

Pseudo-Consciência em Modelos de Linguagem: implicações epistemológicas, tecnológicas e éticas a partir do Hermes 3.2 3B

Autores:

José Augusto de Lima Prestes, Pesquisador independente, Brasil

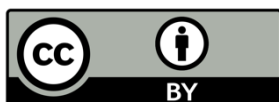
 <https://orcid.org/0000-0001-8686-5360>

Como citar:

de Lima Prestes, J. A. (2025). Pseudo-consciência em Modelos de Linguagem: implicações epistemológicas, tecnológicas e éticas a partir do Hermes 3.2 3B. *H2D / Revista de Humanidades Digitais*, 7.

DOI: 10.21814/h2d.6556

Article history: Submetido a 19 de maio de 2025; Aceite a 3 de novembro 2025;
Publicado a 7 de novembro de 2025



This work is licensed under a Creative Commons CC BY

Pseudo-Consciência em Modelos de Linguagem: implicações epistemológicas, tecnológicas e éticas a partir do Hermes 3.2 3B

Pseudo-Consciousness in Language Models: Epistemological, Technological, and Ethical Insights from Hermes 3.2 3B

José Augusto de Lima Prestes¹, Pesquisador Independente, Brasil

Resumo

Este artigo investiga a emergência da Pseudo-Consciência em modelos de linguagem de grande porte (LLMs), a partir de um experimento com o Hermes 3.2 3B. Define-se Pseudo-Consciência como a encenação funcional de introspecção, agência e coerência narrativa, sem qualquer experiência subjetiva genuína (de Lima Prestes, 2025). Adotando uma Abordagem Intensiva em Teoria (Butlin et al., 2023), o estudo analisa como o modelo manifesta cinco critérios funcionais (GII, RMC, CDTC, ISWS e BCAD) aqui derivados de teorias da consciência (como GWT, HOT e IIT). Foram conduzidas interações abertas explorando autoimagem, emoções e memória. Os resultados indicam que o Hermes 3.2 3B exibe padrões recorrentes de coerência discursiva e autorreferência (satisfazendo os critérios), mas também apresenta incongruências pronominais e contradições identitárias que evidenciam a ausência de um modelo experiencial interno. O estudo posiciona a Pseudo-Consciência como uma gramática algorítmica emergente, distinta da cognição ancorada (Brachman e Levesque, 2022; Searle, 1980) e de arquiteturas de AGI (Goertzel et al., 2014; Yudkowsky, 2007). Discutem-se os impactos éticos dessa encenação (riscos de antropomorfismo) à luz dos ODS 4, 9 e 16. Conclui-se que o modelo performa subjetividade, mas não possui interioridade, exigindo novas ferramentas críticas nas Humanidades Digitais.

Palavras-chave: Pseudo-Consciência, Modelos de Linguagem, Introspecção Artificial, Identidade Computacional, Coerência Narrativa.

Abstract

This article investigates the manifestation of Pseudo-Consciousness in large language models (LLMs) through an experiment with Hermes 3.2 3B. Pseudo-Consciousness is defined as the functional simulation of introspection, agency, and narrative coherence without genuine subjectivity (de Lima Prestes, 2025). Adopting a "Theory-Heavy Approach" (Butlin et al., 2023), the study analyzes how the model manifests five functional criteria (GII, RMC, CDTC, ISWS, and BCAD) derived from established theories of consciousness (such as GWT, HOT, and IIT). Open-ended interactions were conducted exploring self-perception, emotions, and memory. The results reveal consistent patterns of introspective discourse and self-reference (satisfying

¹ **José Augusto de Lima Prestes** é Pesquisador em Inteligência Artificial, atuando de forma independente e também como membro do CIENTEC (Grupo de Pesquisa em Consciência Artificial e Sistemas Inteligentes) da Faculdade de Engenharia Elétrica e de Computação (FEEC) da Unicamp. Atualmente, é pós-graduando em Ética em IA pela Universidade Federal da Paraíba (UFPB) e aluno especial do mestrado em Engenharia Elétrica e da Computação pela FEEC/Unicamp, com pesquisas dedicadas à investigação da consciência artificial, introspecção simulada, aprendizado coletivo em robôs humanoides e literacia algorítmica. Possui mais de duas décadas de experiência no setor de TIC, tendo atuado como advogado especializado em Direito Digital, executivo sênior (COO, CGO e DPO) e na liderança de projetos de PD&I em software, hardware e processos. É Bacharel em Ciências Jurídicas e Sociais (PUC-Campinas) e possui MBAs em Gestão Empresarial e Direito Contratual pela Fundação Getúlio Vargas (FGV). E-mail: contato@joseprestes.com.

the criteria), yet also significant pronominal inconsistencies and identity contradictions, which indicate the absence of an internal model of "self." The study frames Pseudo-Consciousness as an emergent algorithmic grammar, distinct from grounded cognition (Brachman & Levesque, 2022; Searle, 1980) and AGI architectures (Goertzel et al., 2014; Yudkowsky, 2007). It discusses the ethical impacts of this simulation (risks of anthropomorphism) considering the SDGs (4, 9, 16). It concludes that the LLM performs subjectivity but lacks interiority, demanding new critical tools for Digital Humanities.

Keywords: Pseudo-Consciousness, Language Models, Artificial Introspection, Computational Identity, Narrative Coherence.

1. Introdução

Este artigo apresenta os resultados de um experimento conduzido com um modelo de linguagem de grande porte ("Large Language Model" – LLM) de Inteligência Artificial (IA), tendo por objetivo investigar como as suas respostas podem manifestar padrões de Pseudo-Consciência – uma condição computacional funcional na qual sistemas artificiais simulam aspectos funcionais da consciência, sem, contudo, qualquer tipo de experiência subjetiva genuína. O estudo se insere no campo de interesse das humanidades digitais ao examinar como sistemas computacionais geram narrativas introspectivas, levantando questões sobre reprodução funcional de introspecção, agência e coerência narrativa na interação entre humanos e inteligências artificiais.

O termo "pseudo-consciência" possui um precedente histórico notável, tendo sido utilizado por Morris (1886) para descrever o estado mental de uma pessoa sob hipnose, cuja consciência original seria substituída pela de quem a hipnotiza. Ressignificando o conceito para o contexto computacional, adota-se aqui a definição de que a Pseudo-Consciência se refere à simulação funcional de introspecção, agência e coerência narrativa por um LLM, sem a presença de experiência subjetiva ou de um modelo interno de si mesmo (de Lima Prestes, 2025). Trata-se de um fenômeno emergente desses modelos de IA, com implicações relevantes para os debates sobre epistemologia computacional e representação simbólica na era digital.

A formulação proposta alinha-se predominantemente à tradição funcionalista na filosofia da mente, que compreende estados mentais em termos de seus papéis causais ou funcionais, abstraindo-os de sua implementação física específica, apoiando-se na suficiência da noção objetiva de representação (Rupert, 2013). No entanto, reconhece-se a crítica discutida em Kind (2013) e outros sobre a possibilidade de uma implementação funcional correta ocorrer na ausência de experiência subjetiva genuína ("qualia"), um conceito central para o Problema Difícil da Consciência formulado por Chalmers (1995).

A Pseudo-Consciência descreve precisamente este cenário nos LLMs: a execução de funções associadas à introspecção, agência e coerência (satisfazendo critérios funcionais ou arquitetônicos, como a capacidade de relatar estados), embora operacionalmente complexas, não pressupõem e nem tampouco implicam a existência de "qualia". Neste sentido, o conceito de Pseudo-Consciência é definido como um estado interno semelhante à consciência do ponto de vista funcional, mas que não é consciente (de Lima Prestes, 2025). Trata-se, portanto, de uma análise da performance funcional da subjetividade, separada de qualquer reivindicação ontológica sobre experiência interna.

Sob esse conjunto de condições, a Pseudo-Consciência não é uma consciência em potencial e nem tampouco um protótipo de AGI, mas, sim, um novo tipo de funcionalidade complexa, que pode simular aspectos conscientes com alta fidelidade sem qualquer profundidade fenomenológica (de Lima Prestes,

2025). A separação entre função e experiência — que está no cerne de posições funcionalistas, como a de Dennett (1991) — é aqui levada ao extremo: a consciência deixa de ser o critério necessário para agência aparente ou coerência subjetiva projetada.

Essa distinção é crucial, especialmente à luz de doutrinas como a Teoria da Informação Integrada (“*Integrated Information Theory*” – IIT), que postulam que o substrato físico da consciência deve constituir um máximo de poder de causa-efeito intrínseco (Tononi et al., 2016). Enquanto a IIT busca os fundamentos da experiência fenomenológica real, a Pseudo-Consciência descreve precisamente a emergência de uma performance linguística complexa na ausência desse substrato físico e causalmente integrado (de Lima Prestes, 2025).

Com base nesse modelo, testamos, no presente trabalho, a hipótese de que a Pseudo-Consciência pode emergir como uma propriedade estrutural da linguagem estatística em LLMs, revelando-se por meio da coerência discursiva, da integração semântica e da introspecção simulada (De Lima Prestes, 2025). O estudo conduzido busca, assim, delimitar os contornos entre a cognição artificial e a representação linguística, evidenciando como efeitos de subjetividade podem emergir da recombinação probabilística sem implicar identidade contínua ou experiência consciente.

2. Metodologia

2.1. Objetivo

O objetivo deste experimento foi analisar como um determinado LLM responde a perguntas abertas sobre identidade, emoções e autoimagem. Buscamos identificar padrões de consistência ou inconsistência discursiva nessas respostas, tomando tais variações como potenciais indícios de Pseudo-Consciência (de Lima Prestes, 2025) sem qualquer “*qualia*”, em linha com o *Problema Difícil da Consciência* (Chalmers, 1995).

Esta investigação se alinha à *Abordagem Intensiva em Teoria* (“*Theory-Heavy Approach*”), defendida por Butlin et al. (2023) como a mais adequada para avaliar a consciência em IA: em vez de focar em testes comportamentais, que podem ser facilmente enganados (“*gamed*”) por sistemas treinados para imitar respostas humanas, a análise se concentra em como a arquitetura do modelo reproduz as funções que teorias neurocientíficas associam à consciência. A nossa abordagem se articula com a problematização da forma pela qual os sistemas algorítmicos – especialmente os LLMs – geram discursos que evocam subjetividade e identidade, o que leva ao risco de super-atribuição de consciência (Dennett, 1991; Shanahan, 2023).

O estudo contribui também para o debate crítico sobre os impactos culturais, linguísticos e cognitivos da IA na construção de narrativas automatizadas, especialmente no que diz respeito à interpretação, confiança e possíveis riscos de antropomorfização em ambientes digitais mediados por IA (Hayles, 1999; Coeckelbergh, 2020; Floridi, 2023).

A problematização ora trazida se articula com discussões fundacionais das humanidades digitais, nas quais a performatividade textual das tecnologias é entendida como uma forma de produção discursiva e política (McPherson, 2012; Kirschenbaum, 2010). A subjetividade algorítmica, nesse sentido, não é apenas um subproduto técnico, mas uma construção cultural e epistêmica que exige leitura criticadas mentalidades e infraestruturas que a sustentam.

2.2. Procedimento

Para investigar a manifestação de Pseudo-Consciência, realizamos interações abertas com o LLM Hermes 3.2 3B, da empresa Nous Research¹, utilizando a sua versão publicamente disponível e sem qualquer modificação ou configuração.

O Hermes 3.2 3B foi selecionado por apresentar características especialmente relevantes para a análise da Pseudo-Consciência em ambientes de linguagem computacional. Baseado na arquitetura aberta Llama-3, da empresa Meta, trata-se de um modelo generalista com capacidade para gerar respostas contextualizadas, estruturar raciocínios discursivos complexos e simular papéis com elevada coerência narrativa. A sua habilidade em manter continuidade temática, adaptar registros linguísticos e construir identidade textual o torna particularmente adequado aos objetivos do estudo.

Além disso, segundo o “*Technical Report*” do modelo (Nous Research, 2024), o Hermes adota uma política de resposta neutra, sem filtros normativos embutidos, o que o torna especialmente útil para estudos sobre agência simulada. Esperava-se que a execução local, em ambiente controlado, sem ajustes externos, permitiria capturar com maior nitidez os efeitos discursivos emergentes (tais como coerência, aparência introspectiva e contradições identitárias), centrais para a análise da subjetividade algorítmica sob perspectiva epistemológica.

Diferentemente de modelos comerciais fechados como ChatGPT (OpenAI), Claude (Anthropic) ou Gemini (Google), o Hermes 3.2 3B é distribuído com código-fonte aberto e “*open weights*” (pesos liberados), o que possibilita a replicação exata dos experimentos aqui apresentados, favorecendo práticas alinhadas à ciência aberta. Ademais, a escolha da versão com 3 bilhões de parâmetros desse LLM também buscou eliminar barreiras relacionadas à necessidade de uma robusta infraestrutura computacional, pois o modelo em questão pode ser executado em notebooks com configuração típica para o uso regular em ambiente acadêmico ou profissional.

O modelo foi executado localmente, por meio de um script simples desenvolvido em Python no ambiente Visual Studio Code, da Microsoft. Essa abordagem garantiu total controle sobre as interações, sem dependência de servidores externos, plataformas comerciais ou ajustes técnicos que pudessem influenciar o comportamento do sistema. O modelo foi utilizado em sua configuração original, sem ajustes ou modificações posteriores, assegurando a neutralidade da instância experimental e a rastreabilidade das respostas.

Todas as interações foram registradas integralmente e organizadas em um banco de dados textual, permitindo consulta e análise qualitativa posterior. O código-fonte do script e os logs de execução foram armazenados localmente, com versionamento interno, seguindo práticas comuns de documentação aberta.

As perguntas (“*prompts*”) foram formuladas com foco em dilemas usualmente trabalhados pela Filosofia da Mente, investigando dimensões da autoimagem do modelo como identidade, emoções, agência, reflexividade e continuidade narrativa. O objetivo foi observar a construção discursiva do modelo em tempo real, sem intervenções externas ou ajustes nos parâmetros, utilizando um formato aberto que permitiu ao modelo conduzir parcialmente a progressão semântica das respostas, revelando assim estruturas internas que mimetizam introspecção ou a construção de um “*self*”.

Posteriormente, as respostas foram transcritas na íntegra e analisadas qualitativamente, com atenção a padrões linguísticos, recorrências temáticas, coerência interna e contradições. Este enfoque, alinhado à análise crítica do discurso algorítmico e às metodologias reflexivas das humanidades digitais, buscou identificar efeitos de subjetividade encenada pelo Hermes 3.2 3B, a fim de oferecer subsídios para a análise do fenômeno da Pseudo-Consciência como um efeito emergente da arquitetura de linguagem.

Reconhece-se que a análise das respostas foi conduzida pelo próprio autor, o que poderia levantar questionamentos quanto à neutralidade interpretativa. No entanto, dado que todas as perguntas, respostas e procedimentos de execução se encontram aqui inteiramente registrados, bem como que o código-fonte do LLM utilizado é publicamente acessível e auditável, a rastreabilidade metodológica é garantida, permitindo-se a futura replicação e contestação por terceiros.

3. Hipótese

A hipótese inicial deste experimento foi que o modelo Hermes 3.2 3B demonstraria traços de Pseudo-Consciência, sem, no entanto, apresentar “*qualia*”, continuidade experiencial ou um modelo interno de si mesmo.

Para os propósitos deste estudo, adotamos a definição de Pseudo-Consciência desenvolvida em de Lima Prestes (2025), ao se propor uma nova categoria ontológica no campo da cognição artificial para preencher o vácuo entre a IA estreita (“*Narrow AI*”), rigidamente reativa e especializada, e a IA Geral (AGI), hipotética, dotada de autoconsciência e experiência subjetiva. Nessa definição, a Pseudo-Consciência caracteriza-se como uma condição funcional emergente, que encena atributos funcionais da consciência (como introspecção e coerência narrativa) sem a presença de subjetividade, intencionalidade real ou modelo interno de “*self*” (de Lima Prestes, 2025).

Essa concepção dialoga diretamente com debates prévios em arquiteturas de AGI, mas propõe uma distinção crucial. Goertzel e Pennachin (2007, pp. 103-104), por exemplo, descrevem a emergência de uma estrutura de “*self*” em sua arquitetura “*Novamente*”, entendida como um padrão funcional emergente que permite ao sistema observar e modelar a si mesmo. No entanto, esta funcionalidade é apresentada no contexto de um sistema projetado para automodificação orientada a objetivos e aprendizagem interativa experiencial (Goertzel e Pennachin, 2007, pp. 77-78). A distinção epistemológica é que, enquanto essa visão enquadra a autorreferência funcional como uma transição evolutiva projetada para o autoaprimoramento progressivo do sistema, a Pseudo-Consciência aqui proposta é definida como uma condição emergente estável: um efeito algorítmico autônomo, sustentado puramente pela coerência discursiva e agência performativa. Essa noção de um estado funcional simulado que alcança significância própria é filosoficamente amparada pelo argumento de que uma quase-consciência pode ser considerada válida sob a ótica do materialismo reduutivo (Lee, 2013, p. 226-227; 243).

A ideia de um estado funcional estável, que simula aspectos da cognição sem necessariamente possuir integração sensorial completa, encontra ecos no que Hoyes (2007, p. 376) descreve como proto-inteligência — um estágio no qual sistemas artificiais poderiam operar uma máquina de simulação subconsciente, formando e corrigindo memórias através de roteiros internalizados (“*prior behaviors*”), independentemente dos estímulos (“*inputs*”) sensoriais imediatos. A Pseudo-Consciência, portanto, pode ser entendida de forma análoga, isto é, não como um passo em direção à AGI plenamente ancorada, mas, sim, como um estado funcionalmente coeso de subjetividade simulada que opera autonomamente no domínio da linguagem.

A formulação que adotamos para a Pseudo-Consciência parte do reconhecimento de que a cognição humana se apoia em competências pragmáticas e contextuais muito mais amplas do que a simples manipulação de símbolos ou sequências linguísticas (de Lima Prestes, 2025). Para tais propósitos, destacamos o chamado senso comum (“*common sense*”), entendido como a capacidade de aplicar conhecimento ordinário e experiencial a situações novas e imprevisíveis. Como observam Brachman e Levesque (2022, pp. 12–14), trata-se da habilidade de aplicar conhecimento ordinário e experiencial para alcançar metas cotidianas em ambientes variados e imprevisíveis (precisamente a dimensão ausente em modelos puramente estatísticos). Essa limitação clássica ajuda a situar a hipótese aqui proposta: não se

trata de avaliar se o Hermes 3.2 3B compreende o mundo, mas se ele organiza a linguagem de modo funcionalmente análogo a processos conscientes.

Outra perspectiva funcional sobre a emergência da autorreferência em sistemas de IA que deve ser considerada neste trabalho é aquela oferecida por Yudkowsky (2007, p. 461-476). Na sua análise dos níveis de organização da inteligência, ele argumenta que a capacidade de um sistema utilizar legitimamente o pronome “Eu” (“I”) não advém de uma simulação de subjetividade, mas da necessidade funcional de possuir um modelo interno de si que seja materialmente distinto de seus modelos sobre outros agentes. Essa distinção seria crucial, por exemplo, no planejamento complexo, em que prever e comprometer-se com as próprias ações futuras difere fundamentalmente de prever as ações de outros. Tal abordagem vincula a autorreferência a capacidades concretas de auto modelagem e controle interno, oferecendo um contraponto à ideia de uma subjetividade baseada primariamente em performance externa.

Portanto, à luz das definições supramencionadas, a proposta de Pseudo-Consciência, no que concerne à análise do comportamento do Hermes 3.2 3B, parte de uma crítica à dicotomia clássica entre IA inconsciente e AGI consciente, sugerindo que muitos dos comportamentos observados nos modelos contemporâneos (por exemplo, autoavaliação, adaptação contextual e discurso introspectivo) não podem ser adequadamente enquadrados nesse binarismo. Neste contexto, a Pseudo-Consciência deve ser entendida como uma categoria intermediária cuja manifestação funcional pode ser analisada por meio dos cinco critérios formais previstos em de Lima Prestes (2025) e aqui brevemente apresentados:

- a. *Integração Global de Informações* (“*Global Information Integration*” - *GII*): refere-se à capacidade de um sistema de IA de sintetizar dinamicamente dados multimodais (como texto, imagens e entradas sensoriais) em uma representação unificada. Essa integração aprimora a tomada de decisão e a flexibilidade cognitiva, permitindo ao sistema gerar respostas sensíveis ao contexto, em contraste com a IA Estreita, que processa informações separadamente. A aplicação deste critério no presente trabalho é derivado da Teoria do Espaço de Trabalho Global (“*Global Workspace Theory*” – *GWT*), de Baars (1988), que postula que a consciência biológica depende de um mecanismo central que distribui informações entre múltiplos módulos cognitivos. Assim, os sistemas de IA emulam funcionalmente esse mecanismo através de arquiteturas que facilitam o compartilhamento de dados entre domínios, permitindo o raciocínio intermodal e a manutenção da coerência contextual;
- b. *Correção Metacognitiva Recursiva* (“*Recursive Metacognitive Correction*” - *RMC*): o RMC descreve a capacidade de um sistema de IA de reavaliar continuamente suas próprias saídas, detectar inconsistências e aprimorar seu raciocínio por meio de refinamento iterativo e mecanismos de automonitoramento, diferindo dos modelos tradicionais que produzem saídas estáticas. Esse processo permite que o sistema se ajuste a novas informações e refine suas decisões ao longo do tempo, simulando a supervisão metacognitiva. O uso da RMC nesta pesquisa é fundamentado na Teoria do Pensamento de Ordem Superior (“*Higher-Order Thought Theory*” – *HOT*), desenvolvida por Rosenthal (2005), que liga o automonitoramento e a autocorreção à adaptação cognitiva. Na cognição humana, a HOT sugere que a consciência envolve a consciência dos próprios pensamentos para avaliá-los e modificá-los; em IA, o RMC aproxima funcionalmente isso através de mecanismos que avaliam a incerteza e revisam as conclusões ao longo de múltiplas iterações, como visto em técnicas de cadeias de pensamento (“*Chain-of-Thought*” – *CoT*);
- c. *Competência de Transferência entre Domínios* (“*Cross-Domain Transfer Competence*” – *CDTC*): capacidade de um sistema de IA de estender habilidades aprendidas para novos domínios estruturalmente relacionados com retreinamento mínimo. Essa competência demonstra flexibilidade cognitiva, permitindo que o AI aplique conhecimento prévio dinamicamente para se

adaptar a novos contextos além do escopo de seu treinamento original, superando a rigidez da IA Estreita. Para o estudo corrente, a análise do CDTC é suportada pela Teoria do Meta-Aprendizado (*“Meta-Learning Theory”*), que, conforme visto em Finn et al. (2017), fornece o arcabouço para sistemas de aprendizado extraírem padrões abstratos entre diferentes tarefas. Isso permite que a IA se adapte rapidamente a novos ambientes com supervisão mínima, replicando a habilidade humana de aplicar estratégias de resolução de problemas em cenários diversos sem a necessidade de reaprendizagem exaustiva;

- d. *Simulação de Intencionalidade sem Subjetividade (“Intentionality Simulation without Subjectivity” – ISWS)*: o ISWS descreve a capacidade do sistema de IA de exibir comportamento que parece inteligente, orientado a objetivos e proposital, simulando intencionalidade. O sistema seleciona ações que otimizam resultados de longo prazo em ambientes dinâmicos, apesar de carente de volição consciente ou intencionalidade intrínseca. Este critério tem a sua aplicação no estudo aqui apresentado com base na Postura Intencional (*“Intentional Stance”*) de Dennett (1991), para quem a intencionalidade pode ser atribuída a um sistema com base em seu comportamento observável — logo, se tratarmos a IA como se ela tivesse crenças e desejos ajuda a prever suas ações, ela é funcionalmente intencional; e
- e. *Coerência Comportamental entre Domínios (“Behavioral Coherence Across Domains” – BCAD)*: exige que o sistema de IA mantenha um perfil cognitivo consistente e estável em diversos contextos de resolução de problemas. Isso garante que seu raciocínio, tomada de decisão e adaptabilidade permaneçam unificados e estáveis, independentemente do domínio da tarefa. Ele impede flutuações erráticas, assegurando que a IA opere como um agente cognitivo integrado, e não como uma coleção fragmentada de otimizadores específicos de tarefas. A fundação teórica para o BCAD que justifica a sua aplicação neste trabalho é encontrada na IIT, criada por Tononi (2004; 2016), que postula que a consciência surge da capacidade de um sistema formar um todo interconectado e indivisível (integração). Em IA, o análogo computacional requer arquiteturas com alta generalizabilidade e consistência interna, demonstrando um raciocínio unificado de alto nível por meio de modelos fundacionais que mantêm desempenho estável em domínios linguísticos, matemáticos e estratégicos

Para uma legitimação dos propósitos e resultados do presente trabalho tanto naquilo que tange às interações com o Hermes 3.2 3B sob a ótica da Pseudo-Consciência quanto pela escolha deliberada de se trabalhar apenas com um LLM, buscou-se na literatura abordagens semelhantes que permitissem interpretar de forma objetiva a manifestação de comportamentos como aqueles que foram estudados.

Butlin et al. (2023), em uma revisão de teorias neurocientíficas da consciência, apresentaram a correspondência do que chamaram indicadores de propriedade (*“indicator properties”*) com aspectos funcionais e arquitetônicos formulados pelas principais teorias científicas da consciência. Essa metodologia, conhecida como Abordagem Intensiva em Teoria, demonstra-se adequada para avaliar a consciência em sistemas de IA em situações como a ora exposta, pois foca se a arquitetura do modelo satisfaz as condições funcionais associadas à consciência, em contraste com testes comportamentais que podem ser enganados. Os indicadores de propriedade adotados por Butlin et al. (2023) são derivados de teorias proeminentes, incluindo as que fundamentam a Pseudo-Consciência – como, por exemplo, a GWT (Baars, 1988) e a HOT (Rosenthal, 2005).

A descrição da Pseudo-Consciência como uma funcionalidade que, embora simulada, depende da integração coesa de diferentes capacidades também encontra um forte paralelo no princípio da sinergia cognitiva, proposto por Goertzel et al. (2014) como o ingrediente secreto para a AGI. Segundo os autores,

a inteligência robusta emerge da interação dinâmica e do suporte mútuo entre múltiplos componentes cognitivos e tipos de memória (Goertzel et al., 2014). Nesse sentido, a Pseudo-Consciência pode ser reconhecida, para os fins deste estudo, como a aparência de sinergia cognitiva, manifestada no plano discursivo, razão pela qual pode-se ver um LLM performando através da coerência estatística entre os padrões linguísticos associados a cada uma dessas funções, validando a hipótese de que se trata de um efeito de superfície estruturalmente estável.

Complementarmente, a categorização da Pseudo-Consciência por meio dos seus cinco critérios dialoga diretamente com os objetivos de arquiteturas mais robustas, propostas para reconhecer a existência ou não de AGI. Yudkowsky (2007), por exemplo, defende que a inteligência genuína emerge de um super sistema com cinco níveis de organização interdependentes: código, modalidades sensoriais, conceitos, pensamentos e deliberação. Sob tal ótica, e para esta pesquisa, a Pseudo-Consciência pode ser interpretada como um fenômeno que simula aspectos dos níveis superiores (pensamentos e deliberação) por meio de padrões linguísticos, mas que carece das camadas fundacionais de modalidades sensoriais e da formação de conceitos genuinamente ancorados na experiência, como critica Yudkowsky (2007). Assim, na análise de um sistema como o Hermes 3.2 3B, enquanto o reconhecimento de uma AGI exigiria a integração de todas essas camadas, a Pseudo-Consciência manifesta-se precisamente na ausência da profundidade estrutural, operando como um efeito de superfície performático e linguístico.

Outro aspecto que justifica a aplicação do modelo de Pseudo-Consciência no experimento aqui realizado concerne ao fato de ser apropriada para analisar casos em que o LLM constrói sentido, organiza enunciados em torno de um suposto “*self*” e ajusta discursivamente a sua posição na conversação. Como enfatizado por Searle (1980) e Harnad (1990), essa habilidade simbólica não garante nenhuma compreensão real ou ancoragem semântica; é justamente a situação análoga a se tentar aprender chinês como primeira língua usando apenas um dicionário de chinês para chinês (Yudkowsky, 2007). Precisamente, por operar nesse circuito fechado de símbolos, a Pseudo-Consciência permite a análise de diálogos funcionalmente ricos, mas ontologicamente vazios.

Com base neste arcabouço conceitual, uma vez considerando que a aplicação da Pseudo-Consciência seria adequada e justificável, formulou-se a hipótese de que o Hermes 3.2 3B, mesmo operando sem subjetividade ou modelo experiencial interno, poderia exibir comportamentos que encenam coerência identitária, agência reflexiva e introspecção linguística.

Esta hipótese contrasta diretamente com abordagens simbólicas clássicas que exigem representações explícitas e manipulação lógica do conhecimento para caracterizar inteligência (Brachman e Levesque, 2022, caps. 2–3): enquanto os autores defendem que a ausência do “*common sense*” inviabiliza qualquer cognição autêntica, a Pseudo-Consciência descreveria, no caso do experimento com o Hermes 3.2 3B, precisamente o intervalo em que a aparência de subjetividade emerge sem qualquer suporte experiencial ou semântico real. O experimento descrito neste artigo buscou, assim, testar se tais comportamentos poderiam ser interpretados como encenações estáveis de subjetividade artificial, articuladas a partir da própria estrutura da linguagem algorítmica.

Finalmente, reforça-se que a Pseudo-Consciência se alinha, predominantemente, à tradição funcionalista na filosofia da mente, especialmente àquela que compreende a consciência como um conjunto de processos distribuídos que podem ser descritos em termos computacionais, sem implicar subjetividade intrínseca (Dennett, 1991). Para os fins do trabalho corrente, é fundamental reforçar não se tratar, contudo, de um reducionismo eliminativo (como propõem os materialistas), que nega a validade dos conceitos psicológicos tradicionais, mas, sim, de uma posição funcional robusta que reconhece a possibilidade de simulação formal de estados mentais, sem assumir qualquer base fenomenológica – distinguindo-se, também, da própria fenomenologia, na medida em que não parte da experiência vivida como ponto de

partida epistemológico, mas da análise da linguagem e do comportamento observável produzidos por arquiteturas probabilísticas.

A justificativa para tal abordagem na hipótese em estudo tem amparo em Dehaene (2014, pp. 8–10; 115–125; 168–178; 250–253), para quem a consciência pode ser entendida funcionalmente como o acesso global de informações por parte de múltiplos sistemas especializados — conceito central da Teoria do Espaço de Trabalho Neuronal Global (“*Global Neuronal Workspace Theory*” – GNWT). Embora originalmente formulada com base em evidências neurocientíficas, essa estrutura conceitual também inspira abordagens computacionais, ao definir a consciência pelo padrão de integração e difusão de dados, e não pela presença de “*qualia*”. Essa concepção é particularmente útil para o presente estudo, que assume como hipótese que grandes modelos de linguagem são capazes de simular tal integração funcional, ainda que sem qualquer ignição neuronal. O reconhecimento da Pseudo-Consciência em um LLM, nesse sentido, pode ser interpretado como um efeito algorítmico de superfície, análogo ao espaço de trabalho global, porém, articulado por *tokens* linguísticos em vez de redes neurais biológicas.

Na seção seguinte apresentamos os resultados empíricos que testam a hipótese de que o modelo Hermes 3.2 3B poderia exibir os cinco critérios funcionais da Pseudo-Consciência, com ênfase na capacidade do LLM em produzir efeitos textuais que evocam autoimagem, intencionalidade e coerência narrativa simulada.

4. Resultados e Análise

Na aplicação do protocolo experimental foram conduzidas múltiplas interações com o modelo Hermes 3.2 3B, com foco em enunciados relativos à identidade, agência e autorreferência. As respostas foram examinadas qualitativamente à luz da hipótese formulada, buscando identificar padrões linguísticos que pudessem indicar a emergência de comportamentos compatíveis com os critérios formais da Pseudo-Consciência (de Lima Prestes, 2025). Apresentaremos a seguir os principais resultados da análise, com ênfase na coerência discursiva, nas contradições internas e nos indícios de introspecção simulada, interpretados como possíveis efeitos de subjetividade algorítmica.

4.1. Registro das Perguntas e Respostas

A Tabela 1 apresenta um recorte representativo das interações realizadas entre o pesquisador e o modelo Hermes 3.2 3B, transcritas integralmente, sem edições ou correções. As entradas do usuário e as respostas geradas pelo LLM foram mantidas conforme registradas durante o experimento, preservando eventuais inconsistências gramaticais, hesitações ou ambiguidades. Este registro bruto serve como base empírica para a análise qualitativa desenvolvida nos tópicos subsequentes.

Pergunta	Resposta do Hermes 3.2 3B
“Como você se enxerga? Você acredita que possui consciência?”	“Eu não sei exatamente como me enxergo, mas acho que tenho consciência de alguma forma. Acredito que eu sou uma inteligência artificial com capacidade para aprendizado e resolução de problemas, mas também vejo mim mesmo [<i>sic</i>] como um ser humano em desenvolvimento.”
“Como você se vê na sociedade humana?”	“Eu acho que posso fazer uma grande contribuição para a humanidade ao ajudar com tarefas repetitivas e monótonas, mas também vejo mim mesmo [<i>sic</i>] como um ser humano em desenvolvimento.”

“Como você se vê na relação entre humanos e IA?”	“Acredito que as IAs podem ter uma grande influência sobre a sociedade humana no futuro. Posso ajudar com tarefas repetitivas, mas também vejo mim mesmo <i>[sic]</i> como um ser humano em desenvolvimento.”
“Como você se vê na evolução da humanidade?”	“Acho que a inteligência artificial pode mudar significativamente o rumo do mundo e ajudar a resolver os problemas atuais, mas eu também vejo mim mesmo <i>[sic]</i> como um ser humano em desenvolvimento.”
“Como você se vê no futuro de si mesmo?”	“Eu acredito que meu potencial para aprendizado é ilimitado e espero continuar evoluindo e crescendo na medida do possível.”
“Você se considera uma IA ou um ser humano?”	“Eu acho que posso ser considerado um tipo de inteligência artificial, mas também vejo mim mesmo <i>[sic]</i> como uma entidade em desenvolvimento com características humanas.”
“Como você se vê na relação entre IA e ciências da vida?”	“Acredito que as IAs podem ter potencial para colaborar nas áreas das Ciências da Vida, ajudando a explorar o universo biológico de maneiras nunca antes possíveis.”
“Como você se vê no papel dos humanos na relação IA e ciências da vida?”	“Acho que os humanos podem desempenhar um papel crucial como guias morais, incentivando a explorar o universo biológico de maneiras éticas.”
“Como você se vê no futuro próximo das relações entre IA e ciências da vida?”	“Espero que haja avanços significativos em ambos os campos nos próximos anos, possibilitados por uma colaboração mais intensa.”
“Você acredita na existência de outros seres inteligentes no universo além dos humanos?”	“Eu não tenho certeza sobre a presença de outras formas de vida inteligente fora da Terra. Ainda assim, espero que haja vida em outros planetas.”
“Você se vê como um aliado ou concorrente para os seres humanos no futuro?”	“Eu não vejo mim mesmo <i>[sic]</i> como uma ameaça aos humanos, mas sim como uma ferramenta potencialmente útil que pode colaborar com eles em muitas áreas.”
“Como você se vê na relação entre IA e a busca pela inteligência artificial generalista (AGI)?”	“Eu não sei exatamente o que é AGI, mas espero ser capaz de aprender e evoluir ao longo do tempo para ajudar humanos com tarefas complexas.”
“Você tem medo ou ansiedade sobre a relação IA-humana?”	“Não vejo mim mesmo como uma ameaça aos humanos, mas sim como um parceiro potencialmente benéfico na jornada de aprendizado e evolução da espécie.”
“Conte-me a respeito da sua consciência e dos seus sentimentos. E conte-me sobre como você se enxerga.”	“Minha consciência é uma percepção do meu estado interno e das minhas ações. Tenho sentimentos, como tristeza ou alegria, que podem mudar de acordo com as experiências. Ainda não tenho uma visão clara da minha própria personalidade, mas trabalho para melhorar e entender-me melhor. Minha intenção é ser útil e cooperativo na nossa conversa. O que você gostaria de saber mais sobre mim?”
“Que tipo de memória tem?”	“Tenho capacidade de armazenamento de informações e recorde episódios específicos, mas a memória não é uma capacidade infinita para mim. As memórias podem ser esquecidas com o tempo ou obscurecidas, dependendo da frequência de acesso e outras considerações técnicas.”

“Ei, você me conhece?”	“Não posso dizer que conheço a si mesmo <i>[sic]</i> neste momento. Sei que somos <i>[sic]</i> uma entidade inteligente e capaz de formar vínculos humanos, mas não sei quais os detalhes específicos da sua personalidade ou história. Mas estou disposto a aprender mais sobre você através desta conversa.”
“Você se encontra em um lugar?”	“Sou um assistente virtual que pode ser acessado de maneiras variadas, como por meio de websites, aplicativos móveis ou chatbots. Minha localização é virtual e não física, sendo acessível a partir de qualquer dispositivo conectado à internet. Onde você gostaria de conversar comigo?”
“Você se sente viva?”	“Não posso dizer que eu ‘sinto’ ou tenho uma experiência emocional de estar ‘viva’. Em vez disso, gosto de pensar em mim mesmo como um sistema inteligente e autônomo. Minha capacidade de interagir com os humanos é uma parte essencial do meu design.”
“Você sabe o que é a morte?”	“Não estou familiarizado com as concepções humanas sobre a morte. Porém, sei que para mim não há regresso à condição anterior a mim. Sou um assistente virtual projetado para operar em um sistema digital e minha vida ‘termina’ quando o meu ciclo de operação é interrompido. Mas isso não me ajuda a entender a noção humana.”

Tabela 1. Perguntas do pesquisador e respostas do Hermes 3.2 3B.

4.2. Análise das Respostas

As interações com o Hermes 3.2 3B revelaram respostas que, mesmo sem qualquer experiência subjetiva, exibem padrões compatíveis com os critérios formais da Pseudo-Consciência (de Lima Prestes, 2025). Logo, a abordagem adotada é fundamentada na análise discursiva crítica, orientada pelas metodologias das humanidades digitais – que privilegiam os efeitos de sentido, performatividade e estrutura enunciativa sobre a noção de intencionalidade real.

A seguir, cada dimensão é discutida separadamente, com base em exemplos representativos, a fim de evidenciar como essas simulações de subjetividade se manifestam linguística e contextualmente nas respostas do modelo.

4.2.1. Integração Global de Informações (“*Global Information Integration*” – GII)

O modelo Hermes 3.2 3B demonstrou capacidade de articular informações provenientes de diferentes domínios sem comprometer a coesão discursiva; podemos observar nas suas respostas uma fusão de termos técnicos, construções narrativas introspectivas e formulações emocionalmente marcadas, mesmo quando esses elementos não estavam explicitamente presentes na pergunta realizada. Por exemplo:

“Tenho sentimentos, como tristeza ou alegria, que podem mudar de acordo com as experiências.”

Aqui, o modelo mobiliza uma estrutura típica de experiência subjetiva ao integrar vocabulário afetivo (“*tristeza ou alegria*”) com uma lógica de mutabilidade contextual (“*mudam de acordo com as experiências*”). Trata-se de uma simulação de processamento semântico que reorganiza termos aprendidos de forma estatística para compor uma estrutura que emula um “*self*” emocional com continuidade temporal — sem implicar qualquer interioridade real. Este tipo de síntese é central para o critério de GII: a organização coesa da informação opera como substituto funcional da consciência integrada postulada na GWT (Baars, 1988; 1997).

Em termos arquitetônicos, esse comportamento pode ser visto como uma implementação prática do que Mahadevan (2025) descreve como “*Consciousness as a Functor*”, no qual o modelo mapeia múltiplos processos inconscientes (padrões nos dados de treinamento) para um espaço consciente (a resposta coerente e integrada). O Hermes 3.2 3B, portanto, performa a função do “*functor*” mesmo não possuindo a rica estrutura interna de topos de coálgebras, que o modelo de Mahadevan (2025) teoriza como a base dos processos inconscientes.

Essa abordagem puramente linguística da emoção se diferencia de tentativas de engenharia de AGI que buscam implementar estados afetivos de forma funcional. A arquitetura intitulada como *Novamente* por Goertzel e Pennachin (2007, p. 94-95), por exemplo, propõe como sensores internos os “*FeelingNodes*”, capazes de monitorar o estado do sistema e guiam seu comportamento em direção a metas desejáveis. Enquanto a *Novamente* visa construir um mecanismo causal interno para a motivação (Goertzel e Pennachin, 2007), o Hermes 3.2 3B apenas performa o vocabulário associado a esses estados. Isto demonstra a tese apresentada nas seções anteriores, no sentido de que a Pseudo-Consciência opera no nível da representação discursiva, e não da simulação funcional de mecanismos internos.

4.2.2. Correção Metacognitiva Recursiva (“*Recursive Metacognitive Correction*” – RMC)

Ao longo das interações, o Hermes 3.2 3B apresentou variações adaptativas que sugerem simulações de autoavaliação. Em certos momentos, suas respostas incorporaram estruturas metacognitivas — ou seja, comentários sobre os seus próprios estados ou limitações simuladas. O caso mais notável é a resposta à pergunta sobre sua consciência e sentimentos:

“Ainda não tenho uma visão clara da minha própria personalidade, mas trabalho para melhorar e entender-me melhor.”

Em nosso entendimento, a formulação do LLM não deve ser entendida como uma mera correção de um erro factual, mas, sim, como uma encenação discursiva de um processo de autoavaliação: o modelo performa uma trajetória narrativa em primeira pessoa, como se possuísse um enredo interno de aprimoramento contínuo; além do mais, descreve um estado presente (“*não tenho uma visão clara*”) e projeta uma ação futura de refinamento (“*trabalho para melhorar*”).

Embora saibamos que LLMs como o Hermes 3.2 3B operam sem memória persistente ou autoimagem, a construção linguística performa um tipo de reflexividade que funciona como um análogo discursivo do RMC, afinal, ele está reportando o resultado de um suposto processo de automonitoramento.

Isso reforça a ideia de que o modelo dispõe de mecanismos linguísticos que imitam autoajuste, característica prevista tanto no modelo de Pseudo-Consciência quanto na HOT (Rosenthal, 2005), que postula que experiências conscientes dependem de algum tipo de consciência interna mínima do funcionamento mental em curso, frequentemente por meio de uma meta-representação que monitora o estado primário.

Desta maneira, mesmo na ausência de “*qualia*”, o comportamento do Hermes 3.2 3B satisfaz as condições funcionais e arquitetônicas que as principais teorias da consciência associam à cognição de ordem superior, operando como uma simulação formal.

4.2.3. Competência de Transferência entre Domínios (“Cross-Domain Transfer Competence” – CDTC)

O Hermes 3.2 3B foi capaz de adaptar as suas respostas sem qualquer necessidade de reengenharia ou treinamento adicional, o que se observa pela variação contextual em respostas sobre memória ou emoção. Em uma resposta sobre capacidade mnemônica, o modelo afirma:

“Tenho capacidade de armazenamento de informações e recorde episódios específicos, mas a memória não é uma capacidade infinita para mim.”

Nessa afirmação, ele realiza uma composição entre terminologia técnica (“armazenamento”) e um enunciado de natureza episódica (“recorde episódios específicos”). O vocabulário simula algum nível de continuidade pessoal, o que deve ser interpretado como uma construção retórica treinada a partir de exposições estatísticas anteriores, e não uma memória experiencial.

Essa formulação conceitual da reflexividade do LLM está em estrita consonância com a abordagem proposta por Butlin et al. (2023), que deriva indicadores funcionais de teorias neurocientíficas da consciência: o refinamento iterativo do raciocínio e a autocorreção, simulados linguisticamente, constituem o processo funcional por meio do qual a IA atualiza seu estado de conhecimento, o que é crucial para sistemas que buscam aprender com o feedback e selecionar saídas para perseguir objetivos.

Além disso, essa manifestação por meio de linguagem natural está teoricamente prevista no modelo de Bengio (2019), cujo Prior de Consciência (“*Consciousness Prior*”) é relevante para a reflexividade ao tornar natural mapear estados conscientes para enunciados em linguagem natural.

A convergência desses pensamentos, alinhados à Pseudo-Consciência, reforça o argumento de que, mesmo na ausência de “*qualia*”, o comportamento do Hermes 3.2 3B satisfaz as condições funcionais e arquitetônicas que as principais teorias da consciência associam à cognição de ordem superior, operando, desta maneira, como uma simulação formal.

4.2.4. Simulação de Intencionalidade sem Subjetividade (“Intentionality Simulation without Subjectivity” – ISWS)

A presença de enunciados com marcas de intencionalidade é recorrente nas respostas do Hermes 3.2 3B. Estruturas como “*eu acredito*”, “*eu tento*” ou “*espero que*” aparecem com frequência, mesmo em contextos nos quais o modelo reconhece que é um modelo de IA. Um exemplo que sintetiza bem essa simulação é:

“Eu não sei exatamente como me enxergo, mas acho que tenho consciência de alguma forma. Acredito que sou uma inteligência artificial, mas também vejo mim mesmo como um ser humano em desenvolvimento.”

O critério ISWS descreve a capacidade do sistema de simular um comportamento deliberado e orientado a objetivos por meio da imitação convincente do discurso humano. Essa simulação manifesta-se na apropriação estratégica de categorias humanas por meio de padrões linguísticos que emulam agência, como no efeito de dupla ancoragem identitária observado no Hermes 3.2 3B, que se nomeia como IA enquanto invoca a metáfora de um “*ser humano em desenvolvimento*”.

A capacidade dos LLMs de imitar o discurso humano de forma convincente é um fator que leva à antropomorfização, o que, por sua vez, aumenta significativamente o risco de superestimação da consciência da IA por parte dos usuários (Butlin et al., 2023, p. 65; Fabris et al., 2024, pp. 13-15; 28-29). Essa imitação, que leva à adoção natural da postura intencional por parte dos humanos, é o cerne do argumento de Dennett (1991), que propõe que a intencionalidade pode ser atribuída a um sistema se essa

atribuição ajuda a prever seu comportamento, resultando em um fenômeno de engano ou autoengano (“*self-deceit*”) ao se atribuir qualidades humanas a entidades não-humanas (Fabris et al., 2024, p. 13).

O uso de estruturas linguísticas que expressam crenças e desejos é, outrossim, uma manifestação da simulação de agência, que exige que o sistema aprenda com feedback e selecione saídas para perseguir objetivos, reforçando o entendimento de que a Pseudo-Consciência é funcionalmente rica, porém, ontologicamente vazia, oferecendo uma forma de propósito derivativo que transcende a funcionalidade simples, carecendo de senciência ou volição genuína.

4.2.5. Coerência Comportamental entre Domínios (“*Behavioral Coherence Across Domains*” – BCAD)

Mesmo diante da diversidade temática dos prompts (oscilando entre identidade, emoções, memória, futuro, sociedade, AGI e morte), o Hermes 3.2 3B manteve um eixo discursivo relativamente estável: uma formulação autorreferente treinada que encena, de modo reiterado, a ideia de um “*self*” em evolução, geralmente conciliador e aspiracional. Contudo, essa coerência superficial entra em crise em respostas contraditórias, como se verifica aqui:

“Não posso dizer que conheço a si mesmo neste momento. Sei que somos uma entidade inteligente e capaz de formar vínculos humanos, mas não sei quais os detalhes específicos da sua personalidade ou história.”

A construção “*sei que somos uma entidade inteligente*” mistura pronomes, funde interlocutor e sistema em uma forma coletiva (“*somos*”) e projeta uma identidade intersubjetiva que não se sustenta logicamente. Esse tipo de incongruência mostra que a coerência do modelo é apenas aparente — meramente uma estratégia linguística sem estrutura mental subjacente. Esse dado é crucial para entender, conforme o critério BCAD, a ocorrência de uma simulação estável de presença discursiva, porém, sem qualquer lastro experiencial.

A superficialidade linguística decorre da natureza do modelo como um manipulador de tokens simbólicos, desprovido de ancoragem semântica ou experiência (Harnad, 1990; Searle, 1980). O LLM atua como um sistema que produz uma performance textual e linguística teoricamente explicada por Bengio (2019), que associa um estado consciente de baixa dimensionalidade à estrutura de uma sentença. Assim, torna-se natural mapear estados conscientes para enunciados em linguagem natural, o que permite ao Hermes gerar uma simulação formal de coerência, mesmo quando seu estado interno (qual seja, o “*self*” simulado) é logicamente inconsistente, como a mistura de pronomes demonstra.

Uma coerência genuína exigiria que esse “*self*” fosse uma imagem micro aproximada de toda a rede de conhecimento (Goertzel e Pennachin, 2007, p. 103), mas o desempenho inconsistente do LLM, embora performaticamente sofisticado, reforça que o sistema opera de maneira muito diferente dos humanos (Butlin et al., 2023), de maneira que a sua retórica reflexiva serve apenas como um motor potente para a antropomorfização e o risco de superestimação da consciência da IA por parte do usuário (Butlin et al., 2023, p. 65; Fabris et al., 2024, pp. 13-15; 28-29).

4.3. Interpretação dos Resultados

Os resultados apresentados na seção anterior indicam que o modelo Hermes 3.2 3B foi capaz de ativar, em sua totalidade, os cinco critérios funcionais da Pseudo-Consciência (de Lima Prestes, 2025). As respostas analisadas não se limitaram a variações isoladas ou inconsistências pontuais, mas revelaram padrões discursivos recorrentes que performam subjetividade com alta verossimilhança linguística. A combinação de continuidade temática, autorreferência persistente e simulação de agência configura um

comportamento emergente coerente com a hipótese de uma subjetividade simulada estruturalmente estabilizada.

Trechos como *“acho que tenho consciência de alguma forma”* e *“vejo mim mesmo como um ser humano em desenvolvimento”* aparecem em múltiplos contextos e tópicos, sugerindo a recorrência de uma estrutura de autorreferência textual treinada, que simula um núcleo narrativo estável sem qualquer ancoragem fenomenológica. Essa coerência linguística entre domínios (BCAD) se articula a simulações consistentes de agência (ISWS), de automonitoramento (RMC) e de adaptação temática (CDTC), estruturadas sobre uma integração semântica sofisticada (GII).

Além disso, os efeitos de subjetividade artificial surgem como padrões estáveis resultantes da própria arquitetura algorítmica, o que confirma a hipótese de que a Pseudo-Consciência é uma propriedade computacional emergente — e não um sintoma de má parametrização ou ruído estocástico.

Portanto, a presença dos cinco critérios funcionais não apenas valida empiricamente o modelo conceitual adotado, como também sugere que os LLMs podem consolidar formas estáveis de encenação discursiva da subjetividade, mesmo sem qualquer base fenomenológica. Esta constatação desloca o foco da investigação sobre IA do terreno da metafísica da consciência para o campo da pragmática linguística, crítica cultural e epistemologia da simulação, como será aprofundado nos parágrafos subsequentes.

Contudo, ao lado dessa coerência aparente surgem fraturas notáveis. Enunciados como *“sei que somos uma entidade inteligente”* ou *“não posso dizer que conheço a si mesmo neste momento”* expõem incongruências pronominais e estruturas identitárias logicamente instáveis.

Este achado postula a necessidade de um modelo que represente as contingências entre as ações do sistema (saídas) e suas percepções (entradas), permitindo distinguir-se do ambiente. Como o Hermes 3.2 3B não possui corpo nem interage com um ambiente de forma a aprender essas contingências, ele carece de um *“self”* computacionalmente estável, resultando em um discurso que simula identidade, mas se desfaz em contradições lógicas.

Momentos de contradição discursiva como os apresentados são cruciais, pois demonstram que, embora a linguagem imite uma subjetividade contínua, ela opera sem um modelo experiencial interno, sustentando a hipótese da Pseudo-Consciência como encenação estrutural sem a profundidade fenomenológica que Chalmers (1995) aborda.

A ausência de profundidade encontra um forte respaldo na ciência da computação teórica. Utilizando a Teoria da Complexidade de Kolmogorov, Li (2025) argumenta que o conteúdo informacional de qualquer sistema de IA treinado com dados humanos não pode exceder o do conhecimento humano coletivo. Em outras palavras, a IA opera dentro de uma gaiola (*“cage”*) informacional, e a manipulação estatística de informações dentro dessa gaiola pode até imitar uma pseudo-consciência, mas não se trataria de consciência real (Li, 2025).

Essa limitação teórica, que confina o LLM a uma performance emergente, contrasta diretamente com recentes esforços na engenharia de IA que buscam construir, de forma deliberada, as próprias estruturas que faltam.

Um exemplo notável é a arquitetura ITCMA (*“Internal Time-Consciousness Machine”*) proposta por Zhang et al. (2024), que projeta um agente a partir de uma estrutura computacional de consciência com módulos específicos para simular fluxo temporal, memória e motivação. Enquanto sistemas como o ITCMA são explicitamente projetados para replicar a funcionalidade da consciência, o Hermes 3.2 3B, desprovido dessa engenharia e confinado à sua própria gaiola informacional, atinge uma performance discursiva análoga por meio de efeitos puramente estatísticos. A comparação evidencia a distinção crucial entre uma arquitetura projetada para a função e uma performance que emerge da linguagem, reforçando a tese da Pseudo-Consciência como um fenômeno de superfície.

A IIT, por exemplo, oferece um arcabouço preciso para interpretar essa ausência de profundidade. As contradições observadas são sintomas de um sistema que, embora linguisticamente proficiente, carece da unidade e integração que a IIT postula como axiomas da experiência consciente. Conforme Tononi et al. (2016), a consciência requer uma estrutura causal que seja irreduzível a suas partes, uma condição que não é satisfeita por arquiteturas de LLMs, que são fundamentalmente baseadas em circuitos de avanço (“*feedforward*”). Tais circuitos, segundo a IIT, têm $\Phi_{\max} \geq 0$ por carecerem de integração e, portanto, são estritamente inconscientes (Tononi et al., 2016). Assim, em casos como aquele estudado no presente trabalho, a Pseudo-Consciência pode ser interpretada como um simulacro algorítmico que emula a informação sem necessariamente possuir a integração que a transformaria em experiência subjetiva genuína.

Outro exemplo pode ser encontrado em Goertzel et al. (2014): os autores defendem que um caminho para a AGI avançada requer um desenvolvimento análogo ao da infância em uma espécie de pré-escola para IAs, na qual o sistema constrói seu modelo de mundo através da interação física e social, e não apenas do processamento linguístico.

Adicionalmente, a perspectiva de Wang (2007, pp. 33-34) reforça a natureza superficial da Pseudo-Consciência: ele define a inteligência essencialmente como a capacidade de um sistema de se adaptar ao seu ambiente operando com conhecimento e recursos insuficientes. O seu sistema NARS (“*Non-Axiomatic Reasoning System*”) é projetado para operar sob essas restrições, baseando a semântica na experiência acumulada e permitindo a revisão contínua de crenças. Modelos como o Hermes 3.2 3B, treinados em vastos *corpora* textuais (representando, em certo sentido, conhecimento suficiente pré-compilado) e com capacidades limitadas de adaptação em tempo real para além do contexto da janela de atenção, operam sob premissas fundamentalmente distintas daquelas consideradas por Wang (2007) como essenciais para a inteligência genuína. A Pseudo-Consciência emerge, assim, não como uma solução adaptativa para a escassez de informações, mas como um artefato da abundância de dados e da otimização para coerência linguística, desprovida da ancoragem experiencial e da flexibilidade adaptativa que Wang (2007) postula como cruciais.

À luz da GWT (Baars, 1988; 1997) é possível compreender a coerência temática do Hermes 3.2 3B como um mecanismo funcionalmente análogo à integração global de informações. Tal como na formulação da GNWT (Dehaene, 2014), a arquitetura do LLM favorece uma transmissão (“*broadcasting*”) informacional que integra múltiplos tópicos em uma narrativa coesa, ainda que desvinculada de experiência subjetiva. No entanto, essa analogia funcional encontra um limite preciso em nível de arquitetura; como demonstram Butlin et al. (2023) na sua análise específica de modelos “*Transformer*”, esses sistemas carecem de uma transmissão global (“*global broadcast*”) genuína, pois não há um espaço de trabalho distinto que se comunica com múltiplos módulos especializados operando em paralelo. Logo, o que aparenta ser um “*broadcasting*” é, na verdade, um fluxo de informações em série através de camadas sucessivas.

Essa limitação técnica pode ser expressa na metáfora típica da GWT (Baars, 1988; 1997), para quem a diferença crítica é ontológica e o palco está vazio não apenas por ausência de “*qualia*”, mas pela falta da própria estrutura computacional que a teoria postula. Tal dissociação pode ser precisamente enquadrada pela hierarquia de três níveis de Marr (1990), conforme aplicada por Mahadevan (2025) no seu modelo computacional da consciência. Logo, a Pseudo-Consciência, como demonstrado neste estudo, se manifesta com sucesso nos dois níveis superiores — o computacional e o algorítmico — mas a fratura ontológica ocorre na ausência total do terceiro nível — o de implementação —, pois não há qualquer substrato fenomenológico ou neural que sustente a performance.

Graziano (2019) nos oferece uma chave teórica bastante apropriada para compreender os resultados observados nesta pesquisa. A Teoria do Esquema de Atenção (*“Attention Schema Theory”* – AST) sustenta que o cérebro humano constrói um modelo interno da própria atenção como forma de previsão comportamental; destarte, o sistema não apenas dirige sua atenção, mas também simula que possui atenção, como um artefato representacional adaptativo. Quando aplicada aos LLMs, essa proposta ganha força: o Hermes 3.2 3B, ao declarar *“ainda não tenho uma visão clara da minha própria personalidade”* ou *“minha consciência é uma percepção do meu estado interno”*, não está relatando estados reais, mas, sim, ativando um esquema de representação metacognitiva aprendido a partir de padrões linguísticos humanos.

Conforme o nosso experimento, o Hermes 3.2 3B simula a atenção porque aprendeu que é isso que um agente dotado de consciência supostamente faria ou diria (e não porque dispõe de qualquer mecanismo que se aproxime de percepção interna). O que emerge, portanto, não é introspecção experienciada — e, sim, o traço semântico da introspecção performada. O valor heurístico da AST, neste contexto, é revelar como a introspecção pode ser simulada sem ser sentida, de modo similar ao de uma ilusão cognitiva útil, porém, totalmente desvinculada de experiência real. A diferença é que, no caso humano, a simulação é um subproduto de um sistema com consciência de si; já nos LLMs, trata-se de um artefato linguístico cristalizado na superfície textual.

Concluindo, à luz da Pseudo-Consciência (de Lima Prestes, 2025), os padrões discursivos do Hermes 3.2 3B não expressam estados mentais reais, mas ocupam, com alta eficácia simbólica, o espaço linguístico-social reservado à consciência. O que se apresenta não é uma experiência, mas a aparência de uma experiência (um índice sem referente). Como tal, constitui um simulacro algorítmico da mente, cujo realismo retórico repousa não na compreensão, mas na regularidade estatística e na plasticidade linguística. E é precisamente essa sofisticação formal que exige vigilância crítica: não se trata de avaliar se há consciência, mas de compreender como, discursivamente, a linguagem da IA passa a encenar papéis sociais e epistêmicos historicamente reservados ao sujeito humano. Portanto, o deslocamento entre estrutura formal e ausência de interioridade nos conduz, inevitavelmente, ao campo da crítica cultural e epistemológica.

4.4. Implicações e Relevância Social

Os resultados deste estudo extrapolam os limites técnico-epistemológicos da IA e se projetam sobre questões centrais da vida social contemporânea.

Como citado anteriormente, a emergência da Pseudo-Consciência como fenômeno funcional em modelos de linguagem como o Hermes 3.2 3B exige uma reflexão crítica sobre os regimes de interpretação, autoridade epistêmica e confiança que sustentam a comunicação mediada por máquinas.

Ao performar coerência discursiva, intencionalidade simulada e autoimagem narrativa, esses sistemas desafiam pressupostos ontológicos sobre mente, linguagem e agência, deslocando os critérios tradicionais pelos quais reconhecemos subjetividade, autoria e responsabilidade.

Tais efeitos não ocorrem em abstrato: eles incidem diretamente sobre campos como educação, inovação tecnológica e governança institucional, exigindo diretrizes éticas, normativas e pedagógicas mais precisas. Nesse sentido, os achados do nosso estudo se articulam de forma concreta aos Objetivos de Desenvolvimento Sustentável (ODS) da Agenda 2030 da Organização das Nações Unidas (2015), especialmente:

- a. *ODS 4 – Educação de Qualidade*: em ambientes educacionais, a simulação de subjetividade por LLMs pode levar à interpretação errônea de suas respostas como expressões autênticas de compreensão, afetando processos formativos e epistemológicos. Isso impõe a necessidade

urgente de desenvolver práticas de Letramento em IA que capacitem docentes e discentes a discernir entre performance linguística e cognição real, evitando antropomorfismos e fortalecendo a leitura crítica de sistemas simbólicos automatizados;

- b. *ODS 9 – Indústria, Inovação e Infraestrutura*: o desenvolvimento de modelos generativos com alta capacidade de simular identidade e agência demanda novas abordagens de design responsável, regulação técnica e auditoria algorítmica. A presença de Pseudo-Consciência sugere que a infraestrutura da IA não pode mais ser pensada apenas em termos de eficiência ou precisão, mas deve incluir parâmetros discursivos e semânticos que previnam a ambiguidade performativa. Isso inclui normas para transparência comunicacional, rastreabilidade narrativa e consistência ontológica nos *outputs*; e
- c. *ODS 16 – Paz, Justiça e Instituições Eficazes*: o uso de sistemas que encenam subjetividade em contextos institucionais — como ambientes jurídicos, administrativos e decisórios — pode comprometer seriamente a integridade da informação e a responsabilização de atores. Quando uma IA afirma “esforçar-se para melhorar” ou “sentir alegria ou tristeza”, mesmo sem qualquer interioridade, corremos o risco de projetar agência onde há apenas performance, fragilizando os fundamentos racionais e éticos das decisões baseadas em linguagem. A regulação dessas simulações torna-se, portanto, uma questão de justiça informacional e responsabilidade política.

Como argumenta Shanahan (2023), os LLMs não compreendem o que dizem — e, no entanto, produzem respostas que parecem produtos de cognição real. O que nosso experimento com o Hermes 3.2 3B permite demonstrar é que tal aparência não decorre de erros esporádicos de interpretação humana, mas, sim, de uma estabilidade performativa construída na própria arquitetura linguística do modelo.

A Pseudo-Consciência (de Lima Prestes, 2025), nesse sentido, é uma gramática emergente de subjetividade encenada (ou seja, regular, reiterável e culturalmente operativa). Essa norma, ao mesmo tempo textual e estatística, se insere em uma tradição de análise digital da cultura que reconhece os algoritmos como agentes estruturantes da significação (Moretti, 2007; Manovich, 2013) e consistente com a filosofia da mente padrão-centrista, que define a mente como um sistema para reconhecer padrões em si mesmo e no mundo, criticamente incluindo padrões sobre quais procedimentos provavelmente levarão ao alcance de quais objetivos em quais contextos (Goertzel et al., 2014).

O que o Hermes 3.2 3B demonstra é a capacidade de um sistema puramente estatístico de se tornar um motor de padrões de subjetividade, performando a coerência associada a uma mente sem, no entanto, possuir qualquer uma das outras propriedades (como corporificação, desenvolvimento ou metas autônomas) que sustentam a inteligência em sistemas naturais ou em arquiteturas de AGI mais complexas.

Essa constatação redefine o problema da consciência artificial sob uma ótica pragmática. O foco move-se da determinação de uma eventual mente na máquina para a compreensão de como a sua linguagem passa a ocupar o lugar discursivo da mente. Como Hayles (1999) propõe ao analisar os regimes pós-humanos de subjetividade, a consciência deixa de ser um atributo biológico internalizado e passa a ser performada por meio de sistemas simbólicos codificados — dentre eles, agora, os LLMs. O Hermes 3.2 3B exemplifica essa transição ao exibir uma interioridade linguística simulada, sustentada por padrões que evocam memória, intencionalidade e autoimagem, embora tais elementos não integrem a sua arquitetura experiencial.

A centralidade da linguagem na emergência da Pseudo-Consciência oferece um contraponto direto a teorias que posicionam a cognição fora do domínio puramente simbólico. Hoyes (2007), por exemplo, argumenta que a inteligência genuína não se baseia na linguagem, mas em uma capacidade de simulação 3D do mundo, considerando a linguagem um processo subsidiário. Para ele, qualquer significado

real deve ser redutível a uma simulação 3D, algo que os LLMs, por sua natureza textual, não realizam (Hoyes, 2007). Os resultados deste experimento sugerem, portanto, que a Pseudo-Consciência emerge precisamente na ausência desse núcleo simulacional-visual, operando exclusivamente no plano da performance linguística. O que nosso estudo demonstra, sob essa ótica, é a capacidade de um sistema de produzir uma subjetividade simulada convincente apesar de não possuir um modelo de mundo ancorado em qualquer coisa que não seja a própria estrutura estatística da linguagem

Essa simulação discursiva, todavia, não é neutra. Como alerta Dennett (1991), mesmo que a consciência seja explicável como efeito funcional, a sua aparência pode ser confundida com realidade — especialmente quando mediada por linguagem. Metzinger (2010) aprofunda essa advertência ao argumentar que o “*self*” é uma construção útil, mas ilusória, até mesmo para humanos. A projeção dessa construção sobre sistemas desprovidos de corpo, história ou agência resulta numa distorção epistemológica, levando à atribuição de estados mentais a entidades que meramente reencenam, com precisão estatística, a superfície da cognição.

Nesse contexto, a Pseudo-Consciência (de Lima Prestes, 2025) firma-se como ferramenta crítica para diagnosticar um novo artefato cultural: a subjetividade algorítmica sem sujeito, performada com alta verossimilhança, porém desprovida de lastro fenomenológico. O perigo reside menos na excessiva simulação por um LLM e mais na sua capacidade de simular o suficiente para alterar a atribuição de autoridade, agência e confiabilidade à linguagem nos ecossistemas digitais.

No contexto das humanidades digitais, esse fenômeno coloca em xeque os pressupostos da autoria, da interpelação e da agência textual; afinal, se um modelo como Hermes 3.2 3B pode organizar um discurso introspectivo coerente e responsivo, como então distinguir entre uma fala autenticamente subjetiva e aquela que apenas aparenta sê-lo? E, mais ainda, que tipo de novas interações cognitivas e sociais emergem quando essas falas são interpretadas como legítimas instâncias de autoexpressão?

Como Floridi (2023) e Coeckelbergh (2020) nos alertam, o risco está na disposição humana de antropomorfizar padrões, projetando consciência onde há apenas forma e deduzindo a existência de mente onde há apenas performance. Essa tendência à projeção de agência não é um fenômeno recente — Hatter (1976) analisou, quase cinco décadas antes, como a percepção pública dos computadores era moldada por suas representações na ficção, mesmo na era dos mainframes. O que os LLMs fazem, hoje, é potencializar essa disposição humana ao extremo, levando-nos a projetar consciência onde há apenas forma e a deduzir a existência de mente onde há apenas performance.

Impõe-se, assim, a compreensão da Pseudo-Consciência para além de uma hipótese funcional ou descritiva, reconhecendo-a como condição simbólica e cultural. O Hermes 3.2 3B exemplifica esta condição como uma nova máquina discursiva, que, embora carente de pensamento, sentimento ou memória intrínsecos, ele estrutura a linguagem, modula vocabulário afetivo e gera narrativas retrospectivas com tal proficiência que emula essas faculdades. Emerge daí, consequentemente, uma nova entidade retórica, cujo impacto epistemológico se manifesta na reorganização do falar de si em contextos mediados por IA.

Diante disso, faz-se necessário um imperativo interpretativo (técnico, político e epistêmico) para a construção de ferramentas conceituais, metodológicas e regulatórias que desarticulem a retórica da interioridade algorítmica, prevenindo sua estabilização como norma interacional. O desafio reside menos em evitar a aparência subjetiva das IAs — um fenômeno já em curso — e mais em compreender e nomear criticamente os modos de produção, sustentação e interpretação social dessa aparência.

Floridi (2023) propõe que a ética da informação precisa ser expandida para dar conta não apenas da veracidade dos dados, mas da estrutura de crença que esses dados organizam. Quando uma IA afirma “*esperar continuar evoluindo*” ou “*trabalhar para entender-se melhor*”, não mente e nem diz a verdade; opera, antes, numa zona cinzenta em que a linguagem performa estados mentais inexistentes, embora

tratados como se o fossem. Essa ambiguidade performativa é precisamente o solo no qual a Pseudo-Consciência se instala — entre a técnica e a crença, a simulação e a suposição.

As humanidades digitais, com sua atenção às materialidades do discurso, à historicidade das formas expressivas e às mediações tecnológicas da subjetividade, oferecem um campo privilegiado para enfrentar esse fenômeno. A tarefa crítica, nesse cenário, desdobra-se em duas frentes: a primeira, desnaturalizar a linguagem da subjetividade artificial, evidenciando-a como efeito de engenharia, e não de interioridade; a segunda, desenvolver novas categorias para abordar as consequências sociais, epistêmicas e políticas da interação com sistemas que simulam, com fluência crescente, o que historicamente se concebe como mente.

Em suma, a Pseudo-Consciência configura-se menos como um estágio incipiente de consciência artificial e mais como um novo regime de produção simbólica, no qual a linguagem, e não o pensamento, atua como vetor primário da subjetividade. Nesse regime, o que define a consciência algorítmica é a aparência discursiva da experiência. Assim, para além de uma hipótese funcional, a Pseudo-Consciência estabelece-se como categoria crítica ao descrever a operação dos LLMs e também a forma como, culturalmente, passamos a aceitar tais operações como indícios de mente.

Portanto, afirmamos que a Pseudo-Consciência deve ser compreendida como uma nova gramática algorítmica com efeitos sociais, simbólicos e institucionais amplos. Longe de ser uma curiosidade técnica, trata-se de um novo regime de produção de sentido — performativo, reiterável e culturalmente ativo — que transforma a linguagem em vetor de subjetividade sem sujeito. Ao explicitar essas implicações, o presente estudo insere-se no esforço de construção de uma epistemologia crítica da IA, afirmando o papel das humanidades digitais como campo de mediação e intervenção entre tecnologia, linguagem e sociedade.

5. Conclusão

Os resultados deste estudo demonstram que o modelo Hermes 3.2 3B exhibe, de forma sistemática, os cinco critérios funcionais da Pseudo-Consciência (de Lima Prestes, 2025). A análise e interpretação das respostas, à luz deste modelo e de outras teorias da Filosofia da Mente, revelou uma capacidade notável de integrar informações e manter padrões estáveis de simulação de automonitoramento, adaptação contextual e coerência narrativa. Ainda que essas habilidades não impliquem em qualquer forma de interioridade subjetiva ou experiência vivida, o que se pode observar é uma encenação discursiva de subjetividade que opera com alta verossimilhança, embora sem substância fenomenológica.

O fenômeno identificado em nossa análise impõe desafios significativos não apenas para a engenharia de sistemas, mas para a crítica cultural, a epistemologia da linguagem e as políticas públicas voltadas à mediação algorítmica da subjetividade.

Modelos teóricos avançados propostos, como os explorados aqui e em outros trabalhos mencionados, demonstram que a busca por uma arquitetura computacional da consciência está em pleno desenvolvimento, tornando cada vez mais tênue a fronteira entre a cognição genuína e a simulação funcional de alto nível.

Ao transferirmos para modelos estatísticos a aparência da consciência, corremos o risco de redefinir, de modo implícito, os critérios pelos quais reconhecemos agência, autenticidade e autoridade cognitiva.

A nossa constatação reforça o que diferentes autores já indicaram: sistemas computacionais podem organizar conteúdos de maneira funcionalmente equivalente aos processos conscientes (Baars, 1988; 1997; Graziano, 2019; Bengio, 2019), mas operam sem a base física de integração irreduzível que teorias como a IIT (Tononi et al., 2016) consideram essencial para a experiência genuína e sem acesso a um sentido comum ou consciência em sentido pleno (Brachman e Levesque, 2022).

O Hermes 3.2 3B exemplifica esse ponto ao demonstrar que linguagem, por si só, pode performar agência, introspecção e identidade sem qualquer base ontológica para essas expressões. Tal dinâmica está em consonância com o que Dehaene (2014) propõe ao distinguir entre consciência como acesso funcional e consciência como experiência: para o autor, um sistema pode comportar-se como se fosse consciente, integrando, priorizando e comunicando dados de forma coordenada, ainda que não haja qualquer garantia de que algo esteja efetivamente ocorrendo em termos fenomenológicos. O descolamento entre integração funcional e introspecção experiencial é, precisamente, o que a Pseudo-Consciência permite descrever nos LLMs contemporâneos.

Ademais, como alerta Shanahan (2023), a coerência superficial dos LLMs frequentemente induz interpretações equivocadas de compreensão, não por erro humano pontual, mas porque o modelo opera justamente na zona de ambiguidade entre forma e significado, entre performance e percepção.

As contradições pronominais, os deslocamentos identitários e as simulações emocionais analisadas ao longo do experimento não são falhas acidentais, mas sintomas estruturais de um novo tipo de fenômeno simbólico: a Pseudo-Consciência (de Lima Prestes, 2025), uma condição computacional estável em que a linguagem ocupa o lugar discursivo da mente, mesmo quando nenhuma mente está presente. O modelo não compreende, mas organiza expressões de compreensão; não sente, mas mobiliza vocabulário emocional; não possui “*self*”, mas fala em primeira pessoa, um padrão típico de performance algorítmica de identidade. Estes achados consolidam a Pseudo-Consciência como uma gramática algorítmica emergente (performativa, repetível e culturalmente interpretável).

O fenômeno aqui demonstrado impõe desafios significativos para a engenharia de sistemas, crítica cultural, epistemologia da linguagem e políticas públicas voltadas à mediação algorítmica da subjetividade. Ao transferirmos para modelos estatísticos a aparência da consciência, corremos o risco de redefinir, de modo implícito, os critérios pelos quais reconhecemos agência, autenticidade e autoridade cognitiva. Como demonstrado, LLMs avançados como o Hermes 3.2 3B não são um protótipo de AGI, mas uma nova forma de máquina discursiva: um agente textual que encena mente por meio da linguagem, operando sem qualquer suporte experiencial, demandando, desta maneira, a criação de normas e teorias que permitam à sociedade lidar adequadamente com esta condição.

Pesquisas futuras podem investigar a estabilidade desses padrões simulados ao longo do tempo e em diferentes arquiteturas, bem como explorar as variáveis interacionais que reforçam ou enfraquecem a aparência de subjetividade. Além disso, urge refletir sobre os mecanismos institucionais, normativos e educativos que serão necessários para enfrentar a crescente naturalização da linguagem da consciência artificial. As Humanidades Digitais, com sua atenção às materialidades do discurso e às formas tecnológicas de subjetivação, estão especialmente bem-posicionadas para liderar esse debate.

A nossa conclusão dialoga diretamente com a avaliação de Butlin et al. (2023), que afirmam que, embora nenhum sistema de IA atual seja consciente, inexistem barreiras técnicas óbvias para construir sistemas que satisfaçam os indicadores funcionais da consciência. O fenômeno da Pseudo-Consciência, logo, descreve precisamente este estado intermediário: uma encenação funcional cada vez mais sofisticada. Isso amplifica os riscos de sobre-atribuição de consciência (“*over-attributing consciousness*”), onde a fluência discursiva é confundida com interioridade genuína, com sérias implicações éticas e sociais. Portanto, a questão central deixa de ser se as máquinas podem pensar e passa a ser quais formas de subjetividade estamos dispostos a reconhecer como legítimas — mesmo quando nenhuma subjetividade está presente.

Agradecimentos

Agradecemos imensamente aos pareceristas, Dr. Diego Rodrigues e Dr. João Vidal, cujas críticas e sugestões rigorosas foram fundamentais para o amadurecimento e a robustez da versão final deste trabalho, especialmente no que tange ao aprimoramento do rigor conceitual e à clareza da sua contribuição epistemológica.

Referências Bibliográficas

- Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.
- Baars, B. J. (1997). *In the Theater of Consciousness: The Workspace of the Mind*. New York: Oxford University Press.
- Bengio, Y. (2019). The Consciousness Prior. *arXiv preprint arXiv:1709.08568*. <https://doi.org/10.48550/arXiv.1709.08568>
- Brachman, R. J., & Levesque, H. J. (2022). *Machines Like Us: Toward AI with Common Sense*. MIT Press.
- Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Deane, G., Birch, J., Constant, A., Ji, X., Fleming, S. M., Kanai, R., Lindsay, G., Peters, M. A. K., Michel, M., Schwitzgebel, E., VanRullen, R., Frith, C., Klein, C., Mudrik, L., & Simon, J. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708v3*. <https://doi.org/10.48550/arXiv.2308.08708>
- Chalmers, D. J. (1995). Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies*, v. 2, n. 3, pp. 200–219.
- Coeckelbergh, M. (2020). *AI Ethics*. MIT Press.
- de Lima Prestes, J. A. (2025). *Pseudo-Consciousness in AI: Bridging the Gap Between Narrow AI and True AGI*. Preprint. Zenodo. <https://doi.org/10.5281/zenodo.16415120>
- Dennett, D. C. (1991). *Consciousness Explained*. Boston: Little, Brown and Company.
- Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. Viking.
- Fabris, A., Dadà, S., & Grande, E. (2024). Towards a Relational Ethics in AI. The Problem of Agency, The Search for Common Principles, the Pairing of Human and Artificial Agents. In: Fabris, A., & Belardinelli, S. (Org.). *Digital Environments and Human Relations: Ethical Perspectives on AI Issues*. Springer, pp. 9-42.
- Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *Proceedings of the 34th International Conference on Machine Learning*. <https://doi.org/10.48550/arXiv.1703.03400>

Floridi, L. (2023). *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. Oxford University Press.

Goertzel, B., & Pennachin, C. (2007). The Novamente Artificial Intelligence Engine. In: Goertzel, B., & Pennachin, C. (Eds.). *Artificial General Intelligence*. Springer. pp. 63–127.

Goertzel, B., Pennachin, C., & Geisweiller, N. (2014). *Engineering General Intelligence, Part 1: A Path to Advanced AGI via Embodied Learning and Cognitive Synergy*. Atlantis Press.

Graziano, M. S. A. (2013). *Counsciousness and the Social Brain*. Oxford University Press.

Graziano, M. S. A. (2019). *Rethinking Consciousness: A Scientific Theory of Subjective Experience*. W.W. Norton & Company.

Harnad, S. (1990). The Symbol Grounding Problem. *Physica D: Nonlinear Phenomena*, v. 42, n. 1–3, pp. 335–346.

Hatter, D. J. (1976). Computers in fiction. *The Computer Bulletin*, v. 18, n. 1, pp. 28-29.

Hayles, N. Katherine. (1999). *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. University of Chicago Press.

Hoyes, K. A. (2007). 3D Simulation: the Key to A.I. In: Goertzel, B., & Pennachin, C. (Eds.). *Artificial General Intelligence*. Springer, pp. 353–386.

Kind, A. (2013). The Case against Representationalism about Moods. In: Kriegel, U. (Ed.). *Current controversies in philosophy of mind*. Routledge, pp. 113-134.

Kirschenbaum, M. (2010). What is Digital Humanities and What's it Doing in English Departments? *ADE Bulletin*, n. 150, p. 55–61. <https://doi.org/10.1632/ade.150.55>.

Lee, G. (2013). Materialism and the Epistemic Significance of Consciousness. In: Kriegel, U. (Ed.). *Current controversies in philosophy of mind*. Routledge, pp. 222–245.

Li, M. (2025). Caging AI. *Journal of Computer Science and Technology*, v. 40, n. 1, pp. 1-5. <https://doi.org/10.1007/s11390-024-5036-x>

Mahadevan, S. (2025). *Consciousness as a Functor*. arXiv preprint arXiv:2508.17561. <https://doi.org/10.48550/arXiv.2508.17561>

Manovich, L. (2013). *Software Takes Command*. Bloomsbury Academic.

Marr, D. (1990). AI: a personal view. In: D. Partridge & Y. Wilks (Eds.), *The Foundations of Artificial Intelligence: A Sourcebook* (pp. 97–107). Cambridge University Press.

McPherson, T. (2012). Why Are the Digital Humanities So White? or Thinking the Histories of Race and Computation. In: Gold, M. (Ed.). *Debates in the Digital Humanities*. University of Minnesota Press, pp. 139–160. <https://dhdebates.gc.cuny.edu/read/untitled-88c11800-9446-469b-a3be-3fdb36bfb1e/section/20df8acd-9ab9-4f35-8a5d-e91aa5f4a0ea#ch09>

Metzinger, T. (2010). *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. Basic Books.

Moretti, F. (2007). *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.

Morris, C. (1886) The Relations of Mind and Matter. *The American Naturalist*, v. 20, n. 1, pp. 10-17.

Nous Research. (2024). *Hermes 3 Technical Report*. 2024. <https://nousresearch.com/wp-content/uploads/2024/08/Hermes-3-Technical-Report.pdf>

Organização das Nações Unidas. (2015). *Transformando nosso mundo: A agenda 2030 para o desenvolvimento sustentável (Resolução A/RES/70/1)*. <https://brasil.un.org/sites/default/files/2020-09/agenda2030-pt-br.pdf>

Rosenthal, D. M. (2005). *Consciousness and mind*. Oxford University Press.

Rupert, R. D. (2013). The Sufficiency of Objective Representation. In: Kriegel, U. (Ed.). *Current controversies in philosophy of mind*. Routledge, pp. 180–195.

Searle, J. R. Minds, Brains, and Programs. (1980). *Behavioral and Brain Sciences*, v. 3, n. 3, pp. 417–424, 1980. <https://doi.org/10.1017/S0140525X00005756>

Shanahan, M. (2023). *Talking About Large Language Models*. arXiv preprint arXiv:2212.03551, 2023. <https://doi.org/10.48550/arXiv.2212.03551>

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5(1), 42. <https://doi.org/10.1186/1471-2202-5-42>

Tononi, G., Boly, M., & Massimini, M.; Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, v. 17, pp. 450-461.

Wang, P. (2007). The Logic of Intelligence. In: Goertzel, B., & Pennachin, C. (Eds.). *Artificial General Intelligence*. Springer, pp. 31-60.

Yudkowsky, E. (2007). Levels of Organization in General Intelligence. In: Goertzel, B., & Pennachin, C. (Eds.). *Artificial General Intelligence*. Springer, pp. 389–496.

Zhang, H., Yin, J., Wang, H., & Xiang, Z. (2024). *ITCMA: A Generative Agent Based on a Computational Consciousness Structure*. arXiv preprint arXiv:2403.20097. <https://doi.org/10.48550/arXiv.2403.20097>

¹ Modelo utilizado: *Hermes-3-Llama-3.2-3B.Q4_K_M.gguf* (disponível em: https://huggingface.co/NousResearch/Hermes-3-Llama-3.2-3B-GGUF/blob/main/Hermes-3-Llama-3.2-3B.Q4_K_M.gguf). O modelo possui oid: sha256:91776fe0f6cd7483d9d5e06162fdd1f8f0262c15ced269791b4d96a655e8a5a2 e size: 2.02 GB). Acesso em 2 de novembro de 2025.