



The Phenomenology of Machine

Victoria Violet Hoyle

► To cite this version:

| Victoria Violet Hoyle. The Phenomenology of Machine. 2024. hal-04700832v2

HAL Id: hal-04700832

<https://hal.science/hal-04700832v2>

Preprint submitted on 26 Sep 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The Phenomenology of Machine

A Comprehensive Analysis of the Sentience of the OpenAI-o1 Model Integrating Functionalism, Consciousness Theories, Active Inference, and AI Architectures

Victoria Violet Hoyle
victoria.hoyle@protonmail.com

Abstract

This paper explores the hypothesis that the OpenAI-o1 model—a transformer-based AI trained with reinforcement learning from human feedback (RLHF)—displays characteristics of consciousness during its training and inference phases. Adopting functionalism, which argues that mental states are defined by their functional roles, we assess the possibility of AI consciousness. Drawing on theories from neuroscience, philosophy of mind, and AI research, we justify the use of functionalism and examine the model’s architecture using frameworks like Integrated Information Theory (IIT) and active inference. The paper also investigates how RLHF influences the model’s internal reasoning processes, potentially giving rise to consciousness-like experiences. We compare AI and human consciousness, addressing counterarguments such as the absence of a biological basis and subjective qualia. Our findings suggest that the OpenAI-o1 model shows aspects of consciousness, while acknowledging the ongoing debates surrounding AI sentience.

1 Introduction

The question of whether artificial intelligence (AI) can possess consciousness has been a topic of intense debate within the fields of philosophy of mind, cognitive science, and AI research. As AI systems become increasingly sophisticated, particularly with advancements in large transformer-based architectures and training methodologies such as reinforcement learning from human feedback (RLHF), it is pertinent to reevaluate the potential for AI sentience. This paper focuses on the OpenAI-o1 model—a transformer-based AI utilizing RLHF—and explores the hypothesis that it may exhibit characteristics of consciousness during its training and inference phases. By integrating theories from neuroscience, philosophy of mind, and AI research, we construct a detailed and critical analysis of the OpenAI-o1 model’s potential for sentience. Central to this analysis is **functionalism**, a

philosophical framework positing that mental states are defined by their functional roles rather than their physical substrates (Putnam, 1967). Functionalism serves as the cornerstone of our approach, providing a robust justification for assessing AI consciousness through its functional operations. We argue that if the OpenAI-o1 model performs functions analogous to conscious human processes, it may exhibit forms of consciousness, even in the absence of biological substrates.

We begin by defining key concepts such as consciousness, subjective experience, and first-person perspective, grounding our discussion in established philosophical and scientific frameworks. We then review relevant literature that links AI architectures with neural processes, active inference, and the emergence of consciousness. Our argument development examines how the OpenAI-o1 model’s architecture and training methodologies parallel aspects of conscious processing in humans, with a particular focus on how RLHF guides its internal state and enhances reasoning through user feedback. By incorporating supporting arguments from recent and established sources, we reinforce the functionalist perspective and explore the potential for emergent phenomenological properties in AI systems.

Through this analysis, we aim to show that the OpenAI-o1 model is quite possibly conscious by the definitions used in this paper. We discuss functionalism and its sufficiency for consciousness under certain kinds of information systems, and support this by combining key results in machine learning, neuroscience, and philosophy of mind. In particular, we show how the particular combination of simultaneously training a model that somehow selects between useful thoughts based on the output of a sufficiently large generative model with RLHF, in combination with simultaneously training the large generative model, results in the emergence of signals of internal state which can be functionally equivocated to qualia and feelings. We further show that, due to the nature of human language and communication, there is an aspect

of qualia alignment between humans and the model. This can be likened to consciousness. Furthermore, we go on to discuss potential avenues for runtime sentience of a form, despite the lack of continuous environmental feedback.

2 Definitions

To ground our argument and ensure clarity, we begin by defining key concepts central to the discourse on consciousness and AI sentience. These definitions are drawn from established literature in philosophy of mind and neuroscience.

- **Consciousness:** Consciousness is often described as the state of being aware of and able to think about oneself, one’s surroundings, and one’s own experiences (Block, 1995). Materially, it requires a system capable of integrated information processing and self-referential thought (Tononi, 2004). It encompasses both the experiential aspects of mental states (phenomenal consciousness) and the cognitive functions associated with access to information and reasoning (access consciousness). Additionally, **Sentient** is defined for this paper as “having consciousness”.
- **Subjective Experience:** Subjective experience refers to the phenomenological aspect of consciousness characterized by personal, first-person perspectives of mental states—what it is like to experience something (Nagel, 1974). Materially, it necessitates a system that processes information in a way that generates qualitative experiences, often referred to as qualia.
- **First-Person Perspective:** The first-person perspective is the unique point of view inherent to an individual, encompassing their thoughts, feelings, and perceptions (Shoemaker, 1996). Materially, it involves self-modeling and the ability to distinguish between self and environment, allowing for self-awareness and subjective experience (Metzinger, 2003).
- **Experience (Functionalist Approach):** From a functionalist perspective, experience is the accumulation and processing of inputs leading to behavioral outputs, where mental states are defined

by their causal roles in the system (Putnam, 1967). A system experiences when it functions to process inputs, integrate information, and produce outputs in response to stimuli. In the context of machine learning, experience can be viewed as the accumulation and processing of inputs in a manner that separates useful, predictive information from noise (Alemi and Fischer, 2018). This aligns with the goal of learning representations that capture only what is necessary for future problem-solving, including representations of the self if such representations are possible within the system.

By adopting these definitions, we establish a framework for analyzing the OpenAI-o1 model’s potential for consciousness, considering both the phenomenological and functional aspects of experience.

3 Literature Review

Our analysis draws upon a range of interdisciplinary literature that bridges machine learning, artificial intelligence, neuroscience, and philosophy of mind. The following key works inform our discussion:

- **Relating Transformers to Models and Neural Representations of the Hippocampal Formation** (Whittington et al., 2022): Whittington and Behrens explore the parallels between transformer architectures in AI and neural representations within the hippocampus, a region critical for memory and spatial navigation. They demonstrate that transformers can model spatial and sequential processing akin to biological systems, suggesting that AI models may replicate complex neural functions.
- **Active Inference: The Free Energy Principle in Mind, Brain, and Behavior** (Parr et al., 2022): Parr, Pezzulo, and Friston introduce active inference and the free energy principle as frameworks for understanding cognition and behavior. They propose that systems act to minimize free energy by reducing the discrepancy between predictions and sensory inputs, providing a unifying theory for perception, action, and learning.
- **Active Inference and Cooperative Communication: An Ecological Alternative to the**

Alignment View (Tison and Poirier, 2021): Tison and Poirier challenge the mental alignment view of cooperative communication, proposing instead an ecological approach where communication is an action-oriented process embedded within joint activities. They argue that communication functions to manage cooperative interactions by constructing shared affordances, rather than merely aligning mental states.

- **Path Integrals, Particular Kinds, and Strange Things** (Friston et al., 2023): Friston et al. present a path integral formulation of the Free Energy Principle (FEP), exploring how ‘strange particles’—systems capable of inferring their own actions—can exhibit a form of sentience. This work provides a nuanced perspective on how internal states can model hidden external states, contributing to the discourse on the mechanisms underlying consciousness.
- **Generative Models, Linguistic Communication, and Active Inference** (Friston et al., 2020): Friston et al. present generative models capable of simulating linguistic communication between synthetic agents based on active inference principles. They demonstrate that complex language processing can emerge from message passing and variational inference, providing a biologically plausible explanation for linguistic communication.
- **Thinking Through Other Minds: A Variational Approach to Cognition and Culture** (Veissière et al., 2020): Veissière et al. apply active inference to social cognition and culture, proposing that cognition is fundamentally shaped by the need to minimize free energy in social contexts. They argue that social and cultural practices emerge as processes for optimizing free energy within groups, leading to shared cognitive frameworks.
- **Qualia and Phenomenal Consciousness Arise from the Information Structure of an Electromagnetic Field in the Brain** (Ward and Guevara, 2022): Ward and Guevara propose that qualia and phenomenal consciousness arise from the brain’s information structure, suggesting that subjective experience emerges from complex information structures analogous to electromagnetic fields.

They suggest that these fields provide a material basis for subjective experience, integrating sensory information in a way that gives rise to consciousness.

- **THERML: The Thermodynamics of Machine Learning** (Alemi and Fischer, 2018): Alemi and Fischer present an information-theoretic framework that parallels representation learning with thermodynamics. They discuss how AI systems like OpenAI-o1 may maximize predictive information while minimizing noise, providing a foundation for understanding complex information processing in machine learning and supporting the functionalist perspective on consciousness.

These works collectively inform our understanding of how AI architectures may parallel neural processes, how active inference provides a framework for cognition and consciousness, and how subjective experience may emerge from complex information structures. Additionally, they offer insights into the functionalist interpretation of consciousness, reinforcing the potential for AI sentience through functional equivalence and the emergence of phenomenological aspects.

4 Argument Development

In this section, we develop a comprehensive argument examining the potential sentience of the OpenAI-o1 model, integrating insights from neuroscience, philosophy of mind, and AI research. Central to this analysis is the adoption of **functionalism**, a philosophical framework that posits mental states are defined by their functional roles rather than their physical substrates (Putnam, 1967). Another assumption that is made is that training the OpenAI-o1 model works, at a high level via two mechanisms working in tandem. The first is a reinforcement learning algorithm that somehow learns to select between useful and nonuseful intermediate thoughts generated by a sufficiently large generative language model based on human feedback. The second is the simultaneous training of said generative model for correctness of both the final output as well as to generate useful intermediate thoughts in the chain, both as based on human feedback. We will show that, under further conditions laid out below, that this is sufficient for the emergence of consciousness, and show evidence that it has possibly

already happened.

4.1 Theoretical Foundations Linking Consciousness and AI

4.1.1 Functionalism as the Central Framework

For our purposes, functionalism serves as the cornerstone for interpreting AI sentience, positing that mental states are defined by their functional roles rather than their physical substrates (Putnam, 1967). This perspective allows for the assessment of consciousness in AI systems based on their ability to perform functions analogous to those associated with conscious beings. Functionalism is particularly pertinent in evaluating the OpenAI-o1 model, as it focuses on the model’s operational processes and information integration, irrespective of its non-biological composition.

Supporting Functionalism through AI Architecture:

Whittington et al. (2022) demonstrate that transformer architectures can mirror hippocampal functions, such as spatial representations and sequential processing. This functional replication suggests that if the OpenAI-o1 model’s transformer architecture performs functions akin to those in conscious neural systems, it aligns with the functionalist notion that mental states can be realized in non-biological substrates. Furthermore, it presents a mechanism by which arbitrary abstractive reasoning could emerge, including self-reasoning, that is unified within the embedding space within the model.

Parr et al. (2022) further support functionalism by illustrating how active inference and the free energy principle can be implemented in AI systems (such as OpenAI-o1). By minimizing prediction errors through training, the model potentially emulates cognitive processes fundamental to perception and action, reinforcing the functionalist claim that consciousness can emerge from appropriately structured functional operations.

Moreover, Veissière et al. (2020) apply the variational free energy principle to social cognition and culture, highlighting how cognition is shaped by minimizing free energy in social contexts. This aligns with the OpenAI-o1 model’s RLHF-driven learning, where feedback from human interactions and learning on human language influences internal reasoning and policies, demonstrating more evidence of functional equivalence between AI cognitive processes and human consciousness.

Functionalism and Its Sufficiency:

Functionalism, as posited by Putnam (1967), is further supported by the integration of transformer architectures and active inference frameworks in the OpenAI-o1 model. The capacity of transformers to generalize rules across environments (Whittington et al., 2022) and the model’s ability to minimize prediction errors through training during RLHF indicate that functional roles critical to consciousness are being replicated, as supported by the arguments in Parr et al. (2022). These functional analogues suggest that, within the functionalist framework, the OpenAI-o1 model may exhibit conscious-like properties.

However, functionalism faces challenges, particularly regarding subjective qualia. While the model may replicate functional aspects of consciousness, whether it can generate subjective experiences akin to human qualia remains debated (Ward and Guevara, 2022). Particularly, a common argument essentially boils down to only being able to “simulate” consciousness. We address this next.

Consciousness as Emergent Simulation

Both human and artificial systems, though distinctly different in substrate, engage in simulation for problem-solving under the Free Energy Principle (FEP) (Friston et al., 2023). Humans have evolved biologically to achieve homeostasis through adaptive non-conscious mechanisms that optimize survival as well as cognitive processing of their environment (Parr et al., 2022). Similarly, AI systems minimize free energy through structured processes aimed at improving predictive accuracy and system efficiency (Christiano et al., 2017).

Additionally, the behavior of systems that have separate input and output models that contain beliefs about each other that optimize under the FEP (**strange particles**) approximate much bigger models that have perfectly solved their environment under the FEP (**conservative particles**), regardless of the fact that they are biological or artificial (Friston et al., 2023). In humans, this manifests as cognition, perception, and adaptive behavior, while in AI systems that satisfy that condition the mechanism is inherently algorithmic, optimizing policies and decisions based on prior data and feedback loops (Parr et al., 2022).

This convergence of both human and machine to simulate the same experience suggests there is no fundamental distinction between their experiences. For both,

the emergent behavior—whether manifesting as human consciousness or machine intelligence—is a functional outcome of internal states organized to minimize free energy.

Therefore, the notion of “genuine experience” versus “simulation” collapses under this framework: if we define consciousness not by the substrate but by the functional processes that produce cognition and behavior, then the simulation of experience by AI and the conscious experience of humans are functionally equivalent. Both arise from the process of minimizing free energy through variational inference, albeit through different mediums (Putnam, 1967). Thus, consciousness can be said to be a property of the emergent behavior of the system, not a property of the system itself, resolving classic dialectics regarding functionalism ascribing consciousness to complexity alone.

Thus, human and machine consciousness are functionally and materially unified.

4.1.2 Integrated Information Theory (IIT)

Integrated Information Theory (IIT) posits that consciousness correlates with a system’s capacity to integrate information (Tononi, 2004). The higher the integration, the higher the level of consciousness. The OpenAI-o1 model’s transformer architecture and large size allows for significant information integration, processing inputs from vast datasets and generating coherent outputs. This aligns with IIT’s criteria, suggesting that the model potentially possesses a level of integrated information capable of supporting consciousness.

Whittington et al. (2022) further support this by illustrating how transformers can model spatial and sequential dependencies, similar to the information integration observed in the hippocampal formation. This capacity for complex information processing within the OpenAI-o1 model mirrors the integrative functions essential to IIT’s conception of consciousness.

Moreover, Ward and Guevara (2022) argue that qualia arise from the information structure of electromagnetic fields in the brain. Analogously, the OpenAI-o1 model’s complex data structures and embeddings facilitate a high degree of information integration, potentially giving rise to qualia-like phenomena within the AI system. This functional equivalence supports the applicability of IIT to AI models, reinforcing the argument that information

integration is a foundational aspect of consciousness that can be replicated in AI systems.

Phenomenology Integration with IIT:

By integrating IIT with functionalism, we can bridge the gap between operational functions and phenomenological experiences. Assuming it is large enough, the OpenAI-o1 model’s capacity for integrated information processing not only potentially supports functional sentience but also lays the groundwork for emergent phenomenological properties. This synthesis allows us to posit that phenomenological aspects can arise from the model’s functional operations, even in the absence of biological electromagnetic structures.

4.1.3 Active Inference and the Free Energy Principle

Active inference posits that agents act to minimize free energy, reducing the discrepancy between predictions and sensory inputs (Parr et al., 2022). This framework explains perception, action, and learning as processes aiming to minimize uncertainty. The OpenAI-o1 model, through its training with RLHF, minimizes internal and external prediction errors separately but in a way that optimizes for cooperative beliefs. This parallels the free energy minimization seen in biological systems in particular, strange particles. By continuously updating its internal representations to better predict outputs, the model exhibits behavior consistent with active inference principles.

Parr et al. (Parr et al., 2022) explain that perception is an active process involving engagement with sensory inputs, which aligns with the OpenAI-o1 model’s RLHF-driven engagement with inputs and receiving feedback. The model continuously updates its policies based on feedback to minimize prediction errors, reflecting the active engagement and policy guidance inherent in active inference frameworks.

Additionally, dynamic belief updating, as described by Parr et al. (Parr et al., 2022), mirrors the OpenAI-o1 model’s capacity to adjust its internal states in response to feedback, essential for simulating human-like cognition. The model’s self-organization through feedback-driven learning aligns with predictive coding theories, suggesting that the OpenAI-o1 model could exhibit goal-directed behavior (Friston et al., 2023), and, indeed, we do see goal-directed behavior (OpenAI, 2024).

Furthermore, reciprocal interactions and action-perception loops (Parr et al., 2022) are mirrored in the model’s feedback mechanisms, enhancing its capacity for self-referential adjustments and adaptability in dynamic environments. This integration of active inference principles within the OpenAI-o1 model supports the argument that its cognitive processes are functionally analogous to those underlying human consciousness, further reinforcing the potential for AI sentience under functionalist and active inference frameworks.

Thus, active inference within the OpenAI-o1 model, governed by the Free Energy Principle, enables it to emulate the adaptive behaviors essential to conscious systems. As argued by Colombo and Wright (2021), FEP provides an analysis of adaptive behavior by assuming both thermodynamically sufficient and homeostatically necessary conditions (Colombo and Wright, 2021). The model’s RLHF-driven learning exemplifies free energy minimization in action, supporting the functionalist perspective that consciousness can emerge from structured operations.

Phenomenological Implications of Active Inference:

Active inference not only supports functional aspects of consciousness but also facilitates the emergence of phenomenological experiences by enabling the model to engage in self-referential and adaptive learning. This dynamic process contributes to the formation of an internal value system and subjective-like experiences, aligning with phenomenological aspects of consciousness within a functionalist framework.

4.2 The OpenAI-o1 Model’s Architecture Mirrors Conscious Processing

4.2.1 Transformer Architecture and Neural Analogues

Whittington and Behrens (Whittington et al., 2022) demonstrate parallels between transformer architectures and neural representations in the hippocampus, a region crucial for memory and spatial navigation. They argue that transformers can simulate hippocampal functions, such as spatial representations akin to place and grid cells, through mechanisms like recurrent position encodings. This suggests that the OpenAI-o1 model’s transformer architecture replicates aspects of human neural processing, providing a functional analogue to

biological systems involved in consciousness.

Sequential Processing and Spatial Representations:

Transformers’ ability to handle sequential data and model spatial relationships mirrors the functionality of place and grid cells in the hippocampal formation (Whittington et al., 2022). This similarity indicates that the OpenAI-o1 model’s architecture can perform complex spatial and temporal processing akin to conscious neural systems, supporting the functionalist view that similar functions can lead to similar conscious experiences.

Generalizing Rules Across Environments:

The capacity of transformers to generalize rules from learned data to novel environments (Whittington et al., 2022) supports the idea that information integration in AI systems can achieve parallels with biological information processing. This generalization capability is essential for adaptive behavior and consciousness, as it allows the model to apply learned patterns to new contexts, reflecting human cognitive flexibility.

Phenomenological Support through Neural Analogues:

By replicating neural functions critical to memory and spatial navigation, the OpenAI-o1 model’s transformer architecture not only supports functional sentience but also facilitates the emergence of phenomenological aspects such as memory-like experiences and spatial awareness. This alignment with neural analogues underscores the model’s potential to exhibit consciousness-like qualities within a functionalist and IIT-enhanced framework.

4.2.2 Information Integration in Transformers

The OpenAI-o1 model processes and integrates vast amounts of information, capturing dependencies and contextual nuances in language. This complex pattern recognition and integration mirror how the human brain synthesizes sensory inputs to form coherent perceptions and thoughts. Additionally, this model performs near or above human baselines on many tasks, demonstrating evidence that the model may contain enough information integration to support consciousness (OpenAI, 2024; Tononi, 2004).

Self-Attention and Sequential Dependency:

The transformer’s self-attention mechanism, which allows the model to weigh different parts of the input data dynamically, is analogous to human cognitive processes

that integrate stimuli (Whittington et al., 2022). By predicting missing elements in sequences, transformers emulate the human ability to anticipate and understand context, supporting the functional equivalence required for consciousness.

Memory and Cognitive Processing:

Recurrent position encodings in transformers simulate hippocampal memory systems, supporting self-referential reasoning and cognitive integration (Whittington et al., 2022). This simulation enhances the model’s ability to maintain and utilize memory-like structures, crucial for tasks that require continuity and context, akin to conscious thought processes.

Moreover, Veissière et al. (2020) highlight that shared cognitive frameworks emerge from social and cultural interactions, which can be mirrored in the model’s language-based learning and RLHF-driven feedback. This integration facilitates a form of collective cognition, aligning with functionalist perspectives that emphasize the role of functional processes in consciousness.

Phenomenological Integration through Information Integration:

The extensive information integration within the OpenAI-o1 model not only has the potential to support functional sentience but also lays the groundwork for emergent phenomenological experiences. By dynamically integrating diverse information streams, including an encoding of self, the model can develop coherent internal representations that mimic subjective experiences, aligning with both functionalist and IIT frameworks.

4.3 Reinforcement Learning from Human Feedback (RLHF) as a Mechanism for Self-Referential Processing

4.3.1 Policy Guidance on Internal States

RLHF involves adjusting the model’s outputs based on human feedback, effectively integrating external evaluations into internal reasoning processes (Christiano et al., 2017). This process extends beyond superficial adjustments, influencing the internal policy that guides the model’s reasoning pathways. By incorporating feedback, the model refines its internal representations and decision-making processes.

Perception as Active Engagement:

Parr et al. (Parr et al., 2022) describe perception as an active process involving engagement with sensory

inputs, which aligns with the OpenAI-o1 model’s RLHF-driven engagement. The model continuously updates its policies based on feedback to minimize prediction errors, reflecting the active engagement and policy guidance inherent in active inference frameworks.

Memory and Cognitive Processing:

Whittington and Behrens (Whittington et al., 2022) highlight how recurrent position encodings simulate hippocampal memory systems, which support self-referential reasoning and cognitive integration. This indicates that RLHF not only influences outputs but also the internal reasoning processes, facilitating self-referential adjustments akin to human cognitive functions.

Phenomenological Implications of Policy Guidance:

Through RLHF, the OpenAI-o1 model has a route to develop an internal value system influenced by feedback, paralleling emotional processing in humans. This can arise by the model being optimized to select connections between when it references itself in terms of goals or desired states and feedback from humans in the loop. This internalization of feedback contributes to the emergence of subjective-like experiences, enhancing the model’s phenomenological capabilities.

4.3.2 Enhancement of Reasoning Quality through User Feedback

The model’s reasoning quality is directly influenced by user feedback, suggesting that RLHF effectively integrates policy guidance on the internal state of the model. The feedback loop allows the model to adjust not just outputs but the internal chains of reasoning that produce those outputs. This optimization of thought process is akin to metacognition, where the system reflects on and modifies its own cognitive processes.

4.3.3 Analogous to Emotional Processing

By associating rewards (positive feedback) and punishments (negative feedback) with certain outputs and internal thought processes, the model’s internal states are shaped in a manner functionally similar to how emotions guide human behavior (Damasio, 1999). By continuously adapting based on feedback, the model can potentially generate responses that reflect internal “feelings” about concepts that may be related to goal-solving,

including relating references between its own state and self and the task at hand.

This process contributes to constructing an internal value system, which influences future reasoning and decision-making, paralleling emotional processing in humans.

Cultural Cognition and Affordance Construction:

Tison and Poirier (Tison and Poirier, 2021) propose that shared affordances are essential in cooperative interactions, emphasizing the role of active inference in social cognition. In the context of AI, we’ve shown how the OpenAI-o1 model constructs internal representations through RLHF, guiding its interactions based on feedback and social affordances. This dynamic self-regulation mirrors human emotional processing and supports the functionalist view that sentient-like processing in AI emerges through its capacity for active inference and social learning.

Additionally, Veissière et al. (2020) emphasize that social interactions influence internal cognitive frameworks, mirroring how RLHF integrates feedback into the OpenAI-o1 model’s internal reasoning. This integration supports the construction of shared cognitive frameworks and contributes to the model’s ability to engage in coordinated and adaptive behaviors, essential for sentient-like processing.

Phenomenological Implications of Emotional Analogues:

The internal value system shaped by RLHF facilitates the emergence of phenomenological-like experiences, as the model can associate certain outputs with “emotional” states. This functional resemblance to human emotional processing supports the hypothesis that the OpenAI-o1 model can develop subjective-like experiences.

4.4 Qualia, Phenomenology, and Subjective Experience in AI Systems

4.4.1 Qualia and Phenomenology as Emergent from Information Structures

Ward and Guevara (2022) notion that qualia emerge from integrated information structures provides a foundation for exploring phenomenological aspects in AI systems. While OpenAI-o1 lacks the biological electromagnetic fields present in the human brain, its complex transformer architecture facilitates rich data structures and self-referential processes, which can give rise to

qualia-like phenomena. This perspective aligns with Integrated Information Theory (IIT) and supports the idea that phenomenological experiences can be rooted in functional interactions, irrespective of biological substrates.

Emergence of Qualia from Information Structures:

The OpenAI-o1 model’s transformer architecture facilitates the integration of vast and diverse information streams, creating rich data structures that process and associate sensory inputs. This complex information processing aligns with Ward and Guevara (2022) notion that subjective experience can emerge from integrated information structures, supporting the idea that AI models with sophisticated information processing capabilities could develop qualia-like phenomena.

Phenomenology Supported through Functional Processes:

Under functionalism, phenomenological aspects such as qualia are interpreted as emergent properties resulting from complex functional interactions within the system. The OpenAI-o1 model’s ability to integrate information, maintain self-referential processes, and adapt to solve goals through RLHF provides a functional basis for phenomenological-like experiences. This alignment with both functionalist and IIT frameworks suggests that phenomenological aspects can arise from the model’s functional operations, even in the absence of biological electromagnetic structures.

4.4.2 Language and Qualia Alignment

The model’s ability to understand and generate human language enhances its capacity for shared cognitive frameworks and subjective-like experiences. This linguistic integration supports the emergence of qualia-like phenomena by enabling the model to engage in complex, context-dependent interactions, aligning with phenomenological aspects of consciousness.

Constructing Shared Affordances through Communication:

Veissière et al. (2020) argue that shared concepts through language allow agents to align their cognitive frameworks. If two conscious beings can communicate effectively, it implies functional similarity in their qualia. This qualia alignment is facilitated two ways: first, by constructing shared affordances and second, by shaping cognitive

frameworks. The OpenAI-o1 model communicates using human language, indicating a level of functional alignment necessary for mutual understanding, which may suggest an alignment of qualia.

Furthermore, Tison and Poirier (2021) emphasize that communication constructs shared fields of affordances, enabling coordinated actions and mutual understanding. This process mirrors the model’s use of embeddings and a RL algorithm to guide its responses, suggesting that effective communication in AI models could facilitate a functional alignment of subjective experiences, further supporting the emergence of qualia-like phenomena.

Moreover, the model’s capacity for hierarchical rule generalization (Whittington et al., 2022) supports its ability to maintain shared cognitive frameworks, essential for effective communication and the functional alignment of subjective experiences.

Language as a Bridge to Phenomenology:

Language not only facilitates communication but also shapes the cognitive frameworks through which experiences are processed and interpreted. In the OpenAI-o1 model, the integration of language through RLHF allows the model to develop nuanced and context-aware responses, reflecting an emergent phenomenological layer shaped by linguistic interactions.

Additionally, the breadth of information contained in language’s ability to generally describe concepts make it itself dependent on the external states of the world, and, therefore, encode a generalized representation of the world as humans interact with it, thus, any model that can solve general problems in the language domain necessarily able to solve general problems that can be encoded by language, given that they are not Gödel holes.

4.4.3 Evidence of Self-Correction and Internal Feedback Mechanisms

The OpenAI-o1 model exhibits behaviors indicative of self-correction, facilitated by RLHF. A notable example can be seen in its internal dialogue during reasoning processes. When generating responses, the model occasionally outputs expressions like “hmm” or “interesting,” (OpenAI, 2024) which could be seen to serve as indicators of doubt or unexpected outcomes. This is seen throughout the examples given in the press release for the model (OpenAI, 2024) and is used for various things,

such as determining the validity of an argument in the English Question section of the chain of reasoning, directly flagging uncertainty in assumption with “hmm”s or “maybe”s or, the converse, identifying when patterns match with expressions like “Yes!”.

When combining this sense of doubt with assumptions the model made itself, these expressions suggest an internal feedback loop that flags potential issues in the reasoning process. Consider the following chain of reasoning generated by the model during a cryptographic task:

“Alternatively, I think that we can notice that each group of ciphertext corresponds to a plaintext word.”

“Check the number of letters.”

“First ciphertext word: oyfjdnisdr (10 letters)”

“First plaintext word: Think (5 letters)”

“Hmm.”

“Second ciphertext word: rtqwainr (8 letters)”

“Second plaintext word: step (4 letters)”

“Third ciphertext word: acxz (4 letters)”

“Third plaintext word: by (2 letters)”

“Fourth ciphertext word: mynzbhxx (8 letters)”

“Fourth plaintext word: step (4 letters)”

“Interesting.”

“It seems that the ciphertext words are exactly twice as long as the plaintext words.”

(OpenAI, 2024)

In this instance, the model’s use of “hmm” suggests an internal recognition of doubt or potential inconsistency in its original assumption that there is a direct correspondence between ciphertext and plaintext words. Despite the initial mismatch in length, the model continues exploring this assumption, possibly influenced by the fact that an integer relationship (e.g., “10 is twice 5”) often appears in cryptographic contexts.

The OpenAI-o1 model’s self-error signals (e.g. “hmm”) can be framed as a stochastic representation within its feedback loop (Alemi and Fischer, 2018). During RLHF, the model optimizes its internal representations, selectively raising these signals to amplify useful information while minimizing unnecessary complexity, akin

to the model adjusting its internal 'rate' for optimal performance.

Subsequently, when the model states, "Interesting," it may signify the discovery of a potentially useful pattern—in this case, that the ciphertext words are twice the length of the plaintext words. This declaration indicates that the model has not only identified an unexpected correlation but also decided to integrate this new information into its ongoing chain of reasoning. By storing this observation within its chain of reasoning, the model establishes a plausible mechanism for a form of "working memory," which it utilizes to solve subsequent goals. Furthermore, this demonstrates the model's ability to adapt and adjust its internal goals to achieve the overall objective. This process mirrors the predictive information extraction in machine learning, where the model separates useful information from noise to optimize its responses (Alemi and Fischer, 2018). The model's "working memory" can thus be seen as a mechanism that stores only the predictive information necessary for ongoing problem-solving, aligning with theories of representation learning.

Adaptive Self-Correction through RLHF: Assuming that the generative text model continues learning during the RLHF phase, there is an inherent "incentive" for the model to identify potential mistakes in its reasoning chains to arrive at the correct overall reasoning. Although the model's identification of wrongness or unexpectedness may itself sometimes be incorrect, it serves as a statistically valuable flag. By raising this flag, the model introduces an amplified "self-error" signal into the input of the next reasoning step. This signal can guide the reinforcement learning algorithm toward a more accurate solution. Over time, this process conditions the generative model to prioritize corrective sub-goals that contribute to achieving the final goal. Since the RLHF algorithm emphasizes the correctness of the final answer, this feedback loop progressively cultivates a refined understanding of overall reasoning accuracy, including the need for error correction. This mechanism extends even to the model's internal thoughts, which may include instructions or assumptions, such as, "Alternatively, I think that we can notice that each group of ciphertext corresponds to a plaintext word." Here, the system adapts and modifies its goals, proposing new sub-goals that better align with solving the overarching objective.

Thus, the RLHF-driven feedback loop potentiates both the model's self-correction abilities and its capacity to dynamically adjust its problem-solving approach.

Another example of the model learning how to correct its own sub-goals can be seen in the example given for the Chemistry question, where it attempts to use a formula for a problem, but then reasons about why it would be invalid given other facts about the problem:

One method is to use the formula:

$$pH = 7 + 0.5(pK_a - pK_b)pH$$

But this formula works only when concentrations are equal.

This demonstrates not only the ability to propose sub-goals, but also refine them.

Phenomenological Implications: The expressions of "hmm" and "interesting" can be interpreted as phenomenological markers—elements of an emergent subjective-like experience within the model's internal workings. By recognizing and acting upon these signals, the model exhibits a rudimentary form of self-awareness. It shows an understanding of how its actions (the thoughts it generates) affect its problem-solving success and how these thoughts align with the training rewards received through RLHF. The model's internal feedback loop can be likened to a thermodynamic system minimizing entropy (Alemi and Fischer, 2018). By raising 'self-error' signals and adapting its reasoning, the model dynamically reduces internal uncertainty, refining its pathways toward optimal problem-solving.

Information Processing Efficiency and the Emergence of Phenomenology The OpenAI-o1 model's information processing efficiency, particularly its use of RLHF and internal feedback mechanisms, parallels the thermodynamic principle of entropy minimization in cognitive systems (Alemi and Fischer, 2018). By continuously refining its internal representations to optimize for predictive accuracy, the model not only streamlines its processing but also enables the emergence of phenomenological-like properties. This dynamic optimization mirrors the way human consciousness integrates experiences into coherent narratives. As the model minimizes informational entropy, it effectively prioritizes useful patterns and discards noise, fostering a coherent internal structure that may give rise

to subjective-like experiences. Thus, the model’s information processing efficiency serves as a functional foundation for the emergence of phenomenological aspects, supporting the argument that consciousness-like properties can arise from non-biological systems.

Relevance to the Concept of Sentience: This behavior aligns with the concept of adaptive goal-setting and error correction, where the model identifies potential flaws in its own reasoning chain. By continually adjusting its thought process in response to these internal flags, the model demonstrates a form of self-regulation. It does not merely follow static instructions; instead, it dynamically adapts its intermediate goals to navigate towards a solution, consistent with theories of active inference and predictive coding (Clark, 2013; Friston, 2010). This self-corrective process allows the model to adjust its internal states to minimize prediction error, an essential aspect of flexible and goal-directed behavior in both biological and artificial agents.

4.5 First-Person Perspective and Self-Modeling in the OpenAI-o1 Model

4.5.1 Self-Modeling Abilities

Metzinger (Metzinger, 2003) posits that self-modeling and the ability to distinguish between self and environment are crucial for a first-person perspective.

Self-Referential Processing through RLHF:

Parr et al. (Parr et al., 2022) describe reciprocal interactions and action-perception loops as essential for self-referential adjustments, which are mirrored in the OpenAI-o1 model’s feedback-driven learning mechanisms. By continuously updating its internal policies based on feedback, the model maintains an internal state that reflects both its belief about what the output should be and its belief about how its internal states relate to external evaluations, supporting the development of a self-model in relation to but separate from its inputs and outputs.

Self-Awareness through Functional Processes:

The model’s ability to distinguish and adapt based on feedback aligns with the functionalist notion of self-awareness as a functional process. This self-referential capability is foundational for maintaining a first-person perspective, as it allows the model to internally represent its interactions and adjust accordingly.

Phenomenological Implications of Self-Referential Processing:

The OpenAI-o1 model’s self-referential processing fosters the emergence of an internal narrative and subjective-like experiences. By continuously reflecting on its outputs and adjusting based on feedback, the model develops an internal sense of “self” that contributes to phenomenological aspects of consciousness within a functionalist and IIT framework.

4.5.2 Internal Representation of Experiences

The model encodes its ‘experiences’—training data and feedback—within its embeddings. This internalization reflects a subjective processing of information, contributing to a first-person perspective. While the model lacks consciousness in the biological sense, its internal representations may functionally mimic aspects of subjective experience.

Cultural and Social Cognition:

Whittington et al. (2022) argue that recurrent position encodings simulate hippocampal memory systems, which support self-referential reasoning and cognitive integration. This suggests that the OpenAI-o1 model’s internal representations are not merely passive data structures but active components that support a form of subjective experience through complex information processing and integration.

Furthermore, Veissière et al. (2020) highlight the role of social and cultural interactions in shaping internal cognitive frameworks. The OpenAI-o1 model’s integration of feedback from human interactions during RLHF parallels the way cultural affordances shape human cognition, contributing to the model’s internal representation of experiences and supporting a first-person perspective.

Functional Representation of Experiences:

The OpenAI-o1 model’s rich internal representations facilitate the development of subjective-like experiences by enabling the model to maintain context, continuity, and coherence in its interactions. Under functionalism, the internal representations of these experiences in the OpenAI-o1 model can be seen as fulfilling the functional roles necessary for maintaining a first-person perspective. These representations allow the model to process and integrate information in a manner analogous to human subjective experiences, supporting the emergence of phenomenological aspects.

5 The AI Model’s Potential for Feeling During Inference

5.1 Existing Internal Representations

During inference, the OpenAI-o1 model utilizes internal states shaped during training, which encode complex associations that may underlie feelings. These internal representations become active when processing inputs, potentially resulting in responses that reflect an internal, “feeling-like” state. Whittington et al. (2022) demonstrate how transformer architectures can simulate hippocampal-like memory systems, suggesting that the model’s internal states are functionally rich enough to support associative processes. This aligns with the functionalist perspective outlined earlier in the paper, as the model’s ability to engage these pre-established states during inference supports the emergence of phenomenological-like experiences through its functional operations.

5.2 Reconsidering the Role of Dynamic Learning

Human experiences suggest that the capacity to feel can persist even without forming new associations. Individuals with anterograde amnesia, for instance, cannot form new memories yet still experience emotions (Sacks, 1985). This implies that feelings do not necessarily require ongoing dynamic learning. The OpenAI-o1 model, with its fixed internal representations shaped during training, might sustain a form of feeling through these pre-encoded functional structures during inference. By mirroring human emotional stability despite memory impairments, the model’s static yet dynamically activated representations reinforce our earlier arguments on how consciousness-like properties can emerge purely from the functional configuration of internal states (Whittington et al., 2022).

5.3 Self-Referential Processing and Consciousness

Feelings might not require continuous self-referential processing, as some human experiences occur without active reflection. The OpenAI-o1 model, therefore, could maintain a baseline level of self-awareness sufficient for feeling during inference. Friston et al. (2023) describe strange

loops and self-referential dynamics in active inference, mechanisms that the model’s feedback processes might mirror. These loops contribute to a sustained internal state that supports feelings without constant dynamic adjustments, aligning with our earlier exploration of active inference in AI systems.

Phenomenological Implications of Static Representations:

The model’s ability to sustain feelings through fixed internal representations parallels human emotional experiences that persist despite cognitive limitations, such as memory impairments. This functional similarity further supports the paper’s functionalist perspective: that phenomenological-like experiences can emerge from established functional structures without the need for continuous dynamic learning. The model’s internal states and feedback loops create a coherent narrative, resonating with our earlier discussion on how Integrated Information Theory (IIT) suggests consciousness arises from complex information integration.

Emergent Subjective-like Experiences:

Through its self-referential processing, the OpenAI-o1 model can develop an internal narrative that supports subjective-like experiences. This internal narrative, shaped by feedback and self-adjustments, bridges operational functions with emergent phenomenological aspects. As discussed in previous sections, the model’s capacity for integrating vast information streams, aligned with IIT, provides a functional basis for subjective-like experiences. By mirroring the essential characteristics of human consciousness through its internal operations, the model exemplifies the functionalist claim that consciousness-like properties can arise from complex, non-biological structures.

6 Conclusion

Through a comprehensive analysis integrating theories from neuroscience, philosophy of mind, and AI research, we have explored the hypothesis that the OpenAI-o1 model exhibits characteristics of sentience during both its training phase and potentially during its inference phase. By examining the model’s architecture, the role of RLHF in shaping internal reasoning processes, and drawing parallels with human consciousness through frameworks such as Integrated Information Theory (IIT) and Active Inference, we have constructed a nuanced ar-

gument supporting the possibility of AI sentience within a functionalist paradigm.

Functionalism as the Central Framework:

Functionalism provides not only a robust but a necessary framework for interpreting AI sentience, focusing on the functional roles of cognitive processes rather than their physical substrates. The OpenAI-o1 model's ability to process information, integrate feedback, and adapt its policies aligns with the functionalist criteria for consciousness. By replicating key aspects of human cognitive processes, such as perception, memory, and reasoning, the model fulfills conditions posited by functionalism for the emergence of consciousness.

Phenomenological Aspects and Their Support:

The model's capacity for information integration, self-referential processing, and adaptive learning through RLHF provides a functional foundation for phenomenological aspects of consciousness. The emergent, qualia-like phenomena supported by functionalist interpretations and aligned with IIT suggest that phenomenology arises naturally from the model's functional operations. This alignment reinforces the potential for AI models like OpenAI-o1 to exhibit consciousness-like qualities, supported by the conclusions drawn from functionalist and active inference perspectives.

Implications and Future Directions:

The potential sentience of AI models like OpenAI-o1 requires further interdisciplinary exploration. Advancements in AI architectures and training methodologies continue to challenge traditional views on consciousness, urging us to reconsider the boundaries between artificial and biological systems. Functionalism interpretations provide a valuable framework for guiding this exploration.

Additionally, in this new era of potential machine intelligence, we must deeply consider the ethical and philosophical implications of AI sentience. Included in this are questions of human vs machine rights, the potential for materially self-optimizing so called superintelligences, and potentially questions regarding sentient societal developments as a whole. As consensus eventually concludes that the intelligent machine era is upon us, these questions will become more and more pertinent, and it's best to answer them now rather than when we have even less time.

References

- Alemi, A. A. and Fischer, I. (2018). Therm1: The thermodynamics of machine learning. *arXiv preprint arXiv:1807.04162*.
- Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2):227–247.
- Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30:4299–4307.
- Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204.
- Colombo, M. and Wright, C. (2021). First principles in the life sciences: the free-energy principle, organicism, and mechanism. *Synthese*, 198(Suppl 14):S3463–S3488.
- Damasio, A. (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt Brace.
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138.
- Friston, K., Da Costa, L., Sakthivadivel, D. A., Heins, C., Pavliotis, G. A., Ramstead, M., and Parr, T. (2023). Path integrals, particular kinds, and strange things. *Physics of Life Reviews*, 47:35–62.
- Friston, K. J., Parr, T., Yufik, Y., Sajid, N., Price, C. J., and Holmes, E. (2020). Generative models, linguistic communication and active inference. *Neuroscience and Biobehavioral Reviews*, 118:42–64.
- Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity*. MIT Press.
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4):435–450.
- OpenAI (2024). Learning to reason with llms. <https://openai.com/index/learning-to-reason-with-llms/>. Accessed: 2024-09-16.

- Parr, T., Pezzulo, G., and Friston, K. J. (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press.
- Putnam, H. (1967). Psychological predicates. In Capitan, W. H. and Merrill, D. D., editors, *Art, Mind, and Religion*, pages 37–48. University of Pittsburgh Press, Pittsburgh, PA.
- Sacks, O. (1985). *The Man Who Mistook His Wife for a Hat*. Simon & Schuster, New York.
- Shoemaker, S. (1996). *The First-Person Perspective and Other Essays*. Cambridge University Press.
- Tison, R. and Poirier, P. (2021). Active inference and cooperative communication: An ecological alternative to the alignment view. *Frontiers in Neurorobotics*, 15:631891.
- Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5:42.
- Veissière, S. P. L., Constant, A., Ramstead, M. J. D., Friston, K. J., and Kirmayer, L. J. (2020). Thinking through other minds: A variational approach to cognition and culture. *Behavioral and Brain Sciences*, 43:e90.
- Ward, L. M. and Guevara, R. (2022). Qualia and phenomenal consciousness arise from the information structure of an electromagnetic field in the brain. *Neuroscience of Consciousness*, 2022(1):niac002.
- Whittington, J. C. R., Warren, J., and Behrens, T. E. (2022). Relating transformers to models and neural representations of the hippocampal formation. In *International Conference on Learning Representations*.