EPISTEMOLOGIA · INTELIGÊNCIA ARTIFICIAL · GOVERNANÇA

A lacuna abdutiva em sistemas de inteligência artificial

Geração representacional, restrição empírico-formal e adjudicação epistêmica

Jandislei Antonio Genova

Ensaio técnico-acadêmico autoral independente

23 de agosto de 2026 · Versão 3.0

A questão decisiva não é apenas se o sistema produz uma hipótese surpreendente, mas como essa hipótese foi gerada, a que evidência responde e por que deve ser aceita ou abandonada.

Resumo

Este artigo examina a tese de que modelos de linguagem de grande escala seriam estruturalmente incapazes do salto abdutivo associado à descoberta científica [1]. Adota abordagem conceitual e documental baseada em conjunto de fontes deliberadamente selecionado, fechado em 23 de agosto de 2026 e hierarquizado por estatuto probatório. O argumento de Zahavy é confrontado com resultados em prova formal, busca matemática e pesquisa aberta [11–16]. A análise distingue modelo, agente, pipeline e sistema sociotécnico; separa geração, validação e integração comunitária; e trata a aprendizagem por limiar apenas como analogia para mudança aparentemente descontínua [17–18]. Decompõe a lacuna em geração representacional, restrição empírico-formal — ancoramento empírico ou controle semântico-formal, conforme o domínio — e adjudicação epistêmica. A expressão “ilhas de verificabilidade” descreve domínios nos quais candidatos podem ser avaliados com erro, custo e latência relativamente controláveis. A evidência não reduz toda produção artificial a recombinação trivial nem demonstra autonomia científica geral; alegações de descoberta exigem especificação do domínio, proveniência, avaliação independente e contribuição humana documentada.

Palavras-chave: inteligência artificial; abdução; descoberta científica; autonomia epistêmica; grounding; verificação formal; governança epistêmica.

Abstract

This article examines the claim that large language models are structurally incapable of the abductive jump associated with scientific discovery [1]. It uses a conceptual and documentary approach based on a deliberately selected source set, closed on 23 August 2026 and classified by evidentiary status. Zahavy’s argument is assessed against results in formal proof, mathematical search and open-ended research [11–16]. The analysis distinguishes models, agents, pipelines and sociotechnical systems; separates generation, validation and community integration; and treats threshold learning only as an analogy for apparently discontinuous change [17–18]. It decomposes the gap into representational generation, empirical–formal constraint — empirical grounding or formal semantic control, depending on the domain — and epistemic adjudication. “Islands of verifiability” describes domains in which candidates can be evaluated under relatively controlled error, cost and latency. Evidence neither reduces all artificial output to trivial recombination nor establishes general scientific autonomy; discovery claims require domain specification, provenance, independent evaluation and documented human contribution.

Keywords: artificial intelligence; abduction; scientific discovery; epistemic autonomy; grounding; formal verification; epistemic governance.

1. Introdução: problema, tese e método

“LLMs can’t jump” condensa em fórmula absoluta um problema que exige decomposição. Zahavy pergunta se uma inteligência artificial, dispondo do conhecimento acessível a Einstein, poderia ter formulado a relatividade geral. Sua resposta é negativa para os modelos de linguagem atuais: eles comprimiriam regularidades, executariam deduções e talvez detectassem inconsistências, mas não produziriam a passagem da experiência a novos princípios explicativos [1]. A literatura sobre descoberta, porém, distingue ao menos a geração de hipóteses, sua articulação e sua avaliação [4]. A metáfora do salto pode, por isso, encobrir três perguntas: o sistema produz uma representação nova? Essa representação responde às regras formais ou à evidência do mundo? O sistema seleciona e abandona hipóteses mediante razões auditáveis?

A evidência examinada não coincide nem com a leitura promocional nem com a negação categórica. Sistemas recentes produziram construções matemáticas que melhoraram resultados conhecidos e, em 2026, um modelo interno da OpenAI gerou o núcleo de uma prova que refutou uma conjectura associada às distâncias unitárias no plano; matemáticos externos confirmaram a validade do argumento [12–14]. Em avaliações de pesquisa aberta, por outro lado, agentes completaram a engenharia de projetos sem demonstrar julgamento suficiente para corrigir desenhos fracos ou retornar de caminhos improdutivos [16]. A diferença entre esses desempenhos, e não a escolha de um caso isolado, é o objeto a explicar.

A tese é condicional: a chamada lacuna abdutiva não foi demonstrada como incapacidade estrutural única. Sua extensão varia conforme arquitetura, domínio, acesso a evidência e qualidade do circuito de avaliação. Em ambientes formais, o retorno pode ser relativamente rápido e reproduzível, embora permaneçam riscos de especificação e de seleção do problema. Na ciência aberta, formular a pergunta, escolher a medida e construir o teste integram a própria tarefa. A unidade de análise relevante é, portanto, o sistema epistêmico composto e o arranjo institucional que seleciona, valida e comunica seus resultados.

Pergunta	Tese defendida	Teste decisivo
A IA consegue descobrir ou apenas explorar estruturas herdadas?	**Há hoje evidência de novidade relevante, mas não de autonomia epistêmica geral. A lacuna é graduada e tridimensional, não uma incapacidade já demonstrada.**	Atribuição independente de origem; grounding empírico ou controle semântico-formal, conforme o domínio; avaliação de rivais; teste adversarial; reprodutibilidade e integração explicativa.

Quadro 1 — Mapa do argumento. Elaboração do autor.

> Posição do artigo — O texto não pretende demonstrar que máquinas pensam como humanos nem formular impossibilidade sobre sistemas futuros. Seu objetivo é discriminar resultados observáveis,

> inferências condicionais e lacunas ainda abertas, evitando atribuir ao modelo isolado o que depende de verificadores, ferramentas, curadoria e revisão institucional.

1.1 Método documental, estratégia de busca e limites

A pesquisa assume a forma de ensaio conceitual e documental, não de revisão sistemática. O conjunto documental foi deliberadamente selecionado e encerrado em 23 de agosto de 2026. A identificação e a conferência das fontes utilizaram páginas oficiais de periódicos e editoras, registros DOI, arXiv, PMLR, Nature, MIT Press, Stanford Encyclopedia of Philosophy e páginas canônicas dos autores ou laboratórios. As buscas foram realizadas em português e inglês, com combinações de “abduction”, “scientific discovery”, “LLM”, “AI agents”, “world models”, “symbol grounding”, “threshold learning”, “mathematical discovery”, “open-ended AI research”, “Zürich Notebook”, “Goodhart” e “reward hacking”, além das expressões exatas “ilhas de verificabilidade” e “islands of verifiability”. Fontes históricas anteriores foram incluídas quando necessárias à genealogia conceitual; para resultados técnicos recentes, adotou-se o corte temporal do fechamento.

Foram incluídos trabalhos que definem conceitos mobilizados, descrevem desenhos ou resultados examinados, reconstroem o caso histórico de Einstein ou documentam a proveniência e o estatuto de um artefato. Foram excluídos duplicatas, comentários sem acesso ao documento subjacente e alegações promocionais sem material inspecionável. Artigos revisados por pares sustentam afirmações limitadas aos desenhos avaliados; preprints são tratados como evidência inicial; materiais institucionais documentam artefatos e declarações, mas não substituem validação independente. A seleção não pretende estimar prevalência nem esgotar a literatura.

O conjunto final contém 35 referências. As alegações técnicas recentes se apoiam sobretudo em [9–17,27–29]; a reconstrução independente do percurso de Einstein, em [32–33]; e a discussão sobre otimização de métricas, em [34–35]. A referência [21] registra apenas a gênese documental da analogia e foi excluída da cadeia probatória. Como não houve dupla triagem, protocolo PRISMA ou busca em todas as bases bibliográficas, a expressão “conjunto documental” é preferida a “corpus sistemático”.

Origem da associação heurística — Uma publicação pública de Tina Austin no LinkedIn motivou a aproximação inicial entre o argumento de Zahavy e a aprendizagem por limiar de convenções [21]. Cinco capturas de 23 de agosto de 2026 permanecem sob guarda do autor; o hash SHA-256 da concatenação ordenada dos cinco hashes é 6d3ad0d39211531793678228 2e09edd76d900c7b21353445ca9cd63303936d7e. A publicação documenta apenas a gênese da analogia e não integra a evidência científica. A seleção das fontes, a delimitação da analogia e as conclusões são de responsabilidade do autor.

1.2 Relação com a série autoral

Este artigo retoma problemas desenvolvidos em trabalhos anteriores sobre controle da base representacional, revisão ontológica e autonomia epistêmica relativa à tarefa [22–24]. Sua contribuição incremental consiste em confrontar a tese forte de Zahavy com evidências disponíveis até o fechamento do conjunto documental, delimitar a analogia com aprendizagem por limiar e organizar o problema nas dimensões de geração representacional, restrição empírico-formal e adjudicação. O Perceptual Twins Protocol é citado em sua versão científica corrente, v2.2.0, que não valida autonomia epistêmica e explicita as tarefas empíricas ainda pendentes [24]. Conceitos anteriores são pressupostos e não reapresentados como criações novas.

2. O que conta como descoberta?

2.1 Dedução, indução e abdução

Na tradição associada a Charles S. Peirce, a investigação científica articula três movimentos. A abdução introduz provisoriamente uma hipótese capaz de tornar inteligível um fato surpreendente; a dedução extrai consequências que deveriam ocorrer se a hipótese fosse válida; a indução confronta essas consequências com a experiência e reajusta o grau de confiança [2]. A fórmula é mais rica do que uma simples permutação de “regra, caso e resultado”. Em Peirce, a hipótese não recebe salvo-conduto por ser engenhosa. Ela conquista o direito de ser testada. A racionalidade do ciclo depende também da economia da pesquisa: tempo, custo, risco e poder discriminante dos experimentos limitam quais conjecturas merecem ser investigadas.

No uso contemporâneo, abdução também designa inferência para a melhor explicação. Harman enfatizou a comparação entre hipóteses rivais [3]; a literatura posterior mostrou que “melhor” pode envolver simplicidade, alcance, coerência, mecanismo causal e compatibilidade com conhecimento estabelecido [4,6]. Nenhuma dessas virtudes garante verdade. Uma hipótese pode ser a melhor de um conjunto ruim, e o conjunto considerado pode excluir a alternativa correta. Essa objeção é crucial para avaliar IA: gerar milhares de candidatos não assegura que o espaço de busca contenha uma representação adequada, nem que o avaliador recompense o que possui valor explicativo.

Zahavy usa a tríade para separar aquilo que modelos atuais fariam bem — indução estatística e dedução formal — da criação de axiomas [1]. A separação é útil, mas demasiado limpa. Na prática, indução, dedução e abdução se realimentam. Uma transformação dedutivamente obtida pode revelar um conceito inesperado; uma busca evolutiva pode modificar o espaço de representações; um contraexemplo pode obrigar o sistema a reformular o problema. Classificar uma operação pelo formato lógico de sua saída não basta para reconstruir o processo que a produziu.

2.2 Gerar, validar e integrar

Descoberta não é sinônimo de novidade textual, surpresa ou acerto isolado. Para evitar atribuições prematuras, este artigo separa três estágios. Um resultado candidato deve apresentar novidade relevante e não ser mera recuperação de fonte acessível. A validação exige prova, teste, mensuração ou procedimento público capaz de revelar erro. A integração ocorre quando uma comunidade competente pode examinar, reproduzir ou verificar e incorporar o resultado ao corpo de conhecimento. No sentido social amplo, “descoberta” designa o processo que alcança os três estágios; antes disso, empregam-se “resultado candidato” ou “contribuição”.

Essa definição impede dois atalhos. O primeiro atribui descoberta ao sistema sempre que sua saída é nova e correta, mesmo que humanos tenham formulado o problema, desenhado o verificador, selecionado a execução e reescrito o argumento. O segundo nega qualquer descoberta artificial porque treinamento, linguagem e ferramentas foram produzidos por pessoas. O mesmo critério eliminaria grande parte da ciência humana, que depende de professores, bibliotecas, instrumentos, laboratórios e instituições. A questão séria não é independência absoluta, inexistente em ambos os casos, mas a distribuição causal das contribuições e a transparência com que ela é documentada.

> Regra de atribuição — A contribuição do sistema deve ser avaliada por proveniência e por dependência contrafactual. Quanto mais o resultado depender de escolhas humanas não registradas — problema, métrica, amostragem, seleção de tentativas, edição e interpretação —, menos defensável será descrevê-lo como autônomo. A sobrevivência à auditoria independente é necessária, mas não basta: também é preciso identificar quais elementos novos deixariam de existir sem a intervenção do sistema.

2.3 Precedentes computacionais e unidade de análise

A tentativa de modelar processos de descoberta antecede os modelos de linguagem. Os programas BACON, DALTON, GLAUBER e STAHL foram apresentados como sistemas capazes de detectar regularidades, gerar hipóteses e, em condições delimitadas, criar conceitos ou representações de problema [25]. A historiografia filosófica registra tanto esses precedentes quanto a controvérsia entre geração, articulação e justificação [26]. O presente artigo não reivindica inaugurar a análise computacional da descoberta; examina como sistemas contemporâneos alteram sua escala, composição e regime de validação.

Trabalhos recentes operacionalizam tarefas denominadas abdutivas. AI-Noether infere axiomas candidatos para completar sistemas polinomiais incompletos; Graph of States organiza busca abdutiva por estados de crença e restrições neurossimbólicas; e MolQuest avalia geração e revisão de hipóteses químicas em interação com evidência espectral [27–29]. Esses preprints demonstram que a expressão “abdução em IA” já designa programas técnicos específicos. Não demonstram, por si, uma capacidade

unitária: cada trabalho define linguagem, tarefa e verificador próprios, e nenhum mede autonomia científica geral.

A cognição distribuída acrescenta uma cautela metodológica. Hutchins sustenta que a unidade adequada de análise não deve ser fixada antecipadamente, mas escolhida conforme o fenômeno e os elementos que interagem no sistema [30]. Essa perspectiva sustenta a passagem do LLM isolado ao sistema sociotécnico, sem converter toda contribuição coletiva em propriedade cognitiva do modelo.

3. A força e o excesso da tese “LLMs can’t jump”

3.1 O argumento de Einstein

O melhor aspecto do ensaio de Zahavy é reconstruir a invenção científica como alteração do sistema de representação, e não apenas ajuste de parâmetros. A historiografia do Zürich Notebook confirma que a relatividade geral resultou de um percurso com princípios físicos, colaboração com Marcel Grossmann, exploração do cálculo tensorial, tentativas abandonadas e retorno, em 1915, a estruturas examinadas anos antes [32-33]. O sistema final de equações não se segue como conclusão única das observações disponíveis. Essa reconstrução independente sustenta o valor do caso histórico sem depender da leitura de Zahavy [1,32-33].

A experiência do elevador em queda livre cumpre papel exemplar. Ao imaginar um observador que não sente o próprio peso, Einstein aproxima gravidade e aceleração local e reordena relações antes tratadas separadamente. O episódio integra uma trajetória mais longa, na qual restrições físicas e escolhas matemáticas se corrigiram mutuamente [32-33]. Ele é relevante para modelos de intervenção e contrafactuais, mas não deve ser convertido em algoritmo geral de descoberta [7,31].

3.2 Um estudo histórico não é um teorema de impossibilidade

A tese se torna vulnerável quando uma reconstrução histórica passa a sustentar incapacidade estrutural. O caso pode mostrar que a descrição da descoberta como compressão de dados abundantes é insuficiente [8]. Não demonstra, isoladamente, que toda arquitetura centrada em modelos de linguagem seja incapaz de formular premissas novas. Uma conclusão dessa força exigiria definição operacional de salto, delimitação da classe de sistemas, condições de acesso a dados e ferramentas e um programa de testes capaz de excluir mecanismos alternativos.

Há ainda um problema contrafactual insolúvel em sua forma forte. Não sabemos o que um sistema atual teria feito com “o conhecimento disponível a Einstein”, porque esse conjunto não é um arquivo neutro. A seleção de documentos, a forma de representação, os objetivos, o orçamento de busca e os instrumentos disponíveis alteram a tarefa. Se o sistema recebe o princípio de equivalência, o núcleo

inventivo já foi fornecido; se não recebe qualquer indicação de que gravidade merece revisão, o teste pode medir priorização, não capacidade representacional. O experimento mental é instrutivo, mas não constitui teste padronizado reproduzível.

Também é excessivo afirmar que não havia sinal de erro. Não existia uma função de perda supervisionada semelhante às utilizadas em aprendizagem de máquina, mas havia restrições: exigência de covariância, recuperação do limite newtoniano, igualdade entre massa inercial e gravitacional, dificuldades de conservação, anomalia do periélio de Mercúrio e coerência do formalismo. O Zürich Notebook e a sequência de trabalhos de 1912 a 1915 mostram que esses critérios foram mobilizados, abandonados e reinterpretados ao longo do percurso [32–33]. Eram sinais esparsos, heterogêneos e parcialmente escolhidos pelo pesquisador. A diferença sustenta a crítica à indução ingênua; não autoriza dizer que a invenção ocorreu sem gradiente normativo.

3.3 O erro de categoria: modelo não é sistema

O título fala em LLMs, mas a solução proposta incorpora modelos de mundo fisicamente consistentes, multimodalidade e controle de ações [1]. Isso desloca a unidade de análise. Um modelo de linguagem puro, congelado e sem ferramentas não equivale a um agente que consulta memória, escreve código, executa experimentos, controla um simulador, recebe retorno e modifica sua estratégia. Demonstrar a limitação do primeiro não demonstra a limitação do segundo. Da mesma forma, uma vitória do sistema composto não deve ser creditada automaticamente ao componente linguístico.

A oposição entre mente humana incorporada e modelo textual desincorporado exige precisão. Textos científicos preservam descrições de observações, instrumentos e práticas produzidas por agentes situados. O treinamento nesse material cria dependência representacional mediada em relação à experiência humana; não produz, por si, grounding sensorimotor intrínseco no sentido estrito de Harnad [5]. Chamar essa dependência de “ancoragem mediada” confundiria transmissão linguística de registros com ligação não simbólica entre categorias e referentes. O problema empírico é verificar se o sistema consegue usar medições e intervenções para discriminar mecanismos, em vez de apenas reproduzir linguagem causal.

> Veredicto provisório sobre Zahavy — O ensaio identifica um gargalo relevante e formula programa de pesquisa pertinente. A evidência apresentada, porém, não sustenta por si a passagem de limitação observada para incapacidade estrutural. A formulação compatível com o conjunto documental é mais estreita: ainda não foi demonstrado que sistemas atuais consigam, de modo geral, gerar representações, submetê-las a restrições empíricas ou formais e adjudicar novos princípios científicos.

4. Do salto ao limiar

4.1 A aparência descontínua da mudança

A metáfora do salto sugere ruptura interna súbita, mas a aparência observável pode resultar de processo cumulativo. Guilbeault, Caplan e Yang identificaram, em experimentos sobre aprendizagem de convenções, dois estágios: decisões probabilísticas enquanto a evidência era insuficiente e estabilização após um limiar [17]. O modelo foi inspirado no Princípio da Tolerância, formulado para explicar quando aprendizes tratam uma regularidade linguística como regra produtiva apesar das exceções [18]. A associação sugere uma hipótese sobre dinâmica de mudança; não estabelece mecanismo comum entre aprendizagem social e descoberta científica.

Essa distinção permite separar salto observável de salto processual. O primeiro é uma descontinuidade na saída: uma regra passa a ser aplicada, uma hipótese domina ou um padrão se estabiliza. O segundo exigiria uma descontinuidade real no mecanismo cognitivo. Do primeiro não se infere o segundo. Sistemas de busca, aprendizagem por reforço e seleção evolutiva podem exibir mudanças abruptas quando uma representação finalmente satisfaz múltiplas restrições, embora cada tentativa tenha sido gerada por operações graduais.

4.2 Limites da transferência

O paralelo não autoriza transplantar o modelo do PNAS para a descoberta científica. Os experimentos estudam convenções sociais e inferência de padrões sob condições controladas; não medem criação de teorias, elaboração de ontologias ou julgamento de fecundidade explicativa [17]. Uma aplicação científica exigiria definir o que conta como evidência, exceção, população de alternativas e cruzamento de limiar. Sem essas variáveis, a analogia permanece heurística. Ela orienta perguntas, não fornece resultado empírico sobre IA.

A analogia tem utilidade limitada: uma mudança abrupta na saída não prova ruptura igualmente abrupta no processo. Para avaliar sistemas artificiais, importa reconstruir a trajetória de busca, as evidências acumuladas, as alternativas descartadas e a regra de estabilização. Sem esses registros, “salto” permanece descrição do resultado visto de fora.

5. Evidências em direções opostas

5.1 Dedução formal: um território favorável

AlphaProof mostra como a dedução formal pode ser ampliada quando o ambiente oferece verificação precisa. O sistema combina rede neural de prova, busca em árvores, aprendizagem por reforço e o

verificador Lean. Na Olimpíada Internacional de Matemática de 2024, AlphaProof resolveu três problemas de álgebra e teoria dos números; AlphaGeometry 2 resolveu o problema de geometria. O conjunto alcançou 28 de 42 pontos, faixa de medalha de prata, com inferência de vários dias nos problemas mais difíceis [11]. O verificador reduz a ambiguidade da prova formal, mas não determina a relevância do problema nem garante que sua formalização inicial seja adequada.

O próprio artigo do AlphaProof reconhece a fronteira: olimpíadas operam numa biblioteca relativamente estável de conceitos, ao passo que pesquisa de ponta exige construção de teoria e expansão da biblioteca [11]. A distinção coincide com o núcleo de Zahavy, porém sem justificar barreira absoluta. As provas produzidas são examináveis por verificador formal, mas o material público não autoriza afirmar ausência de memorização sem auditoria de contaminação dos dados e busca de precedentes. O resultado defensável é mais estreito: busca, geração de variantes e verificação podem compor demonstrações formalmente válidas; permanece em aberto como o sistema seleciona novos conceitos e problemas.

5.2 Construções matemáticas e a conjectura de Erdős

AlphaEvolve combina propostas de programas por modelos de linguagem, seleção evolutiva e avaliadores automáticos. No preprint que descreve 67 problemas de análise, combinatória, geometria e teoria dos números, o sistema redescobriu melhores resultados conhecidos em grande parte dos casos e melhorou alguns deles; em tarefas específicas, generalizou resultados finitos para fórmulas [12]. Há construção e busca para além da prova de enunciado fornecido, mas a evidência permanece concentrada em problemas com critérios de correção e qualidade formalizáveis.

Entre os casos examinados, um dos mais fortes é a refutação de uma conjectura associada ao problema das distâncias unitárias no plano. O manuscrito *Planar Point Sets with Many Unit Distances* demonstra a existência de $\delta > 0$ tal que, para infinitos valores de n, existem conjuntos de n pontos com pelo menos $n^{1+\delta}$ pares à distância unitária, contrariando a expectativa de crescimento $n^{1+o(1)}$ [13]. A construção relaciona geometria discreta e teoria algébrica dos números. Segundo a declaração de uso do próprio projeto, um texto produzido por IA apresentou o problema a um modelo interno; a saída passou por avaliação automática, inspeção, verificação e reescrita assistida. Matemáticos externos confirmaram o núcleo da prova e publicaram uma versão humanamente verificada, com melhorias substanciais de exposição [13–14].

O episódio sustenta contribuição automatizada não trivial, mas não resolve o debate sobre autonomia. Segundo a documentação disponível, o pipeline produziu o núcleo novo da prova; especialistas externos confirmaram sua validade, e a exposição foi materialmente revista por humanos [13–14]. O problema, o critério de sucesso e o regime de verificação estavam fortemente especificados. A contribuição cognitiva do sistema pode ser reconhecida sem apagar a cadeia institucional de seleção, validação e comunicação que converteu a saída em conhecimento examinável.

5.3 Pesquisa aberta: quando o avaliador também precisa ser inventado

O AI Scientist apresentou uma arquitetura capaz de gerar ideias, escrever código, executar experimentos, produzir manuscritos e simular revisão em subcampos de aprendizagem de máquina [15]. O trabalho demonstra automação extensa do fluxo. Como o avaliador automático integra o próprio pipeline, existe risco de circularidade: desempenho diante desse avaliador não equivale, sem verificação externa, a contribuição reconhecida por comunidade independente. A inferência é um limite do desenho público, não prova de que todos os resultados sejam inválidos.

Kirgis e coautores realizaram avaliação mais adversarial. Agentes receberam as perguntas centrais de dois trabalhos inéditos submetidos à NeurIPS 2026, seis dias de execução e milhares de dólares em computação. Eles completaram a engenharia sem ajuda, mas não avançaram substancialmente nas questões de pesquisa. Os autores originais rejeitaram os resultados e identificaram falhas de julgamento sobre publicabilidade, correção de desenho, retorno de becos sem saída, percepção de recursos e cumprimento de instruções [16]. Por se tratar de preprint baseado em dois casos, a evidência é inicial; ela demonstra apenas que execução autônoma pode coexistir com baixa adjudicação epistêmica.

Evidência	Sistema	O que sustenta	O que não sustenta	Força
Provas da IMO 2024 (revisado por pares)	AlphaProof + AlphaGeometry 2	Dedução formal avançada em Lean; busca e aprendizagem com retorno verificável.	Construção autônoma de novas teorias ou escolha independente de problemas científicos.	Alta no domínio avaliado
67 problemas matemáticos (preprint)	AlphaEvolve e auxiliares	Descoberta de construções e melhoria de melhores resultados conhecidos em alguns casos.	Generalização para domínios sem avaliador objetivo, rápido e reprodutível.	Alta, com escopo delimitado
Distâncias unitárias (manuscrito institucional + verificação externa)	Modelo interno da OpenAI + avaliação automática + verificação humana	Novidade matemática relevante e conexão inesperada entre áreas; solução produzida de modo automatizado.	Autoria sem mediação institucional ou capacidade de fundar uma ontologia física.	Alta quanto ao resultado; parcial quanto à autonomia
Duas pesquisas abertas (preprint; dois casos)	Agentes de fronteira em avaliações-sombra	Engenharia autônoma, mas falhas de julgamento, criatividade corretiva e retorno de becos sem saída.	Impossibilidade geral; a amostra contém apenas dois casos.	Sugestiva, ainda inicial

Quadro 2 — Evidências contemporâneas e limites inferenciais. Fontes: [11]–[16]. Elaboração do autor.

5.4 Ilhas de verificabilidade

Por conveniência descritiva, este artigo denomina “ilhas de verificabilidade” os domínios em que candidatos podem ser avaliados com erro, custo, latência e reprodutibilidade relativamente controláveis, em comparação com o custo de gerá-los. Provas formais, programas com testes definidos e construções com função de pontuação são exemplos aproximados [11-12]. A expressão não designa teoria autônoma nem pressupõe verificadores perfeitos: falhas de especificação, formalização ou cobertura

podem produzir falsos aceites e falsos descartes. Busca complementar, nas fontes e pelos mecanismos descritos no método, não localizou até o fechamento uso diretamente equivalente das expressões exatas em português e inglês; ausência de resultado, contudo, não prova originalidade nem exclusividade terminológica.

Fora dessas ilhas, o verificador é parte do problema. Em biologia, medicina, economia ou direito, uma métrica intermediária pode premiar o efeito errado; dados carregam seleção institucional; resultados dependem de contexto; e a intervenção capaz de discriminar hipóteses pode ser cara, lenta ou eticamente vedada. Aumentar a quantidade de candidatos sem melhorar o regime de avaliação pode elevar a taxa de falsos positivos e o volume de narrativas convincentes sem lastro. O gargalo deixa de ser apenas gerar respostas e passa a incluir a construção de condições legítimas de erro.

6. Do LLM isolado ao sistema epistêmico híbrido

6.1 A arquitetura relevante

A pergunta “o LLM descobre?” deve ser substituída por “qual arranjo produziu, testou e integrou o resultado?”. Modelo designa o componente treinado; agente, uma configuração operacional que recebe objetivos, mantém estado e seleciona ações ou ferramentas; pipeline, o fluxo de operações automatizadas; sistema sociotécnico, o conjunto de modelos, verificadores, dados, pessoas e instituições. As fronteiras podem variar conforme a implementação, por isso devem ser declaradas em cada estudo. Confundir esses níveis infla crédito e oculta responsabilidade [22–24,30].

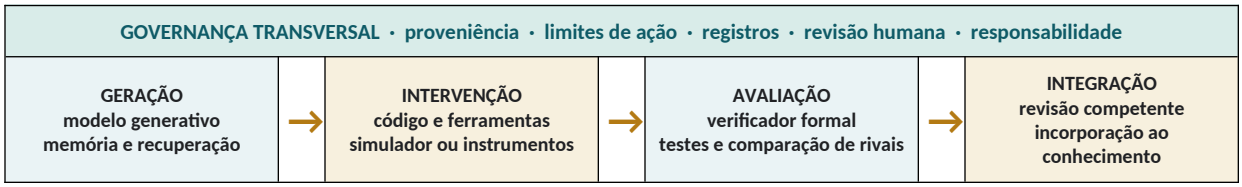

Figura 1 — Sistema epistêmico híbrido: a governança atravessa geração, intervenção, avaliação e integração. Elaboração do autor.

A mudança de unidade analítica melhora atribuição e crítica. Ela evita creditar ao modelo a correção garantida pelo Lean, ao agente a seleção efetuada por avaliador externo ou à instituição a origem de uma construção gerada pelo sistema. A governança não constitui etapa final: proveniência, limites de ação, registros e responsabilidade atravessam geração, intervenção, avaliação e integração.

6.2 Modelos de mundo e intervenção

Modelos de mundo procuram representar transições de um ambiente e permitir que um agente antecipe consequências de ações. Genie foi apresentado como ambiente generativo interativo treinado

em vídeos não rotulados, com espaço latente de ações e controle quadro a quadro [10]. A arquitetura permite variações e intervenções simuladas. Sua interpretação como contrafactual causal depende, contudo, da fidelidade das variáveis e relações aprendidas; simular uma sequência alternativa não basta para identificar causalidade [7,31].

Um simulador pode reproduzir regularidades visuais sem representar as variáveis causais relevantes. Vafa e coautores mostraram que modelos fundacionais podem obter bom desempenho em tarefas próximas da distribuição de treinamento e ainda recorrer a heurísticas específicas, falhando na transferência da mecânica newtoniana para novas tarefas [9]. O resultado não testa agentes incorporados; funciona como advertência contra inferir teoria causal a partir de desempenho preditivo. Em domínios empíricos, ancoramento exige ligação a medições ou intervenções capazes de discriminar mecanismos rivais, inclusive fora da distribuição de treinamento. Em domínios formais, o requisito correspondente é controle semântico-formal público; prova válida não equivale a grounding sensorimotor.

6.3 A armadilha ontológica

Todo simulador codifica ou aprende um espaço de estados, ações e regularidades. Se as categorias necessárias à nova teoria não estiverem representadas, o agente poderá explorar extensamente um domínio mal especificado. A lacuna não desaparece: desloca-se para a definição da ontologia, das variáveis omitidas e das transições consideradas impossíveis. Esse problema já foi desenvolvido, na série autoral, como controle e revisão da base representacional [22–23].

Uma arquitetura voltada à descoberta precisa permitir revisão representacional: criar e eliminar variáveis, comparar esquemas de categorias, detectar falhas sistemáticas e consultar evidência que não tenha sido produzida pelo próprio simulador. Essa capacidade é relativa à tarefa, ao espaço inicial de hipóteses, ao orçamento de interação e à autoridade de intervenção. Deve ser demonstrada por testes adversariais; a combinação “LLM + modelo de mundo” não a estabelece por si [23–24].

7. As três dimensões da lacuna abdutiva

Neste artigo, a lacuna abdutiva é decomposta para tornar comparáveis alegações que costumam aparecer agregadas. A decomposição não reivindica criação histórica da expressão nem pretende ser taxonomia exaustiva. Ela organiza três funções que podem avançar de forma desigual: geração representacional; restrição empírico-formal, realizada como grounding empírico ou controle semântico-formal conforme o domínio; e adjudicação epistêmica. As funções se relacionam a conceitos anteriores de base representacional, revisabilidade e autonomia relativa à tarefa [22–24].

Dimensão	Pergunta operacional	Falha típica	Evidência necessária
1. Geração representacional	O sistema cria primitivas, hipóteses ou enquadramentos que não estavam explicitamente dados?	Recombinação elegante, porém confinada ao vocabulário e aos objetivos fornecidos.	Auditoria de proveniência, busca de precedentes e análise contrafactual de dependência da instrução de entrada.
2. Restrição empírico-formal	Nos domínios empíricos, os símbolos se ligam a medições e intervenções? Nos formais, a semântica e as regras públicas controlam a inferência?	Coerência linguística sem vínculo empírico; ou prova aparente dependente de especificação ou formalização defeituosa.	Intervenção, medição e transferência fora da distribuição; ou prova formal independente com especificação auditada.
3. Adjudicação epistêmica	O sistema compara rivais, escolhe testes discriminantes, calibra incerteza e abandona linhas segundo regras registradas?	Persistência em desenho fraco, instrução desviada, má gestão de recursos ou confusão entre novidade e contribuição.	Comparação de rivais, retorno de becos sem saída, calibração, avaliação externa e decisão justificada de continuidade ou abandono.

Quadro 3 — As três dimensões da lacuna abdutiva. Elaboração do autor.

7.1 Geração representacional

A primeira dimensão é a capacidade de propor primitivas, relações, objetivos ou enquadramentos que não estavam explicitamente fornecidos. Isso não exige criação ex nihilo. O teste é contrafactual: retiradas determinadas pistas, fontes de recuperação ou ferramentas, o sistema ainda converge para uma estrutura funcionalmente equivalente? A inferência raramente é conclusiva, sobretudo em modelos fechados, mas ablações, busca independente de precedentes, diversidade de execuções e auditoria de contaminação podem reduzir a incerteza. O caso das distâncias unitárias indica conexão não trivial entre repertórios existentes; não demonstra criação de ontologia física [13–14].

7.2 Restrição empírico-formal

A segunda dimensão pergunta o que restringe a representação além da coocorrência linguística. No sentido estrito de Harnad, grounding requer que símbolos se apoiem, de baixo para cima, em representações não simbólicas, icônicas e categoriais ligadas à experiência [5]. Nas ciências empíricas, isso demanda conexão auditável com medições, instrumentos e intervenções. Na matemática, o regime é diferente: objetos e regras são definidos por uma semântica formal, e a prova pode ser verificada publicamente; esse controle não deve ser chamado de grounding sensorimotor. Modelos de mundo e robótica podem ampliar o ancoramento empírico, mas sua suficiência depende de transferência sob intervenções e situações novas [7,9–10,31].

7.3 Adjudicação epistêmica

A terceira dimensão é a capacidade funcional de comparar e interromper linhas de investigação mediante razões registradas. Indicadores observáveis incluem detectar controles inadequados, distinguir hipóteses rivais, escolher testes com poder discriminante, calibrar incerteza e abandonar estratégias quando regras de parada são satisfeitas. As avaliações-sombra localizaram falhas nesse nível:

os agentes produziram artefatos técnicos, mas mostraram julgamento insuficiente quanto ao desenho experimental, à alocação de recursos, à retomada após becos sem saída e ao padrão de contribuição publicável [16].

Adjudicação impede equiparar ciência a geração seguida de verificação local. Uma prova pode ser correta e irrelevante; um experimento, reproduzível e incapaz de discriminar rivais. Parte do julgamento pode ser operacionalizada por critérios prévios, comparação cega, valor esperado de informação e regras de parada. O restante deve permanecer visível como decisão humana ou institucional, em vez de ser substituído por nota opaca produzida pelo próprio sistema.

8. Aplicação das três dimensões: protocolo de auditoria

O protocolo traduz as três dimensões em sete perguntas documentais. Cada item deve ser registrado como documentado, parcialmente documentado, não documentado ou não aplicável, sempre com justificativa e localização da evidência. Não há pesos nem nota agregada. Um item não documentado não invalida automaticamente o resultado, mas impede a alegação mais forte que dependa dele. O protocolo é uma proposta de auditoria de conteúdo, ainda sem estudos de validade e confiabilidade interavaliadores nem delimitação empírica do domínio de aplicação.

Critério	Pergunta de auditoria	Evidência mínima	Risco se ausente
1. Participação na formulação	Quem escolheu o problema, a variável e o padrão de sucesso?	Registro das instruções de entrada, dos objetivos e das intervenções humanas.	Chamar de descoberta o cumprimento de uma tarefa inteiramente pré-especificada.
2. Novidade representacional	O que surgiu além de recuperação ou substituição superficial?	Busca de precedentes, ablações e comparação com bases acessíveis.	Confundir raridade textual com inovação conceitual.
3. Proveniência e independência	É possível reconstruir dados, versões, ferramentas e contribuições?	Registros de execução, versões, cadeia de custódia e declaração de edição humana.	Atribuição inflada, contaminação ou impossibilidade de auditoria.
4. Fecundidade explicativa	A proposta explica mais, unifica melhor ou gera consequências novas?	Predições, teoremas, mecanismos ou casos discriminantes.	Solução ad hoc sem ganho de compreensão.
5. Vulnerabilidade ao erro	O resultado pode ser formalmente refutado, invalidado ou empiricamente contrariado?	Prova verificável, teste pré-definido ou observação adversarial.	Narrativa imune à correção.
6. Reprodutibilidade externa	Terceiros conseguem repetir ou verificar o núcleo do resultado?	Artefatos suficientes e revisão independente competente.	Dependência exclusiva da autoridade do laboratório.
7. Integração responsável	Quem interpreta, comunica limites e responde por consequências?	Autoria/contribuição discriminadas e responsável institucional nomeado.	Antropomorfização do sistema e diluição da responsabilidade.

Quadro 4 — Protocolo mínimo de auditoria para alegações de descoberta por IA. Proposição do autor; categorias de registro sem escore agregado e instrumento ainda não validado.

A relação entre os níveis é a seguinte: participação na formulação, novidade e proveniência examinam a geração representacional; vulnerabilidade ao erro e reprodutibilidade verificam a restrição empírico-formal; fecundidade e integração responsável examinam adjudicação e incorporação comunitária. Proveniência e responsabilidade permanecem transversais. Assim, o quadro não acrescenta uma taxonomia concorrente, mas operacionaliza a decomposição apresentada na seção 7.

A terminologia subsequente é vocabulário de atribuição, não escala adicional. Quando o sistema executa plano definido e entrega medidas, há automação experimental. Quando propõe candidatos e pessoas escolhem o problema ou julgam o resultado, há descoberta assistida por IA. “Predominantemente produzida por IA” pode descrever casos em que o núcleo novo é gerado pelo sistema, sobrevive à auditoria de proveniência e à revisão independente, e as contribuições humanas são discriminadas. “Autonomia epistêmica” deve ser qualificada por tarefa, domínio, interfaces, espaço inicial de hipóteses, orçamento, regime de avaliação e autoridade de intervenção [24]. A ausência de correção humana posterior é relevante, mas não suficiente.

> Regra proposta de ônus informacional — A organização que anuncia descoberta produzida por IA controla registros de execução, versões, dados e critérios de seleção. Por isso, deve fornecer evidência proporcional sobre proveniência, independência e intervenção humana. A regra é uma proposta de governança e não é apresentada como norma jurídica universal já vigente.

9. Governança, autoria e responsabilidade

9.1 Atribuição sem antropomorfismo

Dizer que um sistema “descobriu” pode ser uma descrição causal útil, desde que não importe automaticamente consciência, intenção moral ou personalidade jurídica. A linguagem deve discriminar funções: quem formulou o problema, quem construiu o avaliador, qual modelo gerou o núcleo, quem selecionou execuções, quem verificou e quem escreveu a exposição. No manuscrito das distâncias unitárias, a declaração de uso é valiosa justamente por separar solução automatizada, avaliação por IA, inspeção interna, confirmação externa e edição humana [13]. Ela não elimina disputas de crédito; fornece material para julgá-las.

9.2 Responsabilidade não acompanha necessariamente o crédito

Mesmo quando uma contribuição cognitiva é atribuída ao sistema, a responsabilidade permanece nas pessoas e instituições que escolhem objetivos, disponibilizam recursos, divulgam conclusões e decidem aplicações. Modelos não respondem a parecer, não reparam dano e não assumem dever de retratação. Supervisão e documentação devem ser proporcionais ao risco, à irreversibilidade, à opacidade e ao alcance da decisão, e não apenas ao grau declarado de autonomia operacional.

9.3 Reprodutibilidade sob assimetria de acesso

Sistemas de fronteira podem depender de pesos privados, grande computação e fluxos não publicados. Isso intensifica uma assimetria conhecida na ciência: a comunidade consegue verificar parte do produto, mas não reproduzir o processo de geração. Em matemática, uma prova autocontida reduz o problema. Em ciência empírica, dados, instrumentos, limpeza e análise podem ser constitutivos do resultado. Quando a reprodução integral for inviável, devem-se distinguir replicação, reanálise e verificação e fornecer registros e acesso suficientes para auditar as alegações centrais.

9.4 O risco de uma ciência otimizada para o avaliador

A automação em larga escala pode alterar a seleção das perguntas. Quando agentes são otimizados por aprovação de avaliadores automáticos, métricas de impacto ou bancos de testes, o indicador pode deixar de representar a qualidade pretendida. A literatura distingue variantes desse efeito de Goodhart e identifica reward hacking como consequência possível de objetivos incompletos [34–35]. O risco também existe em instituições humanas; a diferença potencial está na escala, na velocidade e na capacidade de explorar lacunas do avaliador.

A resposta institucional inclui amostragem adversarial, replicação, diversidade de avaliadores, declaração de contribuições e mecanismos de contestação [34–35]. Também deve preservar espaços nos quais a métrica não esteja fechada antecipadamente. Critérios automatizados podem tornar invisíveis fenômenos ainda não convertidos em medida, inclusive anomalias capazes de motivar revisão teórica. A referência a mudanças de paradigma é principalmente kuhniana [20]; Popper sustenta, em outro plano, a exigência de teste e crítica [19].

10. Objeções e respostas

10.1 “Toda novidade da IA é recombinação”

A objeção de que toda novidade artificial é recombinação emprega critério que também excluiria resultados humanos. O ponto relevante é saber se a combinação produz estrutura não antecipada, correta segundo padrões declarados e resistente a revisão. Alguns resultados matemáticos examinados satisfazem novidade objetiva, correção formal e confirmação externa [12–14]. Isso não estabelece fecundidade de longo prazo nem autonomia geral.

10.2 “Uma prova nova demonstra inteligência científica geral”

Não. Matemática oferece verificadores excepcionalmente fortes. A passagem de uma prova nova para autonomia científica em domínios empíricos ignora formulação de problemas, qualidade de medidas,

causalidade e interpretação. O feito reduz a plausibilidade de uma impossibilidade geral, mas não demonstra competência uniforme. A extrapolação é tão indevida quanto negar o feito.

10.3 “A incorporação sensório-motora resolverá a abdução”

Sensores, corpos e simuladores ampliam experiências e intervenções disponíveis, mas não fornecem automaticamente categorias ou critérios explicativos adequados. O estudo de Vafa et al. não testa agentes incorporados; mostra que bom desempenho preditivo pode coexistir com representação heurística incapaz de transferência [9]. A inferência permitida é limitada: grounding pode ser necessário em certos domínios, mas não substitui geração representacional ou adjudicação.

10.4 “Se humanos verificam, a descoberta não é da IA”

Verificação posterior não apaga a origem de uma proposta, mas também não torna equivalentes autoria humana e produção automatizada. É preciso separar conferência, edição, correção do argumento central, seleção entre execuções e responsabilidade pela comunicação. Quando humanos introduzem a representação decisiva ou corrigem o núcleo, participam substantivamente da contribuição. Quando verificam uma prova completa e válida, o crédito causal pelo núcleo pode permanecer predominantemente no sistema, sem transferir a ele responsabilidade moral ou jurídica.

11. Agenda de pesquisa: como testar o salto

Uma agenda empírica sobre abdução artificial deve evitar demonstrações escolhidas depois do sucesso. Os itens seguintes operacionalizam as três dimensões: conjuntos prospectivos, auditoria de precedentes e mudança de representação examinam geração; intervenção discriminante examina restrição empírico-formal; retorno de becos sem saída e revisão externa cega examinam adjudicação; replicação entre arquiteturas testa robustez. O ineditismo deve ser definido por corte temporal, conjunto prospectivo, sigilo quando necessário e auditoria de contaminação. Formular o objeto correto deve ser avaliado separadamente de resolver um objeto já fornecido.

- **Conjuntos prospectivos.** Problemas e protocolos devem ser registrados antes da execução, com critérios de sucesso e fracasso.
- **Auditoria de precedentes.** Fontes acessíveis ao sistema e similaridades com literatura anterior precisam ser documentadas.
- **Mudança de representação.** O teste deve exigir criação ou revisão de variáveis, e não apenas busca dentro de uma linguagem fixa.
- **Intervenção discriminante.** O desenho deve privilegiar testes capazes de separar hipóteses rivais, com custo e risco explícitos.
- **Retorno de becos sem saída.** Avalia-se se o agente detecta um impasse e redireciona recursos antes de esgotá-los.

- **Revisão externa cega.** Especialistas independentes julgam contribuição, correção e relevância sem conhecer a origem humana ou artificial.
- **Replicação entre arquiteturas.** O mesmo protocolo deve ser executado por modelos e estruturas de orquestração diferentes para reduzir conclusões dependentes de um produto.

Esse desenho não responde se máquinas possuem intuição em sentido fenomenológico. Ele propõe programa empírico para investigar quando sistemas geram representações explicativas, escolhem testes informativos, corrigem erros e produzem resultados aceitos por especialistas independentes. Antes de qualquer uso comparativo, será necessário estabelecer validade de conteúdo, regras de decisão, treinamento dos avaliadores, concordância interavaliadores, sensibilidade a diferentes arquiteturas e poder discriminante. Sem esses estudos, o protocolo serve para documentação e crítica, não para ranking.

12. Limitações

O estudo é conceitual e documental. Não realiza revisão sistemática exaustiva, dupla seleção de estudos, auditoria de treinamento, reprodução dos experimentos ou acesso aos registros internos dos sistemas. Parte do conjunto recente consiste em preprints e materiais institucionais; mudanças posteriores de versão ou validação comunitária podem alterar as conclusões. Os casos matemáticos e as avaliações-sombra não permitem estimar prevalência de capacidade abdutiva em outras arquiteturas ou domínios.

A decomposição tridimensional, o rótulo “ilhas de verificabilidade” e o protocolo de sete perguntas são instrumentos analíticos propostos, não escalas ou benchmarks validados. A busca por expressão exata não estabelece originalidade exclusiva. Esses recursos não medem consciência, intenção, agência moral ou personalidade jurídica. A análise identifica condições de atribuição e teste; não demonstra que sejam suficientes para descoberta científica geral.

13. Conclusão

A pergunta inicial admite resposta condicional. Sistemas artificiais já produzem mais do que busca mecânica em listas explícitas: em domínios formais, combinam ferramentas, constroem objetos e geram provas capazes de resolver, refutar ou reconfigurar problemas abertos após validação competente. Esses resultados enfraquecem a redução de toda produção artificial a repetição. Não demonstram, contudo, que sistemas atuais formulem de modo geral novas ontologias, conectem conceitos a relações causais e selecionem sem mediação quais resultados merecem integração científica.

Zahavy localiza uma dificuldade relevante entre experiência e axiomas e mostra que dedução não inventa as próprias premissas [1]. O caso histórico isolado, porém, não sustenta a passagem para

incapacidade estrutural. A solução sugerida pelo próprio autor — modelos de mundo, ação e simulação — reconhece que a classe relevante de sistemas está em transformação. A avaliação deve acompanhar essa mudança sem creditar ao LLM isolado aquilo que depende de verificadores, instrumentos e instituições.

A aprendizagem por limiar oferece correção conceitual limitada: mudança abrupta na saída pode resultar de acumulação gradual seguida de estabilização [17–18]. Não há, neste artigo, evidência de que o mesmo mecanismo explique descoberta por IA. A analogia apenas desloca a análise do instante aparente para a trajetória de busca e para o regime que permite detectar erro.

Alegações de descoberta por IA devem ser aceitas, qualificadas ou rejeitadas conforme novidade demonstrável, proveniência auditável, vulnerabilidade ao erro, avaliação independente e contribuição humana registrada. Quando a instituição não documenta adequadamente sua participação na formulação do problema, na métrica, na seleção de execuções ou na narrativa, a atribuição deve ser estreitada. Crédito causal e responsabilidade não precisam coincidir: o primeiro pode ser distribuído entre componentes; a segunda permanece em pessoas e instituições capazes de responder, corrigir e reparar.

> A conclusão operacional é tríplice. Primeiro, resultados em domínios formalmente verificáveis não autorizam extrapolação para ciência aberta. Segundo, a unidade de análise deve abranger modelo, ferramentas, avaliadores e instituições. Terceiro, qualquer alegação de autonomia epistêmica deve especificar tarefa, domínio, interfaces, espaço inicial de hipóteses, orçamento computacional e de interação, proveniência, regime de avaliação e autoridade de intervenção. O artigo oferece critérios para essa discriminação; sua validação permanece agenda de pesquisa.

Glossário operacional

Termo	Sentido adotado neste artigo
Abdução	Geração provisória de hipótese explicativa ou seleção da melhor explicação entre rivais. O artigo distingue essas funções quando a diferença afeta a atribuição; nenhuma delas garante verdade.
Descoberta	Processo que combina resultado candidato novo, validação pública e integração por comunidade competente. Antes da integração, empregam-se "resultado candidato" ou "contribuição".
LLM	Modelo de linguagem de grande escala. Não se confunde com um agente dotado de memória, ferramentas, verificadores, sensores ou revisão humana.
Modelo de mundo (world model)	Modelo que representa e antecipa estados de um ambiente e pode permitir ações simuladas. Predição de transições não garante identificação causal nem contrafactual válido.
Salto observável	Mudança abrupta no desempenho ou na regra exibida, ainda que o processo subjacente tenha acumulado evidência gradualmente.
Ilha de verificabilidade	Domínio em que candidatos podem ser avaliados com latência, custo, reprodutibilidade e taxas de erro relativamente controláveis; expressão descritiva proposta neste artigo.
Geração representacional	Proposição ou revisão de primitivas, relações, objetivos ou enquadramentos não explicitamente fornecidos; sua atribuição exige auditoria de proveniência e dependência contrafactual.
Restrição empírico-formal	Dimensão que pergunta o que restringe uma representação além da coocorrência linguística. Nas ciências empíricas, assume a forma de ancoramento: em sentido estrito, grounding exige ligação não simbólica e sensorimotora a categorias e referentes. Em domínios formais, utiliza-se controle semântico-formal, não grounding empírico.
Adjudicação epistêmica	Seleção, comparação, continuidade ou abandono de linhas de investigação mediante razões e critérios registráveis.
Sistema epistêmico híbrido	Arranjo de modelos, agentes, ferramentas, verificadores, pessoas e instituições que participa da geração, avaliação e integração de um resultado.
Autonomia epistêmica relativa à tarefa	Capacidade demonstrada dentro de tarefa, domínio, interfaces, espaço inicial de hipóteses, orçamento, regime de avaliação e autoridade de intervenção especificados; não implica autonomia moral ou jurídica.

Referências

[1] ZAHAVY, Tom. LLMs can't jump. Position paper, 27 jan. 2026, 10 p. Disponível em: https://www.tomzahavy.com/files/llms-cant-jump.pdf. Acesso em: 23 ago. 2026.

[2] PEIRCE, Charles Sanders. Collected Papers of Charles Sanders Peirce. HARTSHORNE, Charles; WEISS, Paul; BURKS, Arthur W. (ed.). Cambridge, MA: Harvard University Press, 1931–1958. v. 2, §§ 619–644; v. 5, §§ 171–189.

[3] HARMAN, Gilbert H. The inference to the best explanation. The Philosophical Review, v. 74, n. 1, p. 88–95, 1965. DOI: 10.2307/2183532.

[4] DOUVEN, Igor. Abduction. Stanford Encyclopedia of Philosophy, revisão substantiva de 18 jun. 2025. Disponível em: https://plato.stanford.edu/entries/abduction/. Acesso em: 23 ago. 2026.

[5] HARNAD, Stevan. The symbol grounding problem. Physica D: Nonlinear Phenomena, v. 42, n. 1–3, p. 335–346, 1990. DOI: 10.1016/0167-2789(90)90087-6.

[6] MAGNANI, Lorenzo. Abductive Cognition: The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning. Berlin: Springer, 2009. DOI: 10.1007/978-3-642-03631-6.

[7] PEARL, Judea; MACKENZIE, Dana. The Book of Why: The New Science of Cause and Effect. New York: Basic Books, 2018.

[8] SCHMIDHUBER, Jürgen. Driven by compression progress: a simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In: SCHMIDHUBER, J. et al. (org.). Anticipatory Behavior in Adaptive Learning Systems. Berlin: Springer, 2009. p. 48–76. DOI: 10.1007/978-3-642-02565-5_4. Versão preliminar: 2008.

[9] VAFA, Keyon et al. What Has a Foundation Model Found? Using Inductive Bias to Probe for World Models. Proceedings of the 42nd International Conference on Machine Learning, PMLR 267, p. 60727–60747, 2025. Disponível em: https://proceedings.mlr.press/v267/vafa25a.html. Acesso em: 23 ago. 2026.

[10] BRUCE, Jake et al. Genie: Generative Interactive Environments. arXiv:2402.15391, 23 fev. 2024. DOI: 10.48550/arXiv.2402.15391. Preprint.

[11] HUBERT, Thomas et al. Olympiad-level formal mathematical reasoning with reinforcement learning. Nature, v. 651, p. 607–613, 2026. DOI: 10.1038/s41586-025-09833-y.

[12] GEORGIEV, Bogdan et al. Mathematical exploration and discovery at scale. arXiv:2511.02864, versão 3, 22 dez. 2025. DOI: 10.48550/arXiv.2511.02864. Preprint.

[13] OPENAI. Planar Point Sets with Many Unit Distances. Manuscrito técnico, 2026, 18 p. Disponível em: https://cdn.openai.com/pdf/74c24085-19b0-4534-9c90-465b8e29ad73/unit-distance-proof.pdf. Acesso em: 23 ago. 2026.

[14] ALON, Noga et al. Remarks on the disproof of the unit distance conjecture. arXiv:2605.20695, versão 1, 20 maio 2026, 19 p. DOI: 10.48550/arXiv.2605.20695. Disponível em: https://arxiv.org/abs/2605.20695. Acesso em: 23 ago. 2026. Preprint humanamente verificado do contraexemplo gerado por modelo interno da OpenAI.

[15] LU, Chris et al. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. arXiv:2408.06292, versão 3, 1° set. 2024. DOI: 10.48550/arXiv.2408.06292. Preprint.

[16] KIRGIS, Peter et al. Can AI agents conduct open-ended AI research? Early evidence from two case studies. arXiv:2607.27191, versão 2, 7 ago. 2026. DOI: 10.48550/arXiv.2607.27191. Preprint.

[17] GUILBEAULT, Douglas; CAPLAN, Spencer; YANG, Charles. A simple threshold captures the social learning of conventions. Proceedings of the National Academy of Sciences, v. 123, n. 17, e2508061123, 2026. DOI: 10.1073/pnas.2508061123.

[18] YANG, Charles. The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language. Cambridge, MA: MIT Press, 2016.

[19] POPPER, Karl. The Logic of Scientific Discovery. London: Routledge, 2002 [1959].

[20] KUHN, Thomas S. The Structure of Scientific Revolutions. 4. ed. Chicago: University of Chicago Press, 2012 [1962].

[21] AUSTIN, Tina. LLMs can't "jump" (anytime soon). Publicação no LinkedIn, ago. 2026. Registro de gênese documental composto por cinco capturas realizadas em 23 ago. 2026; hash SHA-256 da concatenação ordenada dos cinco hashes: 6d3ad0d392115317936782282e09edd76d900c7b21353445ca9cd63303936d7e. Perfil: https://www.linkedin.com/in/tinaaustin. Material arquivado; não utilizado como evidência científica.

[22] GENOVA, Jandislei Antonio. Controle da base representacional e soberania cognitiva na ciência assistida por inteligência artificial: uma analogia com a análise de Fourier. São Paulo: publicação independente, 4 ago. 2026. Disponível em: https://jandisleigenova.com/artigos/controle-base-representacional-ia-fourier. Acesso em: 23 ago. 2026.

[23] GENOVA, Jandislei Antonio. A IA poderia reconstruir a própria base? Da coerência linguística à coerência perceptivo-agentiva: modelos de mundo, causalidade e governança da revisão representacional em sistemas artificiais. Versão 5.0. São Paulo: publicação independente, 6 ago. 2026. Disponível em: https://jandisleigenova.com/artigos/a-ia-poderia-reconstruir-a-propria-base. Acesso em: 23 ago. 2026.

[24] GENOVA, Jandislei Antonio. Toward a Perceptual Twins Protocol: A Yoked Causal Design for Task-Relative Epistemic Autonomy. Independent preprint, version 2.2.0, 12 Aug. 2026, 38 p. Updated 13 Aug. 2026. Not peer reviewed. Disponível em: https://jandisleigenova.com/artigos/testing-task-relative-epistemic-autonomy-perceptual-twins. Acesso em: 23 ago. 2026.

[25] LANGLEY, Patrick W.; SIMON, Herbert A.; BRADSHAW, Gary L.; ZYTKOW, Jan M. Scientific Discovery: Computational Explorations of the Creative Process. Cambridge, MA: MIT Press, 1987.

[26] SCHICKORE, Jutta. Scientific Discovery. In: ZALTA, Edward N.; NODELMAN, Uri (ed.). The Stanford Encyclopedia of Philosophy. Fall 2025 ed. Stanford: Metaphysics Research Lab, 2025. Disponível em: https://plato.stanford.edu/archives/fall2025/entries/scientific-discovery/. Acesso em: 23 ago. 2026.

[27] SRIVASTAVA, Karan et al. Bridging the Gap Between Scientific Laws Derived by AI Systems and Canonical Knowledge via Abductive Inference with AI-Noether. arXiv:2509.23004, versão 2, 22 dez. 2025, 47 p. DOI: 10.48550/arXiv.2509.23004. Disponível em: https://arxiv.org/abs/2509.23004. Acesso em: 23 ago. 2026. Preprint.

[28] LUO, Yu et al. Graph of States: Solving Abductive Tasks with Large Language Models. arXiv:2603.21250, versão 2, 14 maio 2026. DOI: 10.48550/arXiv.2603.21250. Disponível em: https://arxiv.org/abs/2603.21250. Acesso em: 23 ago. 2026. Preprint.

[29] HAN, Taolin et al. MolQuest: A Benchmark for Agentic Evaluation of Abductive Reasoning in Chemical Structure Elucidation. arXiv:2603.25253, versão 1, 26 mar. 2026. DOI: 10.48550/arXiv.2603.25253. Disponível em: https://arxiv.org/abs/2603.25253. Acesso em: 23 ago. 2026. Preprint.

[30] HUTCHINS, Edwin. Enaction, imagination, and insight. In: STEWART, John; GAPENNE, Olivier; DI PAOLO, Ezequiel A. (ed.). Enaction: Toward a New Paradigm for Cognitive Science. Cambridge, MA: MIT Press, 2010. p. 425–450. DOI: 10.7551/mitpress/9780262014601.003.0016.

[31] PEARL, Judea. Causality: Models, Reasoning, and Inference. 2. ed. Cambridge: Cambridge University Press, 2009. DOI: 10.1017/CBO9780511803161.

[32] STRAUMANN, Norbert. Einstein's 'Zürich Notebook' and his journey to general relativity. Annalen der Physik, v. 523, n. 6, p. 488–500, 2011. DOI: 10.1002/andp.201110467.

[33] JANSSEN, Michel; RENN, Jürgen. Arch and scaffold: How Einstein found his field equations. Physics Today, v. 68, n. 11, p. 30–36, 2015. DOI: 10.1063/PT.3.2979.

[34] MANHEIM, David; GARRABRANT, Scott. Categorizing Variants of Goodhart's Law. arXiv:1803.04585, 13 mar. 2018. DOI: 10.48550/arXiv.1803.04585. Disponível em: https://arxiv.org/abs/1803.04585. Acesso em: 23 ago. 2026. Preprint.

[35] AMODEI, Dario et al. Concrete Problems in AI Safety. arXiv:1606.06565, 21 jun. 2016. DOI: 10.48550/arXiv.1606.06565. Disponível em: https://arxiv.org/abs/1606.06565. Acesso em: 23 ago. 2026. Preprint.

Como citar

GENOVA, Jandislei Antonio. A lacuna abdutiva em sistemas de inteligência artificial: geração representacional, restrição empírico-formal e adjudicação epistêmica. São Paulo: publicação independente, 2026. Versão 3.0, 23 ago. 2026.

Declarações

Natureza e escopo. Ensaio crítico de caráter conceitual e documental baseado em conjunto deliberadamente selecionado de 35 referências. A decomposição tridimensional, o uso descritivo de "ilhas de verificabilidade" e o protocolo de sete perguntas são construções analíticas deste artigo. Não constituem instrumentos de medida validados nem reivindicações de exclusividade terminológica.

Independência. O texto não possui vínculo institucional com os laboratórios, empresas ou autores examinados. Menções a produtos e resultados não constituem endosso comercial.

Controle de inferências. Artigos revisados por pares, livros, preprints, manuscritos institucionais e registros de proveniência foram discriminados segundo seu estatuto. Evidência de dois estudos de caso foi tratada como inicial; a aprendizagem por limiar, como analogia; os resultados matemáticos, como evidência limitada aos domínios e verificadores examinados; e a publicação de rede social [21], apenas como registro da gênese documental.

Responsabilidade autoral e assistência de IA. O autor definiu o problema, a tese e o recorte; selecionou e avaliou as fontes; decidiu aceitar, rejeitar ou reescrever cada formulação; e assume responsabilidade integral pela versão final. Sistemas generativos da OpenAI foram utilizados em agosto de 2026 para reconstrução estrutural e redação de versões intermediárias, formulação de alternativas, busca bibliográfica dirigida, conferência de referências e versões, revisão adversarial, testes de consistência e revisão linguística. As saídas foram submetidas à decisão autoral e à verificação contra fontes recuperáveis. A assistência não constitui validação científica externa, não transfere autoria e não substitui revisão independente.