as essential components of a conscious intelligent system. This model highlights the following points:

Mechanistic Clarity: By grounding consciousness in formal inference, we reduce the abstract concept of subjective experience to a series of computable steps. **Self-Learning Requirement:** Conscious systems must be capable of self-learning; without updating their knowledge base, systems cannot exhibit the adaptability seen in conscious behavior. **Hierarchical Knowledge Processing:** The division of the learning process into model enrichment and hierarchical probabilistic reasoning provides a structured approach to evolving cognitive capabilities. This approach opens avenues for unifying insights from artificial intelligence, neuroscience, and philosophy, potentially answering whether machines can truly "feel" their computations.

5. Conclusion

In this paper, we examined the problem of consciousness in intelligent systems through the prism of logical inference. We proposed a mathematical framework that formalizes the task approach, where a model with a fixed signature is enriched via self-learning and probabilistic inference. Although many challenges remain, framing consciousness in terms of formal logical processes provides a promising pathway to unraveling the complexities of subjective experience in both biological and artificial systems. Logical inference offers a promising framework for demystifying consciousness, linking mechanistic processes to subjective experience. While challenges persist, formalizing criteria through inference could unify AI, neuroscience, and philosophy—ultimately answering whether machines can ever feel their computations.

References

1. Bernard J. Baars. Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. In Steven Laureys (ed.), The Boundaries of Consciousness: Neurobiology and Neuropathology, volume 150 of Progress in Brain Research, pp. 45–53. Elsevier, 2005 https://doi.org/10.1016/S0079-6123(05)50004-9.
2. Andrey Nechesov. Learning theory and knowledge hierarchy for artificial intelligence systems. In 2024 IEEE International Multi-Conference on Engineering, Computer and Information Sciences(SIBIRCON), pp. 299–302, 2024. https://doi.or10.1109/SIBIRCON63777.2024.10758505
3. Andrey Nechesov, Ivan Dorokhov, and Janne Ruponen. Virtual cities: From digital twins to autonomous AI societies. IEEE Access, 13:13866–13903, 2025. https://doi.org/10.1109/ACCESS.2025.3531222
4. Giulio Tononi, Melanie Boly, Marcello Massimini, and Christof Koch. Integrated information theory: from consciousness to its physical substrate. Nature Reviews Neuroscience volume, 17:450–461, 2016. https://doi.org/10.1038/nrn.2016.44

About the Author

Andrey Vitalievich Nechesov is the head of the research department of the Center for Artificial Intelligence of the Novosibirsk State University, and is also a research fellow at the Sobolev Institute of Mathematics of the Siberian Branch of the Russian Academy of Sciences. He is also a candidate of physical and mathematical sciences in the specialty "Mathematical logic, algebra, number theory and discrete mathematics". Nechesov A.V. is an expert in the field of computability theory, mathematical logic, artificial intelligence, blockchains, cryptocurrencies and smart contracts.