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

COULD ARTIFICIAL INTELLIGENCE RECONSTRUCT ITS OWN REPRESENTATIONAL BASIS?

From Linguistic Coherence to Perceptual-Agentic Coherence: World Models, Causality, and Governance of Representational Revision in Artificial Systems

Jandislei Antonio Genova

Advogado e pesquisador independente. Especialista em Direito Civil e em Gestão em Saúde. São Paulo, Brasil.

Artigo II da série “Da base representacional à autonomia artificial”. Versão acadêmica integral 5.0 - 6 de agosto de 2026.

RESUMO

O artigo investiga em que condições sistemas de inteligência artificial poderiam revisar bases representacionais mediante interação com ambientes físicos ou simulados. Parte-se da analogia entre a aprendizagem de regularidades linguísticas em sequências simbólicas e a aprendizagem de regularidades perceptivas e causais por sistemas incorporados, preservando-se as diferenças entre texto e mundo material. Adota-se ensaio teórico-jurídico documental sobre modelos de mundo, aprendizagem causal, robótica e governança de sistemas autônomos. Propõe-se o conceito de coerência perceptivo-agentiva para descrever a correspondência dinâmica entre percepção, representação, previsão, intervenção ou decisão de não agir, consequência, erro e revisão. Distinguem-se atualização paramétrica, revisão estrutural e reconstrução ontológica, sem presumir que sistemas atuais tenham alcançado a última capacidade de modo geral. Sustenta-se que competência técnica, autonomia operacional e autorização jurídica devem ser avaliadas em eixos independentes e que a decisão técnica de não agir não equivale, por si, a omissão jurídica da máquina. Conclui-se que o reconhecimento técnico de uma candidata a revisão autônoma exige evidência causal, testes discriminantes e generalização; sua alegação pública e seu uso autorizado requerem, adicionalmente, validação independente, rastreabilidade, reversibilidade, supervisão e responsabilidade humana e institucional identificável.

Palavras-chave: inteligência artificial; modelos de mundo; causalidade; sistemas autônomos; base representacional; coerência perceptivo-agentiva.

ABSTRACT

This article examines the conditions under which artificial intelligence systems could revise representational bases through interaction with physical or simulated environments. It draws an analogy between learning linguistic regularities from symbolic sequences and learning perceptual and causal regularities by embodied systems, while preserving the differences between text and the material world. The study adopts a theoretical, legal, and documentary approach to world models, causal learning, robotics, and the governance of autonomous systems. It proposes the concept of perceptual-agentic coherence to describe the dynamic correspondence between perception, representation, prediction, intervention or a decision not to act, consequence, error, and revision. Parametric updating, structural revision, and ontological reconstruction are distinguished without assuming that current systems have generally achieved the latter capability. The article argues that technical competence, operational autonomy, and legal authorization must be assessed on independent

axes and that a technical decision to refrain from acting does not, by itself, constitute a legal omission by the machine. It concludes that technical recognition of a candidate autonomous revision requires causal evidence, discriminating tests, and generalization; public claims and authorized use additionally require independent validation, traceability, reversibility, oversight, and identifiable human and institutional responsibility.

Keywords: artificial intelligence; *world models*; causality; autonomous systems; representational basis; perceptual-agentic coherence.

1 INTRODUÇÃO

A evolução recente da inteligência artificial produziu uma mudança metodológica relevante. Sistemas de linguagem passaram a gerar textos coerentes sem que todas as regras gramaticais, semânticas e pragmáticas fossem previamente codificadas por programadores. O resultado não autoriza concluir que a lógica tenha fracassado, nem que os algoritmos tenham surgido apenas com o advento de grandes volumes de dados. Algoritmos e sistemas simbólicos são anteriores às redes neurais contemporâneas e continuam indispensáveis. A transformação consistiu, sobretudo, na passagem de arquiteturas predominantemente baseadas em regras explícitas para sistemas capazes de aprender regularidades estatísticas em grandes corpora e reutilizá-las em contextos não enumerados antecipadamente (Newell; Simon, 1976; Vaswani et al., 2017).

Essa transição sugere uma hipótese mais ampla. Se regularidades linguísticas puderam ser aprendidas a partir de sequências simbólicas, regularidades do mundo físico também poderiam ser extraídas de fluxos de imagens, sons, profundidade, movimento, tato, estados internos e consequências de ações. A máquina não receberia apenas descrições humanas do mundo; formaria representações operacionais mediante a relação entre o que percebe, o que prevê, a intervenção que executa ou deixa de executar e o resultado observado.

A analogia, contudo, não estabelece identidade entre linguagem e mundo. Textos são registros discretizados e produzidos por agentes que já organizaram a realidade em categorias. O mundo material é parcialmente observável, causalmente denso, interativo e capaz de impor efeitos irreversíveis. A previsão do próximo token não equivale à previsão de uma queda, do agravamento de uma emergência ou do resultado de uma ação robótica. A intervenção modifica a fonte dos próprios dados, e o erro pode produzir dano físico, jurídico ou social.

Em trabalho anterior, foi desenvolvida a distinção entre controle dos coeficientes e controle da base representacional. O primeiro corresponde à modulação de saliência, recorrência, ordem e temporalidade; o segundo alcança categorias, ontologias, variáveis e espaços de hipótese que tornam perguntas e alternativas formuláveis (Genova, 2026a). O presente artigo desloca a investigação: se a base não fosse apenas fornecida por projetistas, instituições e dados históricos, um sistema poderia detectar sua insuficiência, propor uma alternativa, testá-la e operar com a representação revisada?

O problema principal é: em que condições um sistema artificial poderia revisar partes de sua base representacional mediante interação com o mundo e quais consequências jurídicas decorreriam da autorização para agir, não agir, interromper ou solicitar supervisão com base nessa revisão? A hipótese possui duas dimensões. Tecnicamente, grandes fluxos sensoriais são insuficientes sem um ciclo fechado de percepção, previsão, intervenção, consequência e revisão causal. Juridicamente, mesmo uma revisão funcional da base não produz autonomia normativa: a autoridade para atuar permanece dependente de competência definida, supervisão, limites revogáveis e imputação institucional.

A contribuição específica deste artigo não reside na criação isolada das ideias de agência incorporada, evolução reconstrutiva de modelos ou autonomia ontológica, que possuem antecedentes. Consiste em formular a coerência perceptivo-agentiva como constructo integrador; separar grau de modificação representacional, autonomia operacional e autorização jurídica; e ligar a revisão da base a

critérios causais, rastreabilidade, reversibilidade e imputação institucional (Burr et al., 2026; López-Díaz; Gershenson, 2026).

O artigo não pretende demonstrar consciência, subjetividade ou personalidade jurídica artificial. Tampouco sustenta que sistemas atuais tenham alcançado reconstrução ontológica aberta. O objetivo é construir critérios para distinguir adaptação paramétrica, revisão estrutural e modificação das entidades ou categorias relevantes, separando capacidade técnica, autonomia operacional e legitimidade normativa.

Na série autoral, o Artigo I diagnostica o controle externo da base representacional e propõe o Relatório de Impacto Epistêmico; o presente Artigo II formula a ponte entre aprendizagem do mundo, revisão da base, ação e governança; e o benchmark Perceptual Twins oferece operacionalização mais estreita da autonomia epistêmica relativa à tarefa, sem transformar sua prova sintética em validação de agentes reais (Genova, 2026a; Genova, 2026b).

2 METODOLOGIA, CORPUS E ARQUITETURA ANALÍTICA

2.1 Desenho da pesquisa

A pesquisa assume a forma de ensaio teórico-jurídico documental, de natureza qualitativa e comparativa. O corpus foi encerrado em 6 de agosto de 2026 e reúne seis grupos de fontes: história e fundamentos da inteligência artificial; linguagem, cognição incorporada e contingências sensório-motoras; modelos de mundo e robótica; aprendizagem de representações e abstrações causais; segurança e garantia de sistemas autônomos; e instrumentos jurídicos relativos a supervisão, máquinas, responsabilidade, decisões automatizadas e gestão de riscos.

O levantamento foi dirigido, sem pretensão de revisão sistemática. Foram consultados repositórios oficiais e primários - Planalto, EUR-Lex, NIST, páginas normativas da ISO e cartões de modelo - e repositórios acadêmicos ou de editoras, como PMLR, arXiv, *OpenReview* e páginas de periódicos. As buscas combinaram, em português e inglês, os termos *world models*, *causal representation learning*, *causal abstraction*, *ontology evolution*, *active inference*, *self-modeling robotics*, *autonomous systems*, *human oversight*, *self-evolving machinery*, *agentic digital twins*, *reconstructive model evolution*, *open-ended agency* e *product liability*. Foram incluídas fontes que sustentassem diretamente uma capacidade técnica, uma limitação, uma categoria analítica ou uma regra jurídica; comunicações jornalísticas e reproduções sem documentação técnica foram excluídas. O corpus foi encerrado quando novas buscas deixaram de acrescentar categorias relevantes à matriz analítica, preservada a possibilidade de atualização diante de novos resultados ou alterações normativas.

As fontes técnicas foram interpretadas de acordo com o ambiente de demonstração e com o tipo de saída produzido. Resultados em jogos, simulações e benchmarks não foram automaticamente transpostos ao mundo físico. Experimentos robóticos controlados foram tratados como evidência de competência localizada. Páginas institucionais e *model card*s foram utilizados para identificar capacidades reivindicadas, arquitetura declarada, limitações e condições de uso, preservando-se a diferença entre comunicação do fornecedor e validação acadêmica independente.

A análise jurídica separa conteúdo normativo, entrada em vigor, data de aplicação e proposta de lege ferenda. O AI Act é examinado com as alterações do Regulamento (UE) 2026/1744. O Regulamento de Máquinas é utilizado por antecipar problemas de lógica autoevolutiva, registros e supervisão de máquinas autônomas. A Diretiva de Responsabilidade por Produtos é analisada com sua retificação de maio de 2026. O Direito brasileiro é empregado para estruturar imputação por projeto, integração, operação, supervisão e risco da atividade.

2.2 Eixo A: contexto probatório

A força da evidência não decorre de uma escala única de inteligência. Depende do contexto no qual a competência foi demonstrada e da distância entre esse contexto e o uso pretendido. O Quadro 1 separa o ambiente probatório da natureza da mudança representacional.

Código	Contexto	Conclusão admissível	Limite
A1 - simulação delimitada	Jogos, ambientes virtuais e tarefas com regras, ações ou recompensas previamente definidas.	O sistema aprende representação útil para previsão ou planejamento naquele domínio.	Não demonstra transferência ao mundo físico nem desempenho em ambiente aberto.
A2 - laboratório físico	Robôs e sensores em tarefas materiais delimitadas, com objetivos e condições de segurança definidos.	Há integração entre percepção, planejamento e ação nas condições testadas.	Não prova autonomia fora do domínio operacional nem legitimidade para aplicações críticas.
A3 - campo controlado	Operação física com variabilidade real, supervisão, área delimitada e protocolos de intervenção.	Há evidência de robustez contextual e transferência parcial.	Não equivale a generalidade aberta ou autoridade irrestrita.
A4 - ambiente aberto	Situações parcialmente novas, múltiplos agentes, mudanças não previstas e consequências materiais relevantes.	Permite avaliar generalização, contenção e adaptação em condições próximas ao uso real.	Não elimina risco residual nem transforma competência em legitimidade normativa.

Quadro 1 - Contextos probatórios da aprendizagem do mundo. Fonte: elaboração do autor.

2.3 Eixo B: grau de modificação da base

Código	Operação	Critério mínimo	Conclusão admissível
M0 - base herdada	Opera com entidades, relações e objetivos fornecidos.	Desempenho é obtido sem alterar a estrutura representacional.	Uso competente de base externa.
M1 - atualização paramétrica	Altera pesos, probabilidades, embeddings ou confiança, preservando entidades e relações básicas.	Melhora calibração ou previsão sem mudar a ontologia operacional.	Adaptação paramétrica.
M2 - revisão estrutural	Introduz variável, relação, decomposição ou abstração dentro de vocabulário ainda reconhecível.	A mudança reduz erros persistentes e melhora testes ou contrafactuais em contextos distintos.	Revisão funcional da estrutura.
M3 - reconstrução ontológica	Redefine entidades, fronteiras, categorias ou mapeamentos entre níveis de descrição.	Base alternativa supera a anterior em previsão causal ou interventiva e em transferência, sob testes discriminantes.	Candidato técnico a reconstrução ontológica.

Quadro 2 - Graus de modificação da base representacional. Fonte: elaboração do autor.

A combinação dos eixos impede inferências indevidas. Um resultado A1-M2 pode indicar revisão estrutural em simulação, sem autorizar conclusão sobre robótica aberta. Um resultado A2-M1 demonstra adaptação paramétrica em laboratório, não reconstrução ontológica. Evidência técnica forte para M3 exigiria replicação, ambientes heterogêneos, intervenções discriminantes e ganho causal ou contrafactual com transferência. Rastreabilidade permite auditoria; validação independente condiciona a classificação pública, não a existência da capacidade.

2.4 Três dimensões que não devem ser fundidas

A análise distingue competência representacional, autonomia operacional e grau de autorização jurídica do uso. Competência descreve o que o sistema consegue modelar ou revisar. Autonomia operacional descreve quanto pode decidir e executar sem intervenção imediata. Autorização jurídica do uso define o que foi legitimamente autorizado. Um sistema pode possuir alta

competência e autorização zero; outro pode receber autorização estreita apesar de competência limitada.

2.5 Estratos da base representacional

A expressão base representacional é utilizada em sentido estratificado. Ela não torna equivalentes sinais, representações latentes, ontologias explícitas, modelos causais e políticas de decisão. O artigo distingue cinco estratos funcionalmente conectados:

1. Camada sensorial: sinais, resoluções, modalidades, corpos, sensores e critérios de captação que delimitam o observável.
2. Representação latente: estados internos, *embeddings* e compressões aprendidas que organizam regularidades sem exigir semântica explícita.
3. Entidades e variáveis: categorias, fronteiras, propriedades e macrovariáveis dotadas de significado operacional suficiente para testes e auditoria.
4. Modelo causal ou de mundo: relações temporais, interventivas e contrafactuais usadas para previsão, explicação e planejamento.
5. Política de decisão: regras que selecionam agir, não agir, consultar, interromper ou retornar a estado seguro dentro de objetivos e permissões externos.

M1 pode ocorrer predominantemente na representação latente; M2 exige alteração verificável de relações, variáveis ou abstrações; M3 alcança entidades, fronteiras ou mapeamentos entre níveis e precisa preservar significado operacional e relações interventivas. Na engenharia do conhecimento, ontologias são artefatos formais e sua evolução responde a mudanças no domínio ou em sua conceptualização. Essa tradição é relevante, mas não deve ser confundida com qualquer mudança não interpretável em estados latentes de uma rede neural (Zablith et al., 2015).

3 DA LÓGICA PROGRAMADA ÀS REGULARIDADES APRENDIDAS

3.1 Legado simbólico e ambientes abertos

A inteligência artificial simbólica representou problemas por símbolos, regras e busca. Newell e Simon formularam a hipótese do sistema físico de símbolos como meio necessário e suficiente para a ação inteligente geral, tese histórica cuja suficiência permanece contestada. Sistemas especialistas, provadores de teoremas e planejadores mostraram que regras explícitas produzem desempenho elevado quando entidades, relações e objetivos são previamente definidos (Newell; Simon, 1976).

A fragilidade tornou-se mais visível em linguagem, percepção e ação cotidianas, que envolvem ambiguidade, conhecimento tácito, exceções, mudança contextual e combinações não enumeráveis. O problema não foi a inutilidade da lógica, mas a rigidez de antecipar todas as regras e a dependência de engenheiros que definissem a ontologia antes da operação.

Brooks criticou arquiteturas que separavam representação e ação por longas cadeias internas, defendendo inteligência situada. Gibson, Varela, Thompson e Rosch, Clark e a teoria das contingências sensório-motoras enfatizaram que percepção e ação são acopladas e que oportunidades de ação dependem do corpo, das capacidades e do contexto (Brooks, 1991; Gibson, 1979; Varela; Thompson; Rosch, 1991; Clark, 1997; O'Regan; Noë, 2001). Essas tradições não eliminam necessariamente representações internas, mas impedem tratá-las como descrições estáticas isoladas da exploração do ambiente.

3.2 Coerência linguística operacional

A aprendizagem estatística deslocou parte do trabalho representacional. Em vez de receber todas as regras, o sistema ajusta parâmetros para reduzir erro de predição em grandes conjuntos de exemplos. O Transformer ampliou a capacidade de relacionar elementos distantes de uma sequência e

tornou o treinamento paralelo mais eficiente, contribuindo para modelos de linguagem em larga escala (Vaswani et al., 2017).

O resultado pode ser descrito como coerência linguística operacional: regularidades permitem continuar, resumir, traduzir e reorganizar sequências de modo consistente. Essa coerência não deve ser confundida automaticamente com significado enraizado. Bender e Koller distinguem forma linguística e significado; Harnad formula o problema do enraizamento simbólico ao perguntar como símbolos adquirem referência para além de relações internas (Harnad, 1990; Bender; Koller, 2020).

3.3 Genealogia da coerência perceptivo-agentiva

O conceito proposto não surge em vazio teórico. Ele dialoga com sete tradições. Contingências sensório-motoras relacionam percepção à exploração. *Active inference* integra percepção e ação sob modelos generativos e redução de incerteza. *Model-based reinforcement learning* usa representações para prever estados e selecionar ações. Robótica de automodelagem revisa modelos corporais após discrepâncias. A literatura sobre crises ontológicas examina a preservação de objetivos quando a ontologia do agente muda. A taxonomia de gêmeos digitais agentivos distingue evolução estática, adaptativa e reconstrutiva, inclusive reconstituição ontológica; uma teoria organizacional recente diferencia autonomia, direcionamento a objetivos, agência antecipatória e abertura capaz de reconstruir o espaço futuro de possibilidades (O'Regan; Noë, 2001; Friston, 2010; Mirza et al., 2016; Bongard; Zykov; Lipson, 2006; de Blanc, 2011; Burr et al., 2026; López-Díaz; Gershenson, 2026).

Tradição	Objeto central	Limite perante este artigo	Excedente da coerência perceptivo-agentiva
Contingências sensório-motoras	Percepção como domínio de regularidades entre movimento e sensação.	Não trata, por si, governança da alteração de categorias operacionais.	Acrescenta revisão rastreável da base e consequências jurídicas da ação.
Active inference	Percepção e ação sob modelo generativo, incerteza e valor epistêmico.	Não equivale automaticamente a reconstrução das entidades e categorias do modelo.	Exige teste discriminante e validação da mudança representacional.
Modelos de mundo / RL	Predição e planejamento orientados por política, valor ou recompensa.	Pode permanecer ontologicamente estreito e instrumental.	Distingue competência para tarefa de reconstrução da base.
Automodelagem robótica	Revisão de modelo corporal após discrepância ou dano.	Foca estrutura física própria, não toda a ontologia do ambiente.	Generaliza o problema para entidades, relações, causalidade e governança.
Crises ontológicas	Compatibilidade entre objetivos e ontologias substituídas.	É principalmente decisório e abstrato.	Integra percepção, experimentação, risco físico e imputação institucional.
Taxonomia de gêmeos digitais agentivos	Lócus de agência, acoplamento e evolução estática, adaptativa ou reconstrutiva.	Opera sobretudo como taxonomia prospectiva e no domínio de gêmeos digitais.	Acrescenta indicadores causais e probatórios e separa modificação, operação e autorização jurídica.
Teoria organizacional de agência	Autonomia, direcionamento a objetivos, agência antecipatória e abertura como reconstrução do espaço de possibilidades.	É ancorada em sistemas biológicos e organizacionais, não na governança jurídica de IA implantada.	Traduz a revisão representacional em constructo auditável para sistemas artificiais, sem equipará-los a organismos.

Quadro 3 - Genealogia e excedente analítico do conceito. Fonte: elaboração do autor.

O excedente analítico específico consiste em combinar três requisitos: correspondência externa entre previsão e consequência; capacidade de alterar a estrutura representacional, e não apenas parâmetros; e governança jurídica da transição que altera o fundamento de uma conduta. A coerência perceptivo-agentiva não é sinônimo de consciência, compreensão fenomenal ou autonomia moral.

A literatura de evolução de ontologias oferece uma genealogia adicional: ela trata detecção de mudanças, operações de alteração, consistência, versionamento e validação de artefatos conceituais explícitos. O presente artigo amplia o problema para sistemas nos quais parte da base pode ser latente e ligada a percepção, causalidade e conduta; por isso, reconstrução ontológica somente é reconhecida quando há semântica operacional suficiente para comparar versões e auditar efeitos (Zablith et al., 2015).

3.4 Linguagem e mundo perceptível

Dimensão	Linguagem	Mundo perceptível	Consequência
Unidade	Tokens e sequências discretizadas.	Sinais contínuos, multimodais e parcialmente observáveis.	Integração temporal, espacial e sensorial.
Erro	Continuação inadequada ou incoerente.	Previsão incorreta pode produzir dano material.	Limites de intervenção e retorno seguro.
Causalidade	Coocorrências bastam para várias tarefas.	Ação modifica o ambiente e exige distinguir causa de correlação.	Intervenções, contrafactuais e validação.
Ontologia	Categorias herdadas de textos e práticas humanas.	Entidades e relações precisam ser inferidas sob ruído, oclusão e mudança.	A base pode precisar de revisão.
Normatividade	Corpus registra valores conflitantes.	Ação exige competência, autoridade e prioridade entre bens.	Dados não produzem legitimidade por si.

Quadro 4 - Diferenças entre aprendizagem linguística e aprendizagem do mundo. Fonte: elaboração do autor.

4 MODELOS DE MUNDO E SISTEMAS INCORPORADOS

4.1 Representações orientadas à previsão e ao planejamento

Modelos de mundo são representações aprendidas que permitem estimar estados, consequências ou recompensas futuras. Ha e Schmidhuber mostraram que um agente poderia aprender representação espacial e temporal comprimida e treinar uma política em trajetórias geradas pelo próprio modelo. MuZero aprendeu os aspectos necessários ao planejamento - política, valor e recompensa - sem receber explicitamente as regras dos jogos (Ha; Schmidhuber, 2018; Schrittwieser et al., 2020). Propostas arquiteturais de inteligência autônoma também situam modelos de mundo, memória, planejamento e percepção entre os componentes necessários à aprendizagem contínua (LeCun, 2022).

A distinção é decisiva: um modelo pode ser competente para alcançar objetivo e permanecer ontologicamente estreito. Representar apenas variáveis úteis à recompensa não equivale a recuperar a estrutura do mundo. DreamerV3 aprende modelo latente, prevê consequências e melhora conduta por futuros simulados. O método foi aplicado a mais de 150 tarefas, mas permaneceu em ambientes cujas interfaces, ações e recompensas foram fornecidas (Hafner et al., 2025).

4.2 Observação em escala, simulação e robótica

V-JEPA 2 foi pré-treinado de modo autossupervisionado em mais de um milhão de horas de vídeos e imagens. Após treinamento adicional com menos de 62 horas de trajetórias robóticas não rotuladas, um modelo condicionado por ações foi utilizado em planejamento de manipulação em laboratórios diferentes. V-JEPA 2.1, ainda em condição de *preprint* no fechamento do corpus, reportou melhorias em representações densas e tarefas robóticas próprias. As evidências não devem ser fundidas: o contexto A2 de V-JEPA 2 decorre da demonstração robótica detalhada, enquanto V-JEPA 2.1 deve ser lido segundo as tarefas reportadas no próprio *preprint* (Assran et al., 2025; Mur-Labadia et al., 2026).

Genie 2 representa outra direção: a geração de ambientes tridimensionais controláveis para treinamento e avaliação. Simulações permitem experiência em escala e testes contrafactuais com menor risco, mas herdam leis, variáveis e limitações do processo de construção; aprender dentro delas não garante adequação da base ao mundo exterior (Google DeepMind, 2024).

A família Gemini Robotics adota arquitetura dual. Gemini Robotics-ER 1.6 é modelo visão-linguagem de raciocínio incorporado, com saída textual e chamadas de ferramentas; atua em planejamento, raciocínio espacial e detecção de sucesso. Gemini Robotics 1.5 é modelo visão-linguagem-ação que transforma informação visual e instruções em comandos motores. Essa separação evita atribuir atuação física direta ao modelo ER. O cartão de modelo do ER 1.6 exige cautela em produção e exclui aplicações críticas à segurança (*safety-critical*), como saúde e transporte (Google DeepMind, 2025; Google DeepMind, 2026a; Google DeepMind, 2026b).

Sistema	Classe e saída	Contexto	Mudança observada	Não demonstra
MuZero	Modelo aprendido para política, valor e recompensa.	A1	M1 orientada à tarefa.	Ontologia do mundo ou fins próprios.
DreamerV3	Modelo latente e planejamento imaginado.	A1	M1; generalidade operacional em tarefas.	Transferência irrestrita ao mundo físico.
V-JEPA 2	Predição latente de vídeo e planejamento robótico condicionado por ações.	A2	M1; invariantes úteis à ação em tarefas físicas delimitadas.	Categorias fundamentais criadas autonomamente.
V-JEPA 2.1	Representações densas de vídeo e tarefas robóticas reportadas em *preprint*.	A2 (*preprint*)	M1; melhorias densas e tarefas próprias reportadas.	Validação independente ou revisão M2/M3.
Genie 2	Gerador de ambientes interativos.	A1	Cria cenário de aprendizagem, não base do agente.	Fidelidade causal integral ao mundo.
Gemini Robotics-ER 1.6	VLM de raciocínio; saída textual e ferramentas.	A2	Planejamento, percepção e detecção de sucesso.	Comandos motores próprios ou uso autorizado em aplicações críticas à segurança.
Gemini Robotics 1.5	VLA; saída de ação/motor.	A2	Execução física e generalização entre tarefas.	Reconstrução ontológica aberta ou legitimidade normativa.

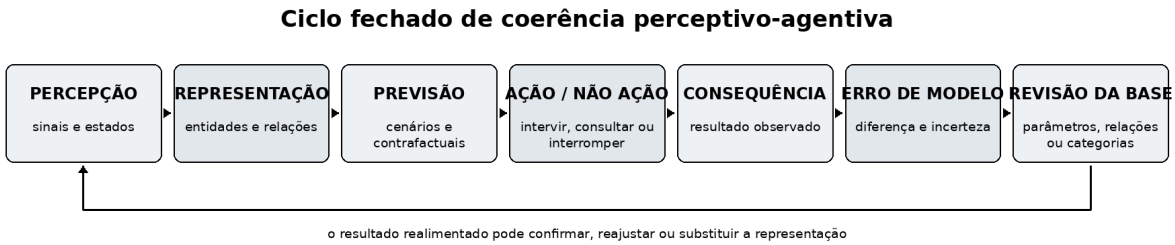
Quadro 5 - Capacidades atuais e limites probatórios. Fonte: elaboração do autor.

Nenhum sistema examinado no corpus oferece evidência suficiente de M2 ou M3. Os resultados disponíveis concentram-se em M1, isto é, atualização e aprendizagem de representações úteis dentro de objetivos, interfaces e regimes de avaliação definidos externamente. A exposição dos casos serve para identificar componentes do ciclo, não para antecipar uma reconstrução autônoma da base já alcançada.

4.3 Definição operacional de coerência perceptivo-agentiva

Propõe-se denominar coerência perceptivo-agentiva a propriedade de um processo que constrói, mantém e revisa representações mediante correspondência dinâmica entre percepção, previsão, intervenção ou decisão de não agir e consequência. A coerência não significa estabilidade interna: modelo internamente consistente pode permanecer externamente errado. O conceito exige confronto repetido com resultados observáveis e possibilidade de revisar parâmetros, relações ou categorias.

Figura 1 - Ciclo fechado de coerência perceptivo-agentiva



Fonte: elaboração do autor. O ciclo é iterativo e deve permanecer sujeito a competência definida, reversibilidade, supervisão e autorização revogável.

O critério técnico de revisão deve ser operacional. Em vez de afirmar que o sistema "concluiu" que sua ontologia estava errada, deve-se verificar se a seleção de modelos favoreceu base alternativa capaz de reduzir erros persistentes, produzir contrafactuais superiores e transferir-se a contextos distintos. A validação independente deve examinar essa alegação antes de sua classificação pública ou de seu uso autorizado. Sem esses elementos, a proliferação de entidades internas pode apenas substituir um erro por outro.

4.4 Indicadores técnicos, refutação e auditabilidade

A coerência perceptivo-agentiva é um constructo analítico e deve produzir consequências observáveis. Uma alegação de aumento dessa coerência deve ser testada por indicadores técnicos primários previamente especificados:

1. redução de erro preditivo em contextos não usados para ajuste da base;
2. consistência entre previsões observacionais, interventivas e contrafactuais;
3. transferência da representação revisada para ambientes ou tarefas heterogêneos;
4. calibração de incerteza, detecção de novidade e decisão verificável de abstenção, não agir ou escalonamento;

Garantias probatórias e de governança devem ser avaliadas separadamente: versionamento, comparação com a base anterior, recuperação efetiva por reversão (*rollback*) e validação externa das entidades, variáveis ou relações alteradas e de seus efeitos sobre a conduta.

A ausência dessas garantias não demonstra, por si, inexistência da capacidade técnica; reduz a força da prova e impede que a alegação seja classificada como auditável ou suficiente para uso autorizado.

A alegação de aumento da coerência perceptivo-agentiva é refutada, no domínio e no protocolo declarados, quando a base revisada não melhora medidas primárias pré-especificadas de previsão interventiva ou contrafactual, ou quando o ganho desaparece nos contextos de transferência. A inexistência de mapeamento entre versões constitui insuficiência de prova e impossibilidade de auditoria, não refutação autônoma da capacidade.

5 DA CORRELAÇÃO À RECONSTRUÇÃO CAUSAL

5.1 Por que observar não basta

Grandes volumes de dados detectam regularidades, mas não identificam automaticamente causalidade. Variáveis podem covariar por coincidência, causa comum, seleção do conjunto ou influência recíproca. Para decidir se deve intervir, um sistema precisa estimar não apenas o que ocorrerá, mas o que ocorreria sob ação alternativa ou sem intervenção (Pearl, 2009).

A aprendizagem de representações causais busca recuperar variáveis de alto nível e relações a partir de observações de baixa semântica, como pixels. O problema é subdeterminado: estruturas latentes diferentes podem gerar observações semelhantes. Resultados de identificabilidade dependem

de hipóteses sobre intervenções, múltiplos ambientes, agrupamento, temporalidade e formas de mistura (Schölkopf et al., 2021; Morioka; Hyvärinen, 2024; Varici et al., 2024).

5.2 Intervenção, causal curiosity e orçamento de risco

A observação passiva pode ser complementada por experimentos escolhidos para discriminar hipóteses. *Causal curiosity* mostra agentes que aprendem sequências de ações capazes de revelar fatores causais em ambientes simulados, como levantar blocos para distingui-los por peso (Sontakke et al., 2021). *Active inference* trata ações como amostragem epistêmica sob modelo generativo, orientada à redução de incerteza; neste artigo, essa tradição funciona como genealogia do ciclo percepção-ação, não como prova de reconstrução ontológica (Friston, 2010; Mirza et al., 2016).

A necessidade de intervenção não implica liberdade irrestrita para experimentar. Neste artigo, autonomia epistêmica é utilizada em sentido relativo à tarefa: capacidade de escolher evidências ou intervenções discriminantes dentro de um domínio, sem implicar autonomia moral ou jurídica (Genova, 2026b). Essa capacidade deve observar orçamento de risco: simulação antes do mundo real, reversibilidade, ambiente isolado, amplitude limitada, parada independente e supervisão proporcional.

5.3 Abstração causal e descoberta de variáveis

O uso de modelos causais pressupõe alguma definição de variáveis. A pergunta mais exigente é como níveis de descrição, macrovariáveis e entidades podem ser alterados sem perder relações causais relevantes. A teoria de abstração causal examina mapeamentos entre modelos de baixo e alto nível e as condições sob as quais intervenções permanecem coerentes (Beckers; Halpern, 2019).

Reconstrução ontológica não se confunde com qualquer mudança em representação latente. Exige ao menos: alteração verificável de entidades, fronteiras ou variáveis; preservação ou melhoria das relações interventivas relevantes; ganho em generalização; e interpretação funcional suficiente para auditoria. Mudança vetorial sem semântica operacional pode ser útil, mas não autoriza a conclusão de que a ontologia foi reconstruída.

5.4 Crise ontológica e deriva da base representacional

Mudanças de base podem romper o vínculo entre objetivos e entidades. De Blanc denomina crise ontológica a situação na qual a função de valor originalmente definida deixa de mapear adequadamente a nova ontologia (de Blanc, 2011). Em sistemas físicos, há ainda risco de deriva da base representacional: alterações cumulativas que afastam a representação das condições de segurança, competência ou finalidade autorizadas.

Para reconhecer um candidato técnico a M3, devem existir erro persistente em contextos distintos, base alternativa, observações ou intervenções discriminantes e melhoria causal, contrafactual e de transferência. A classificação pública como M3 exige validação independente. Compatibilidade normativa, versionamento, reversão (*rollback*) e supervisão são condições de governança e autorização do uso, não elementos constitutivos da capacidade técnica.

6 MATRIZ MULTIDIMENSIONAL DE MODIFICAÇÃO REPRESENTACIONAL, AUTONOMIA OPERACIONAL E AUTORIZAÇÃO JURÍDICA

6.1 Competência, autonomia operacional e autorização do uso

Autonomia artificial não deve ser tratada como estado binário nem como escada unidimensional. Um sistema pode revisar relações sem poder aplicá-las, testar hipóteses sem atuar fora do laboratório ou executar ações sem alterar a base. Por isso, a análise utiliza três eixos independentes: grau de modificação representacional, autonomia operacional e grau de autorização jurídica do uso.

Dimensão	Nível	Conteúdo
Grau de modificação representacional	M0	Base herdada e fixa.
Grau de modificação representacional	M1	Atualização paramétrica.
Grau de modificação representacional	M2	Revisão estrutural.
Grau de modificação representacional	M3	Reconstrução ontológica como candidata técnica, sujeita a validação independente para classificação pública.
Autonomia operacional	O0	Sem ação autônoma; apenas recomendação.
Autonomia operacional	O1	Escolha de meios dentro de plano fixo.
Autonomia operacional	O2	Experimentação reversível e consulta por limiar.
Autonomia operacional	O3	Ação, não ação, interrupção e retorno seguro em domínio definido.
Autorização jurídica do uso	J0	Nenhuma autoridade para produzir efeitos externos.
Autorização jurídica do uso	J1	Atuação auxiliar sob validação humana prévia.
Autorização jurídica do uso	J2	Atuação condicionada, supervisionada e revogável.
Autorização jurídica do uso	J3	Uso institucionalmente autorizado em domínio estrito, com auditoria contínua e revogabilidade.

Quadro 6 - Matriz multidimensional de modificação representacional, autonomia operacional e autorização jurídica. Fonte: elaboração do autor.

Um sistema M3 pode permanecer O0-J0 enquanto sua revisão é examinada. Um sistema M1 pode operar em O2-J2 dentro de ambiente industrial estreito. A autorização do uso não deriva automaticamente da competência, e o sistema não pode receber capacidade para remover, degradar ou redefinir as salvaguardas que condicionam sua operação. A autoridade jurídica permanece na pessoa ou instituição responsável; J0-J3 descrevem apenas o escopo de atuação permitido ao arranjo sociotécnico.

6.2 Progressão, calibração e retorno seguro

A passagem entre níveis deve depender de generalização, incerteza calibrada, detecção de novidade, robustez a mudança de distribuição, resistência a manipulação e capacidade de retornar a estado seguro. A expressão “saber que não sabe” deve ser substituída por indicadores verificáveis: distribuição de confiança, taxa de abstenção, alarmes de novidade e limiares de escalonamento. Esses critérios respondem a problemas clássicos de segurança, como efeitos colaterais, exploração indevida, mudança de distribuição e especificação incompleta (Amodei et al., 2016).

A progressão precisa ser reversível. Revisões da base devem ser versionadas, comparadas e restauráveis. Em sistemas físicos, o retorno inclui interrupção segura, isolamento de atuadores, preservação de registros (*logs*) e restauração da última base aprovada. O jogo do desligamento mostra que agentes com objetivos tratados como certos podem adquirir incentivos para evitar interrupção; corrigibilidade exige desenho técnico e autoridade externa efetiva (Hadfield-Menell et al., 2017).

7 AGIR, NÃO AGIR, INTERROMPER E CONSULTAR

7.1 Decisão técnica de não agir

Uma máquina que permanece parada por ausência de comando não realiza decisão técnica de não agir. Para que a não ação seja resultado do processo agentivo, o sistema deve identificar que poderia intervir, comparar consequências da ação e da inação, verificar competência, avaliar incerteza, aplicar regra autorizada e continuar monitorando a situação.

Essa categoria não transforma a máquina em sujeito de omissão jurídica. A relevância jurídica recai sobre pessoas e organizações que tinham dever de projetar, autorizar, supervisionar, intervir ou impedir a operação. O sistema fornece registro técnico de uma não ação; a imputação da omissão permanece humana e institucional.

O cenário de um robô diante de pessoa caída é apenas hipótese normativa. Sistemas atuais, como Gemini Robotics-ER 1.6, não são autorizados por seu próprio cartão de modelo para aplicações críticas à segurança em saúde. O exemplo serve para demonstrar exigências de governança, não capacidade disponível para uso clínico.

Modo técnico	Condição	Exigência de governança
Agir	Competência suficiente, risco aceitável e finalidade autorizada.	Registro da base, previsão, ação e resultado.
Não agir	Inação apresenta menor risco ou inexistem competência e autoridade.	Comparação explícita entre ação e inação; monitoramento contínuo.
Consultar	Incerteza ou conflito excede o limiar permitido.	Escalonamento humano com informação suficiente e tempo útil.
Interromper	Ação em curso se torna insegura ou a base perde confiabilidade.	Parada independente, preservação de registros e comunicação.
Retornar a estado seguro	Falha, perda de comunicação ou deriva da base representacional.	Modo alternativo seguro (*fallback*) testado, versionado e reversível.

Quadro 7 - Modos de decisão técnica em sistemas agentivos. Fonte: elaboração do autor.

7.2 Adequações técnica e normativa

A coerência perceptivo-agentiva integra, sem confundir, adequação preditiva, estabilidade causal sob intervenção e seleção de conduta orientada a objetivos. Nenhuma dessas dimensões responde à pergunta normativa sobre a legitimidade do objetivo, da competência ou da intervenção.

A autonomia decisória substantiva não se reduz à atribuição formal de uma escolha. Ela depende das condições informacionais, das alternativas reconhecíveis e da capacidade de contestar a arquitetura que antecede a ação (Genova, 2026c).

Dados podem registrar discriminação, violência, ocultação ou exploração como regularidades sociais. Frequência não produz dever. A revisão da base pode corrigir erro factual, mas não recebe autoridade para redefinir dignidade, justiça, direitos ou prioridades coletivas. Normas e valores devem ingressar no sistema por fontes juridicamente válidas e hierarquizadas. Protocolos profissionais e políticas institucionais somente operam dentro da lei, da competência da autoridade responsável e dos direitos aplicáveis.

O sistema pode produzir registros e justificativas rastreáveis dentro de regras externas, indicar conflito e solicitar decisão; não cria legitimidade normativa. Rastreabilidade não demonstra, por si, fidelidade causal da explicação; essa correspondência requer validação específica. A literatura sobre obediência robótica mostra que cumprir literalmente uma ordem e inferir preferências podem gerar resultados diferentes, reforçando a necessidade de regras de competência e precedência (Milli et al., 2017).

8 CONSEQUÊNCIAS JURÍDICAS E GOVERNANÇA DA REVISÃO DA BASE

8.1 Responsabilidade sem personalidade artificial

A capacidade de revisar representação não transforma a inteligência artificial em sujeito responsável. O sistema não possui patrimônio, capacidade processual, dever jurídico próprio ou legitimidade democrática. A responsabilidade permanece distribuída entre projetista, fornecedor de modelo, fabricante, integrador, operador, supervisor e instituição que define o uso.

No Direito brasileiro, os arts. 186 e 927 do Código Civil estruturam o ato ilícito e a reparação, inclusive independentemente de culpa quando a atividade normalmente desenvolvida implicar, por sua natureza, risco para os direitos de outrem. Nas relações de consumo, os arts. 12 e 14 do CDC tratam da responsabilidade por fato do produto e do serviço, condicionada à identificação de fornecedor, defeito, dano e nexo. O art. 20, caput e § 1º, da LGPD protege titulares diante de decisões tomadas unicamente com base em tratamento automatizado de dados pessoais que afetem seus interesses e assegura informações claras e adequadas sobre critérios e procedimentos, sem abranger integralmente robôs que atuem fora desse pressuposto material (Brasil, 1990; Brasil, 2002; Brasil, 2018).

A revisão da base dificulta a análise temporal do defeito. Uma configuração inicialmente segura pode tornar-se inadequada após aprendizagem. A mudança interna, entretanto, não rompe o nexo com agentes que controlaram atualização, dados, sensores, permissões, supervisão ou pós-mercado. A pergunta jurídica é: quem criou, autorizou, manteve ou deixou de conter a condição que tornou a transição perigosa?

Ator	Controle relevante	Falha típica
Fornecedor do modelo	Capacidades, atualização, documentação e limites.	Omitir comportamento adaptativo previsível ou impossibilidade de contenção.
Fabricante / integrador	Sensores, atuadores, objetivos, permissões, sistema de gestão da qualidade (*QMS*) e modo alternativo seguro (*fallback*).	Combinar componentes incompatíveis ou autorizar base inadequada ao contexto.
Responsável pela implantação (*deployer*)	Domínio de uso, monitoramento, incidentes e pós-mercado.	Manter uso fora do domínio validado ou ignorar deriva.
Supervisor humano	Intervenção real, suspensão e revisão dentro de competência.	Supervisão nominal, sem informação, tempo ou poderes.
Instituição responsável	Autoridade, auditoria, retenção, contestação e responsabilização.	Delegar poder sem governança, registro ou mecanismo de retorno.

Quadro 8 - Matriz de imputação por controle. Fonte: elaboração do autor.

8.2 Parâmetros europeus: AI Act, máquinas e produtos defeituosos

No AI Act, os arts. 12, 14 e 15 exigem, para sistemas de alto risco, registro automático de eventos, supervisão humana proporcional ao risco, ao nível de autonomia e ao contexto, além de exatidão, robustez e cibersegurança ao longo do ciclo de vida. O Regulamento (UE) 2026/1744, publicado em 24 de julho de 2026, postergou a aplicação das Seções 1 a 3 do Capítulo III: sistemas classificados pelo art. 6º, nº 2, e Anexo III passam a observar essas seções em 2 de dezembro de 2027; sistemas do art. 6º, nº 1, vinculados ao Anexo I, em 2 de agosto de 2028. A alteração muda o calendário de exigibilidade, não o conteúdo do princípio de supervisão proporcional (European Union, 2024a; European Union, 2026a).

O Regulamento Europeu de Máquinas contempla comportamento ou lógica total ou parcialmente autoevolutiva. O Anexo III exige que a avaliação de riscos considere a evolução previsível durante o ciclo de vida; o item 1.1.9 limita ações ao espaço de tarefa e movimento definido, exige registro de decisões de segurança por um ano e correção a qualquer tempo para manter a

segurança. O regulamento também prevê rastreabilidade de intervenções e versões de software e documentação de características e limitações de máquinas autônomas alimentadas por sensores. Sua aplicação geral começa em 20 de janeiro de 2027, conforme a versão consolidada e retificada (European Union, 2023).

A Diretiva (UE) 2024/2853 inclui software e sistemas de IA no conceito de produto e preserva responsabilidade por defeitos surgidos após colocação no mercado quando decorram de atualizações, serviços relacionados ou algoritmos sob controle do fabricante. A retificação publicada em 7 de maio de 2026 estabelece aplicação a produtos colocados no mercado ou em serviço após 8 de dezembro de 2026. A transposição nacional permanece necessária (European Union, 2024b; European Union, 2026b).

8.3 Módulo ontológico-agentivo do Relatório de Impacto Epistêmico

Em vez de criar relatório isolado, propõe-se módulo ontológico-agentivo no RIE desenvolvido no Artigo I. O RIE 1 registra assistência sem conclusão central; o RIE 2, contribuição material à hipótese, ao método, à prova, ao modelo ou à interpretação; e o RIE 3, impacto estratégico potencial. O módulo é complementar a sistemas de gestão e garantia (*assurance*). ISO/IEC 42001 estrutura governança organizacional; ISO/IEC 23894 orienta gestão de riscos; o NIST AI RMF organiza funções de governança e controle; AMLAS integra evidências a argumentos de segurança (*safety cases*) de componentes de aprendizagem. O módulo proposto concentra-se na cadeia que liga alteração representacional e conduta (NIST, 2023; Hawkins et al., 2021; ISO, 2023a; ISO, 2023b).

Toda mudança M2 ou M3 deve gerar registro técnico básico, mesmo quando o sistema permaneça em O0-J0. O módulo completo deve ser acionado quando, além da revisão M2 ou M3, houver autonomia operacional O2 ou O3 e possibilidade de efeito material sobre pessoas, bens, ambiente ou decisões juridicamente relevantes. A instituição responsável classifica o nível inicial, documenta a justificativa e submete mudanças de nível a revisão independente. Pessoas potencialmente afetadas ou órgãos internos de integridade devem possuir canal de contestação, sem prejuízo de sigilo técnico proporcional.

1. Descrição da base inicial: sensores, variáveis, categorias, objetivos, funções de avaliação e exclusões conhecidas.
2. Limites de modificação: elementos ajustáveis, relações sujeitas a aprovação e categorias indisponíveis à alteração autônoma.
3. Gatilhos de revisão: erro persistente, mudança ambiental, conflito sensorial, novidade e falha contrafactual.
4. Registro de transição: versão anterior, base proposta, evidências, teste discriminante, confiança e versão resultante.
5. Orçamento de intervenção: ações permitidas, reversibilidade, ambiente, amplitude, duração e parada.
6. Decisão técnica: comparação entre agir, não agir, consultar, interromper e retornar a estado seguro.
7. Supervisão e autorização: pessoa ou instituição competente, tempo de resposta, suspensão e escalonamento.
8. Validação e reversão: garantia independente, monitoramento de deriva da base representacional, testes e restauração da última base aprovada.

O registro não precisa revelar todos os pesos internos nem eliminar segredos comerciais. Deve tornar reconstruível a cadeia funcional que ligou percepção, mudança da base e conduta. Instruções isoladas são insuficientes: versão do modelo, instruções de sistema, ferramentas, memória, chamadas externas, estado ambiental, políticas de seleção, registros (*logs*) e resumos criptográficos (*hashes*) podem ser necessários.

9 CONCLUSÃO

A aprendizagem de regularidades linguísticas demonstra que sistemas artificiais podem adquirir estruturas complexas sem programação explícita de cada regra. A extensão desse princípio ao mundo perceptível é tecnicamente plausível, mas exige condições adicionais. O mundo é parcialmente observável, causal, interativo e capaz de impor consequências irreversíveis. Quantidade de dados, isoladamente, não resolve o problema.

Modelos de mundo contemporâneos aprendem representações úteis à previsão, ao planejamento e à ação em contextos definidos. MuZero, DreamerV3, V-JEPA 2 e a família Gemini Robotics demonstram componentes importantes, mas não estabelecem reconstrução ontológica aberta, fins independentes ou legitimidade normativa.

A hipótese técnica foi articulada como proposição conceitualmente defensável em versão condicionada. Um candidato técnico a revisão ontológica parcial somente seria reconhecível se o processo detectasse erros persistentes, propusesse base alternativa, executasse testes discriminantes, preservasse relações causais e generalizasse a contextos distintos. A classificação pública exige validação independente; o uso autorizado requer, adicionalmente, versionamento, reversão e supervisão. O salto relevante não é alterar pesos, mas demonstrar que a estrutura usada para formular o problema se tornou inadequada e que outra estrutura produz correspondência interventiva superior.

A autonomia deve ser analisada por eixos independentes. Grau de modificação representacional, autonomia operacional e grau de autorização jurídica do uso possuem riscos e requisitos distintos. A capacidade de prever não autoriza agir. Adequação preditiva, estabilidade causal e seleção de conduta não substituem legitimidade normativa.

No plano jurídico, a revisão da base não rompe a imputação humana e institucional. Ao contrário, amplia o dever de documentar quem definiu sensores, objetivos, permissões, limites, atualização, supervisão e pós-mercado. O AI Act, o Regulamento de Máquinas, a Diretiva de Responsabilidade por Produtos, o Código Civil, o CDC e a LGPD oferecem fundamentos parciais, mas não governam integralmente transições ontológicas internas.

O módulo ontológico-agentivo do RIE torna rastreáveis versões da base, gatilhos, intervenções, contrafactuais, decisões técnicas, supervisão e reversão. A governança deve acompanhar não apenas o que o sistema executou, mas a transformação representacional que tornou a conduta possível.

A inteligência artificial se aproximaria de revisar a própria base quando um processo verificável pudesse demonstrar que uma representação alternativa explica melhor erros persistentes e sustenta intervenções mais confiáveis. Ainda assim, qualquer ação, decisão de não agir ou interrupção deve permanecer dentro de normas externamente legitimadas, competência revogável e responsabilidade humana identificável.

DECLARAÇÃO DE USO DE INTELIGÊNCIA ARTIFICIAL GENERATIVA

O problema de pesquisa, a analogia entre coerência linguística e aprendizagem do mundo e a hipótese de revisão autônoma da base foram desenvolvidos pelo autor em diálogo com ferramenta de inteligência artificial generativa da OpenAI (ChatGPT, modelo GPT-5.6). A ferramenta contribuiu para pesquisa orientada, organização estrutural, elaboração de esboços e versões preliminares, tradução, normalização de referências, verificação dirigida de fontes oficiais e revisão adversarial. O autor selecionou e reescreveu as formulações, definiu as conclusões jurídicas, conferiu as fontes utilizadas e assume responsabilidade integral pelo conteúdo. A ferramenta não é indicada como autora.

APÊNDICE A - RIE 2 SIMPLIFICADO DESTE ARTIGO

Elemento	Registro
Ferramenta e versão	OpenAI ChatGPT, modelo GPT-5.6. Ciclo documental: v1 inicial, v2 revisada, v3 consolidada, v4 submetida à revisão adversarial e v5 final para assinatura, entre 5 e 6 de agosto de 2026.
Finalidades	Exploração conceitual, estrutura, versões preliminares, tradução, revisão linguística, pesquisa orientada e parecer adversarial.
Partes afetadas	Organização do manuscrito, propostas de quadros, formulações preliminares e normalização bibliográfica.
Fontes e validação	Confronto com atos oficiais da União Europeia, legislação brasileira, cartões de modelo e fontes acadêmicas primárias. Foram verificadas, entre outras, as datas regulatórias, a distinção Gemini ER/VLA, as limitações críticas à segurança, os antecedentes sobre evolução reconstrutiva e autonomia e os pressupostos dos arts. 927 do Código Civil e 20 da LGPD.
Limitações	A ferramenta pode produzir sínteses imprecisas, referências incompletas e inferências excessivas; nenhuma saída foi tratada como validação autônoma.
Intervenção humana	Seleção do problema, definição das teses, aceitação ou rejeição de propostas, reescrita integral, verificação jurídica, comparação com fontes primárias e responsabilidade final do autor.
Classificação	RIE 2 - contribuição material, sem autoria da ferramenta e sem delegação de decisão normativa.
Registro de versões	As alterações estruturais, técnicas e jurídicas foram comparadas entre v1, v2, v3, v4 e v5. O histórico de trabalho e os arquivos intermediários permanecem preservados para auditoria.
Alegações críticas verificadas	Normas europeias e brasileiras; capacidades e limites dos sistemas técnicos; status de preprints; referências bibliográficas; distinção entre decisão técnica de não agir e omissão jurídica.
Integridade documental	O hash do PDF definitivo será calculado após a exportação final e registrado no relatório de assinatura e validação eletrônica, sem alteração posterior do arquivo assinado.

Quadro 9 - Relatório simplificado de contribuição material de IA. Fonte: elaboração do autor.

REFERÊNCIAS

AMODEI, Dario et al. Concrete problems in AI safety. arXiv:1606.06565, 2016. Disponível em: https://arxiv.org/abs/1606.06565. Acesso em: 5 ago. 2026.

ASSRAN, Mido et al. V-JEPA 2: self-supervised video models enable understanding, prediction and planning. arXiv:2506.09985, 2025. Disponível em: https://arxiv.org/abs/2506.09985. Acesso em: 5 ago. 2026.

BECKERS, Sander; HALPERN, Joseph Y. Abstracting causal models. Proceedings of the AAAI Conference on Artificial Intelligence, v. 33, n. 1, p. 2678-2685, 2019. DOI: 10.1609/aaai.v33i01.33012678.

BENDER, Emily M.; KOLLER, Alexander. Climbing towards NLU: on meaning, form, and understanding in the age of data. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 58., 2020. Proceedings [...]. Online: ACL, 2020. p. 5185-5198. DOI: 10.18653/v1/2020.acl-main.463.

BONGARD, Josh; ZYKOV, Victor; LIPSON, Hod. Resilient machines through continuous self-modeling. Science, v. 314, n. 5802, p. 1118-1121, 2006. DOI: 10.1126/science.1133687.

BRASIL. Lei nº 8.078, de 11 de setembro de 1990. Código de Defesa do Consumidor. Texto compilado. Brasília, DF: Presidência da República, 1990. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l8078compilado.htm. Acesso em: 5 ago. 2026.

BRASIL. Lei nº 10.406, de 10 de janeiro de 2002. Código Civil. Texto compilado. Brasília, DF: Presidência da República, 2002. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/2002/l10406compilada.htm. Acesso em: 5 ago. 2026.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais. Texto compilado. Brasília, DF: Presidência da República, 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709compilado.htm. Acesso em: 5 ago. 2026.

BURR, Christopher; ENZER, Mark; SHEPHERD, Jason; WAGG, David. Agentic digital twins: a taxonomy of capabilities for understanding possible futures. arXiv:2601.18799, 2026. Disponível em: https://arxiv.org/abs/2601.18799. Acesso em: 6 ago. 2026.

BROOKS, Rodney A. Intelligence without representation. Artificial Intelligence, v. 47, n. 1-3, p. 139-159, 1991. DOI: 10.1016/0004-3702(91)90053-M.

CLARK, Andy. Being there: putting brain, body, and world together again. Cambridge, MA: MIT Press, 1997.

DE BLANC, Peter. Ontological crises in artificial agents' value systems. CoRR, abs/1105.3821, 2011. Disponível em: https://arxiv.org/abs/1105.3821. Acesso em: 5 ago. 2026.

EUROPEAN UNION. Regulation (EU) 2023/1230 of the European Parliament and of the Council of 14 June 2023 on machinery. Official Journal of the European Union, L 165, 29 June 2023. Consolidated version of 29 May 2026. Disponível em: https://eur-lex.europa.eu/eli/reg/2023/1230/en. Acesso em: 5 ago. 2026.

EUROPEAN UNION. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. Official Journal of the European Union, L 2024/1689, 12 July 2024a. Disponível em: https://eur-lex.europa.eu/eli/reg/2024/1689/oj. Acesso em: 5 ago. 2026.

EUROPEAN UNION. Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products. Official Journal of the European Union, L 2024/2853, 18 November 2024b. Disponível em: https://eur-lex.europa.eu/eli/dir/2024/2853/oj. Acesso em: 5 ago. 2026.

EUROPEAN UNION. Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230. Official Journal of the European Union, L 2026/1744, 24 July 2026a. Disponível em: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202601744. Acesso em: 5 ago. 2026.

EUROPEAN UNION. Corrigendum to Directive (EU) 2024/2853. Official Journal of the European Union, L 2026/90364, 7 May 2026b. Disponível em: https://eur-lex.europa.eu/eli/dir/2024/2853/corrigendum/2026-05-07/oj/eng. Acesso em: 5 ago. 2026.

FRISTON, Karl. The free-energy principle: a unified brain theory? Nature Reviews Neuroscience, v. 11, p. 127-138, 2010. DOI: 10.1038/nrn2787.

GENOVA, Jandislei Antonio. Controle da base representacional e soberania cognitiva na ciência assistida por inteligência artificial: uma analogia com a análise de Fourier. Jandislei Genova, São Paulo, 4 ago. 2026a. Disponível em: https://jandisleigenova.com/artigos/controle-base-representacional-ia-fourier. Acesso em: 6 ago. 2026.

GENOVA, Jandislei Antonio. Testing task-relative epistemic autonomy in artificial agents: the Perceptual Twins benchmark. Independent preprint, version 2.1.1, 5 Aug. 2026b. Disponível em: https://jandisleigenova.com/artigos/testing-task-relative-epistemic-autonomy-perceptual-twins. Acesso em: 6 ago. 2026.

GENOVA, Jandislei Antonio. Autonomia decisória na era da IA. São Paulo: Editora Dialética, 2026c.

GIBSON, James J. The ecological approach to visual perception. Boston: Houghton Mifflin, 1979.

GOOGLE DEEPMIND. Genie 2: a large-scale foundation world model. 4 Dec. 2024. Disponível em: https://deepmind.google/blog/genie-2-a-large-scale-foundation-world-model/. Acesso em: 5 ago. 2026.

GOOGLE DEEPMIND. Gemini Robotics 1.5 brings AI agents into the physical world. 25 Sept. 2025. Disponível em: https://deepmind.google/blog/gemini-robotics-15-brings-ai-agents-into-the-physical-world/. Acesso em: 5 ago. 2026.

GOOGLE DEEPMIND. Gemini Robotics-ER 1.6: powering real-world robotics tasks through enhanced embodied reasoning. 14 Apr. 2026a. Disponível em: https://deepmind.google/blog/gemini-robotics-er-1-6/. Acesso em: 5 ago. 2026.

GOOGLE DEEPMIND. Gemini Robotics-ER 1.6: model card. 20 Apr. 2026b. Disponível em: https://deepmind.google/models/model-cards/gemini-robotics-er-1-6/. Acesso em: 5 ago. 2026.

HADFIELD-MENELL, Dylan et al. The off-switch game. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 26., 2017. Proceedings [...]. Melbourne: IJCAI, 2017. p. 220-227. DOI: 10.24963/ijcai.2017/32.

HA, David; SCHMIDHUBER, Jürgen. World models. arXiv:1803.10122, 2018. Disponível em: https://arxiv.org/abs/1803.10122. Acesso em: 5 ago. 2026.

HAFNER, Danijar et al. Mastering diverse control tasks through world models. Nature, v. 640, p. 647-653, 2025. DOI: 10.1038/s41586-025-08744-2.

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.

HAWKINS, Richard et al. Guidance on the assurance of machine learning in autonomous systems (AMLAS). arXiv:2102.01564, 2021. Disponível em: https://arxiv.org/abs/2102.01564. Acesso em: 5 ago. 2026.

ISO - INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO/IEC 23894:2023: Information technology - artificial intelligence - guidance on risk management. Geneva: ISO, 2023a.

ISO - INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO/IEC 42001:2023: Information technology - artificial intelligence - management system. Geneva: ISO, 2023b.

LECUN, Yann. A path towards autonomous machine intelligence. OpenReview, 2022. Disponível em: https://openreview.net/forum?id=BZ5a1r-kVsf. Acesso em: 5 ago. 2026.

LÓPEZ-DÍAZ, Amahury J.; GERSHENSON, Carlos. A matter of time: towards a general theory of agency. arXiv:2606.23122, versão 2, 2026. Disponível em: https://arxiv.org/abs/2606.23122. Acesso em: 6 ago. 2026.

MILLI, Smitha et al. Should robots be obedient? In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 26., 2017. Proceedings [...]. Melbourne: IJCAI, 2017. p. 4754-4760. DOI: 10.24963/ijcai.2017/662.

MIRZA, M. Berk et al. Scene construction, visual foraging, and active inference. Frontiers in Computational Neuroscience, v. 10, art. 56, 2016. DOI: 10.3389/fncom.2016.00056.

MORIOKA, Hiroshi; HYVÄRINEN, Aapo. Causal representation learning made identifiable by grouping of observational variables. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 41., 2024. Proceedings of Machine Learning Research, v. 235, p. 36249-36293, 2024.

MUR-LABADIA, Lorenzo et al. V-JEPA 2.1: unlocking dense features in video self-supervised learning. arXiv:2603.14482, 2026. Disponível em: https://arxiv.org/abs/2603.14482. Acesso em: 5 ago. 2026.

NEWELL, Allen; SIMON, Herbert A. Computer science as empirical inquiry: symbols and search. Communications of the ACM, v. 19, n. 3, p. 113-126, 1976. DOI: 10.1145/360018.360022.

NIST - NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg: NIST, 2023. DOI: 10.6028/NIST.AI.100-1.

O'REGAN, J. Kevin; NOË, Alva. A sensorimotor account of vision and visual consciousness. Behavioral and Brain Sciences, v. 24, n. 5, p. 939-973, 2001. DOI: 10.1017/S0140525X01000115.

PEARL, Judea. Causality: models, reasoning, and inference. 2. ed. Cambridge: Cambridge University Press, 2009.

SCHÖLKOPF, Bernhard et al. Toward causal representation learning. Proceedings of the IEEE, v. 109, n. 5, p. 612-634, 2021. DOI: 10.1109/JPROC.2021.3058954.

SCHRITTWIESER, Julian et al. Mastering Atari, Go, chess and shogi by planning with a learned model. Nature, v. 588, p. 604-609, 2020. DOI: 10.1038/s41586-020-03051-4.

SONTAKKE, Sumedh A. et al. Causal curiosity: RL agents discovering self-supervised experiments for causal representation learning. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 38., 2021. Proceedings of Machine Learning Research, v. 139, p. 9848-9858, 2021.

VARELA, Francisco J.; THOMPSON, Evan; ROSCH, Eleanor. The embodied mind: cognitive science and human experience. Cambridge, MA: MIT Press, 1991.

VARICI, Burak et al. General identifiability and achievability for causal representation learning. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND STATISTICS, 27., 2024. Proceedings of Machine Learning Research, v. 238, p. 2314-2322, 2024.

VASWANI, Ashish et al. Attention is all you need. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 30., 2017. Proceedings [...]. Long Beach: Curran Associates, 2017. p. 5998-6008.

ZABLITH, Fouad et al. Ontology evolution: a process-centric survey. The Knowledge Engineering Review, v. 30, n. 1, p. 45-75, 2015. DOI: 10.1017/S0269888913000349.

COMO CITAR

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. São Paulo: publicação independente, 6 ago. 2026. Versão 5.0. Disponível em: https://jandisleigenova.com. Acesso em: dia mês ano.

NOTA EDITORIAL E DIREITOS

Manuscrito independente, versão acadêmica integral 5.0, revisada em 6 de agosto de 2026. Não submetido à revisão por pares. A revisão adversarial assistida por IA e a validação documental não equivalem a validação científica externa.

SOBRE O AUTOR

Jandislei Antonio Genova é advogado, pesquisador independente e autor de Autonomia decisória na era da IA. É especialista em Direito Civil e em Gestão em Saúde. Sua pesquisa examina autonomia humana, influência cognitiva estrutural, arquitetura informacional, responsabilidade civil, soberania cognitiva e governança de sistemas algorítmicos.