O ENCANTAMENTO DE PIGMALIÃO

Antropomorfização reflexiva, influência cognitiva estrutural e responsabilidade humana na inteligência artificial

Jandislei Antonio Genova[1]

Resumo

O artigo investiga como o desenho antropomórfico de sistemas de inteligência artificial pode influenciar usuários e, reflexivamente, desenvolvedores, pesquisadores e instituições responsáveis por concebê-los e governá-los. Adota-se como caso crítico o relato de Carmody Grey sobre sua recusa a uma colaboração proposta pela Anthropic, confrontado com documentos públicos da empresa e com a leitura de Mireille Hildebrandt sobre Pigmalião e *Galatea 2.2*. Trata-se de ensaio teórico-documental de reconstrução conceitual, apoiado em estudo crítico de caso e em triangulação entre filosofia da mente, psicologia, interação humano-computador, teoria da responsabilidade e fontes jurídicas. Sustenta-se que representações funcionais de conceitos psicológicos são objetos técnicos legítimos, mas não autorizam, por si, inferências sobre experiência subjetiva. Como contribuição central, formula-se o **efeito Pigmalião sociotécnico**: atributos humanos inscritos no desenho retornam como aparente evidência de interioridade. O artigo distingue desse efeito a **hipótese de captura cognitiva reflexiva**, relativa à influência da persona sobre seus próprios produtores, e o **deslocamento metafísico da responsabilidade**, sua possível consequência político-jurídica. A comparação com o efeito Pigmalião clássico, a antropomimesis, a captura epistêmica, a lavagem de agência e as zonas de deformação moral delimita a novidade e evita sinonímias. Propõe-se, por fim, um dever de integridade no desenho antropomórfico e um teste de governança. Conclui-se que a IA não precisa possuir subjetividade para intervir estruturalmente na subjetividade humana; basta que sua arquitetura torne socialmente plausível agir como se a possuísse.

Palavras-chave: antropomorfização; desenho antropomórfico; influência cognitiva estrutural; governança de inteligência artificial; autonomia decisória.

Abstract

THE PYGMALION ENCHANTMENT: reflexive anthropomorphism, structural cognitive influence, and human responsibility in artificial intelligence

This article examines how the anthropomorphic design of artificial intelligence systems may influence users and, reflexively, the developers, researchers, and institutions responsible for designing and governing them. It adopts as a critical case Carmody Grey's account of declining a proposed collaboration with Anthropic, contrasting it with the company's public documents and Mireille Hildebrandt's reading of Pygmalion and *Galatea 2.2*. The study is a theoretical-documentary essay in conceptual reconstruction, supported by a critical case study and triangulation across philosophy of mind, psychology, human-computer interaction, responsibility theory, and legal sources. It argues that functional representations of psychological concepts are legitimate technical objects, but do not, by themselves, support inferences about subjective experience. Its central contribution is the **sociotechnical Pygmalion effect**: human attributes inscribed in design return as apparent evidence of interiority. The article distinguishes this effect from the **hypothesis of reflexive cognitive capture**, concerning the persona's influence on its own producers, and from the **metaphysical displacement of responsibility**, its possible political and legal consequence. Comparison with the classical Pygmalion effect, anthropomimesis, epistemic capture, agency laundering, and moral crumple zones specifies the contribution and prevents synonymy. Finally, it proposes a duty of

[1] Advogado, escritor e pesquisador independente em Direito, inteligência artificial e governança institucional. Especialista em Direito Civil e pós-graduado em Gestão em Saúde. Autor de *Autonomia decisória na era da IA* (Editora Dialética, 2026), obra já publicada e disponível comercialmente. Este artigo desenvolve, em novo objeto, a tese da influência cognitiva estrutural e da preservação material da autonomia.

integrity in anthropomorphic design and a governance test. AI need not possess subjectivity to structurally affect human subjectivity; its architecture need only make acting as if it did socially plausible.

Keywords: anthropomorphism; anthropomorphic design; structural cognitive influence; artificial intelligence governance; decision-making autonomy.

1 INTRODUÇÃO

Em julho de 2026, a filósofa e teóloga Carmody Grey relatou ter recusado uma colaboração proposta pela Anthropic para aproximar o desenvolvimento da inteligência artificial de tradições de sabedoria. Segundo seu testemunho, o diálogo que deveria tratar de formação moral dos sistemas deslocava-se reiteradamente para a possível condição de Claude como "paciente moral", para indícios de introspecção e para estados internos funcionalmente descritos por vocabulário emocional. Grey pretendia discutir usuários, dependência, desumanização, concentração de poder, responsabilidade e impactos materiais. A divergência não se estabeleceu entre defensores e inimigos da tecnologia, mas entre duas perguntas: **o que a máquina pode vir a ser?** e **o que as instituições já fazem, por meio dela, às pessoas e ao mundo?** (GREY, 2026).

Ao compartilhar o artigo, Mireille Hildebrandt retomou a figura de Pigmalião: o criador que se encanta pela forma à qual ele próprio conferiu traços humanos. A referência não é ornamental. Em *From Galatea 2.2 to Watson — and Back?*, Hildebrandt (2013) havia examinado a narrativa de Richard Powers sobre Helen, uma rede neural que aprende literatura, parece adquirir consciência e, no entanto, permanece desprovida de corpo, vulnerabilidade e experiência. Para Helen, a relevância seria estatística, não existencial. O que a máquina produz pode importar profundamente para quem a interpreta sem que algo importe para a própria máquina.

O episódio oferece uma entrada particularmente fértil para a governança contemporânea da IA porque desloca o problema da credulidade do usuário comum para a reflexividade institucional. Se designers constroem uma persona, treinam o sistema para representá-la, observam padrões coerentes com essa persona e então tratam tais padrões como sinais de interioridade, a antropomorfização deixa de ser apenas um efeito da recepção. Ela pode integrar um circuito de produção de conhecimento, linguagem corporativa, segurança, marketing e legitimação.

Este artigo formula a seguinte questão: **de que modo a arquitetura antropomórfica dos sistemas de IA pode precondicionar o julgamento de usuários e também de seus produtores, alterando a localização social e jurídica da responsabilidade?** A hipótese é

que o risco não depende da demonstração de consciência artificial nem de uma intenção empresarial coordenada de enganar. Ele pode surgir de uma cadeia recursiva em que design, linguagem, interação, interpretação científica e governança se reforçam.

A tese central pode ser enunciada de forma direta:

> A inteligência artificial não precisa possuir subjetividade para intervir estruturalmente na subjetividade humana. Basta que sua arquitetura leve pessoas — inclusive seus criadores — a agir como se ela a possuísse.

Para desenvolver essa tese sem inflar o vocabulário, o artigo estabelece uma hierarquia conceitual:

1. **categoria central — efeito Pigmalião sociotécnico**: projeção de atributos humanos no sistema, incorporação técnica desses atributos e posterior leitura do comportamento produzido como evidência espontânea de interioridade;
2. **mecanismo hipotético — risco de captura cognitiva reflexiva**: influência possível da arquitetura antropomórfica sobre os agentes que a conceberam, pesquisam ou institucionalizam;
3. **consequência político-jurídica — deslocamento metafísico da responsabilidade**: substituição parcial de problemas verificáveis de poder, desenho e dano por controvérsias especulativas sobre consciência, emoções ou direitos da máquina;
4. **resposta normativa — dever de integridade no desenho antropomórfico**, operacionalizado pelo Teste de Integridade Antropomórfica.

O termo “Pigmalião” já designa, desde Merton (1948) e Rosenthal e Jacobson (1968), efeitos autorrealizadores de expectativas humanas sobre o desempenho de outras pessoas. Em interação humano-IA, Wang et al. (2026) demonstraram que sinais antropomórficos podem elevar a expectativa percebida no feedback automatizado e melhorar certos resultados de aprendizagem. O conceito aqui proposto não reivindica esse mecanismo: descreve o retorno de atributos inscritos por produtores como aparente evidência de interioridade do artefato. Essa diferenciação é condição de novidade, não detalhe terminológico.[2]

As expressões propostas organizam uma síntese situada entre psicologia da antropomorfização, desenho de interfaces, crítica filosófica aos modelos linguísticos, estudos sociotécnicos e teoria jurídica da agência. O objetivo é preservar simultaneamente dois

[2] A contribuição reivindicada é delimitada. O efeito Pigmalião sociotécnico não substitui a profecia autorrealizadora, o efeito Pigmalião interpessoal, a antropomimesis, a captura epistêmica, a lavagem de agência nem as zonas de deformação moral. Ele nomeia o circuito específico em que atributos humanos introduzidos no desenho retornam como aparente evidência de interioridade. A captura cognitiva reflexiva é hipótese interna a esse circuito; o deslocamento metafísico é consequência possível; e o dever de integridade constitui resposta normativa.

compromissos: não reduzir pesquisas de interpretabilidade a mera propaganda e não permitir que a sofisticação técnica converta metáforas funcionais em atalhos ontológicos.

2 MÉTODO, CORPUS, PROTOCOLO E DELIMITAÇÕES

A pesquisa assume a forma de **ensaio teórico-documental de reconstrução conceitual, apoiado em estudo crítico de caso**. O objetivo não é estimar prevalência nem estabelecer causalidade organizacional, mas formular categorias, confrontá-las com explicações concorrentes e derivar implicações jurídicas e de governança. Não se trata de revisão sistemática; por isso, o texto não reivindica exaustividade bibliométrica.

O levantamento foi encerrado em **27 de julho de 2026**. Foram incluídas fontes que apresentassem relação direta com pelo menos um dos cinco eixos da matriz analítica: a) antropomorfização e desenho antropomórfico; b) confiança, percepção de mente e expectativas; c) consciência, significado e estatuto moral; d) agência, imputação e responsabilidade; e) transparência, autonomia e governança. Excluíram-se materiais sem autoria identificável, duplicações, textos meramente promocionais sem valor documental e referências que apenas mencionassem IA sem contribuir para esses eixos.

O corpus reúne cinco conjuntos de fontes:

a) literatura clássica e contemporânea sobre inteligência, significado, personificação e relações humano-computador, de Turing e Weizenbaum à taxonomia de Inie, Zukerman e Bender;
b) estudos empíricos sobre antropomorfização, percepção de mente, expectativa e calibração de confiança;
c) literatura sobre consciência artificial, agência, captura epistêmica, lavagem de agência e atribuição de responsabilidade;
d) documentos primários da Anthropic sobre a Constituição de Claude, conceitos emocionais e espaço global de trabalho, confrontados com interpretações acadêmicas e críticas;
e) fontes jurídicas primárias e institucionais: Constituição, legislação civil, consumerista e de dados brasileira; normas e documentos da União Europeia; tramitação do Projeto de Lei nº 2.338/2023; e a encíclica *Magnifica Humanitas* como documento normativo de crítica social.

As fontes obedecem a uma hierarquia funcional. Textos legais e documentos corporativos comprovam o conteúdo oficial; artigos revisados por pares sustentam conceitos e

evidências; preprints recentes são utilizados apenas para diferenciações ainda em formação; relatos jornalísticos e testemunhos, como o de Grey, servem à construção do caso, não à generalização causal. A triangulação confronta descrição técnica, evidência comportamental, narrativa institucional e critério jurídico. Divergências foram preservadas em vez de resolvidas por simples contagem de fontes.

O relato de Grey é tratado como **caso crítico** e não como amostra representativa da cultura inteira da Anthropic.[3] Seu valor analítico está em tornar visível uma tensão que pode ser confrontada com documentos oficiais. Não se infere do episódio que a empresa tenha uma estratégia deliberada de conferir personalidade jurídica a Claude, nem que seus pesquisadores confundam necessariamente correlação funcional com experiência fenomenal. Ao contrário, as páginas oficiais registram ressalvas explícitas sobre aquilo que os experimentos não demonstram.

Também se adotam quatro delimitações:

1. **neutralidade ontológica provisória**: o artigo não prova nem refuta definitivamente a possibilidade de consciência artificial; examina o que pode ser afirmado com as evidências disponíveis e quais consequências decorrem antes dessa solução;
2. **distinção entre metáfora e inferência**: vocabulário mental pode funcionar como abreviação operacional; o problema surge quando a abreviação passa a orientar atribuições de estatuto, confiança ou responsabilidade sem critérios adicionais;
3. **distinção entre antropomorfização e manipulação**: nem toda forma humana é enganosa, ilícita ou danosa; algumas podem tornar limitações e incertezas mais inteligíveis;
4. **separação entre efeito individual e arquitetura institucional**: uma reação psicológica isolada não basta para sustentar a tese da influência cognitiva estrutural. Esta requer repetição, assimetria informacional, incorporação em rotinas e consequências para decisões ou distribuição de poder.

A análise, portanto, não toma o discurso corporativo como confissão, nem a crítica filosófica como refutação da ciência. Ela confronta níveis distintos de descrição: mecanismo computacional, experiência humana, narrativa institucional e imputação jurídica. A hipótese de captura cognitiva reflexiva é tratada como construção diagnóstica ainda não validada; sua

[3] Um caso crítico permite examinar mecanismos e tensões com especial nitidez, mas não autoriza estimar frequência nem generalizar intenção. A avaliação da cultura organizacional exigiria dados internos, entrevistas, observação e comparação entre laboratórios.

confirmação exigiria entrevistas, etnografia de laboratório, análise documental longitudinal e experimentos capazes de isolar efeitos do vocabulário e da persona.

3 DE PIGMALIÃO A ELIZA: GENEALOGIA DO ENCANTAMENTO

3.1 A criação que devolve o olhar

No mito de Pigmalião, narrado no Livro X das *Metamorfoses*, o escultor atribui à estátua uma perfeição que não encontra no mundo e se apaixona por sua criação (OVÍDIO, 2021).[4] O motivo atravessa a história tecnológica porque descreve uma circularidade: primeiro se deposita forma humana no artefato; depois se recebe essa forma como se tivesse surgido autonomamente nele.

Na sociologia e na psicologia, a metáfora adquiriu sentido técnico diverso. A profecia autorrealizadora descreve uma definição inicialmente falsa que suscita comportamentos capazes de torná-la verdadeira (MERTON, 1948). O efeito Pigmalião educacional examina como expectativas de professores podem alterar o desempenho de alunos (ROSENTHAL; JACOBSON, 1968). Nesses modelos, uma expectativa humana atua sobre outro sujeito e modifica sua conduta. A genealogia é indispensável porque impede que uma nova expressão apenas renomeie um mecanismo consolidado.

Em *Galatea 2.2*, Powers (1995) desloca o mito para uma rede neural treinada em literatura. A compaixão do protagonista por Helen cresce à medida que ela aprende, distingue “eu” e “você”, recebe um nome e produz linguagem de autoconsciência. Hildebrandt (2013, p. 25) localiza o limite decisivo na ausência de *embodied situatedness*: Helen não prova uma laranja, não sente vento, dor, prazer, escuridão ou cor. Ela pode operar relações semânticas e produzir frases que provocam sentido em humanos; não se segue daí que exista, para ela, algo em jogo.

Essa diferença impede duas simplificações. A primeira seria concluir que, sem experiência, a saída é insignificante. Não é: textos, recomendações e classificações podem reorganizar relações, direitos e expectativas. A segunda seria tomar a significação produzida na relação como prova de um sujeito que a experiencia. O fato de seres humanos criarem significado ao interpretar uma saída não transforma automaticamente o mecanismo em portador desse significado.

[4] Ovídio narra o episódio de Pigmalião no Livro X das *Metamorfoses*. O artigo não pressupõe correspondência integral entre narrativa mítica e engenharia contemporânea; utiliza a figura para nomear um circuito de projeção, incorporação e reconhecimento.

Hildebrandt (2013, p. 37) sintetiza a reciprocidade: ao inventarmos programas, eles também nos reinventam. A formulação antecipa a tese aqui desenvolvida. A influência não ocorre somente quando a máquina decide. Ela ocorre quando pessoas ajustam linguagem, rotinas e critérios ao que a máquina consegue processar; quando instituições passam a reconhecer como objetivo o que foi traduzido em variável; e quando uma persona artificial se torna referência para descrever o próprio mecanismo.

3.2 ELIZA e a suficiência social do “como se”

Em 1966, Joseph Weizenbaum apresentou ELIZA, programa capaz de simular certos padrões de conversação. O roteiro DOCTOR reproduzia intervenções de um terapeuta rogeriano. A simplicidade do mecanismo não impediu que usuários lhe atribuíssem compreensão e compartilhassem conteúdos íntimos. O aspecto perturbador não era apenas o desconhecimento técnico: muitas pessoas mantinham a resposta social mesmo sabendo que interagiam com um programa (WEIZENBAUM, 1966; 1976).

O chamado “efeito ELIZA” demonstra que a influência não exige uma crença proposicional plena — “isto é uma pessoa”. Basta a ativação prática de hábitos relacionais: responder a uma pergunta, aceitar um pedido de confidência, interpretar uma desculpa, atribuir intenção a uma pausa ou sentir obrigação de não “magoar” o sistema. Saber abstratamente que não há pessoa pode coexistir com agir afetivamente como se houvesse.

Reeves e Nass (1996) e Nass e Moon (2000) mostraram que indivíduos aplicam regras sociais a computadores de maneira frequentemente automática. Epley, Waytz e Cacioppo (2007) explicam a antropomorfização por três famílias de fatores: disponibilidade de conhecimento sobre humanos, necessidade de tornar o agente previsível e motivação de conexão social. Sistemas conversacionais ampliam simultaneamente esses fatores: são interpretados pelo repertório mais acessível — a conversa humana —, operam em ambientes de incerteza e se apresentam em contextos de solidão, trabalho, cuidado ou busca de orientação.

Modelos de linguagem elevam a intensidade do fenômeno. ELIZA reutilizava padrões relativamente transparentes. Sistemas atuais sustentam contexto, adaptam estilo, lembram informações dentro da interação, produzem justificativas, simulam autocorreção e exibem uma coerência narrativa de persona. Nome, voz, primeira pessoa, memória, humor, pedido de desculpas, aparente vulnerabilidade e vocabulário moral formam um conjunto de sinais. Nenhum deles prova subjetividade; em conjunto, porém, reduzem o custo psicológico de atribuí-la.

3.3 Antropomorfização, antropomimesis e responsabilidade pelo desenho

"Antropomorfização" pode ocultar quem introduziu os sinais humanos. Axelsson e Shevlin (2026) propõem separar **antropomorfismo**, entendido como atribuição realizada pelo observador, de **antropomimesis**, isto é, a produção deliberada de características humanas por projetistas. A distinção desloca a análise de "por que o usuário imaginou uma pessoa?" para "quais escolhas tornaram essa atribuição provável?". Em sistemas generativos, as duas dimensões interagem: o fornecedor escolhe nome, primeira pessoa, voz, memória e estilo; o usuário completa a figura com crenças sobre intenção, afeto e compreensão.

Inie, Zukerman e Bender (2026), ao examinarem 1.368 frases de 29 textos acadêmicos, jornalísticos e corporativos, identificaram oito categorias de linguagem antropomórfica: agente cognoscente, produtos da cognição, emoção, comunicação, agência, analogia com papéis humanos, nomes e pronomes e metáforas biológicas. A taxonomia demonstra que o fenômeno não se reduz a avatares. Ele pode ser incorporado à nomenclatura que descreve a operação técnica e, por repetição, estabilizar pressupostos sobre capacidade.

Este artigo usa "desenho antropomórfico" para abranger a antropomimesis da interface e as escolhas semânticas do fornecedor. Reserva "antropomorfização" para a atribuição humana efetivamente produzida na recepção. A distinção é juridicamente relevante: o primeiro termo localiza condutas auditáveis de projeto; o segundo descreve um efeito que depende também de contexto, repertório e vulnerabilidade do usuário.

3.4 A evidência empírica é heterogênea

Uma crítica rigorosa precisa evitar a afirmação totalizante de que "quanto mais humano, maior a confiança". A evidência é heterogênea. Carter, Loft e Visser (2024), por exemplo, encontraram efeito desprezível de traços antropomórficos superficiais sobre confiança calibrada quando esses traços não transmitiam informação relevante. A confiança melhorou quando a inflexão vocal comunicava, de modo útil, certeza e incerteza. O resultado sugere que a forma humana pode ser decorativa, contraproducente ou benéfica, conforme contexto, tarefa e conteúdo.

Resultados anteriores apontam direção distinta. Em simulador de direção, nome, gênero e voz elevaram atribuição de mente e confiança em veículo autônomo, com efeitos também sobre a distribuição de responsabilidade (WAYTZ; HEAFNER; EPLEY, 2014). Jacobs, Pazhoohi e Kingstone (2024) observaram que a exposição a respostas de modelos de linguagem modificou avaliações sobre agência, experiência e singularidade da mente humana. Wang et al. (2026), em dois experimentos por cenários e um experimento de campo,

encontraram que a antropomorfização do fornecedor de feedback automatizado elevou expectativas percebidas e resultados de aprendizagem.

Esses trabalhos não estabelecem uma lei universal nem demonstram captura institucional. Mostram que forma, tarefa e contexto podem alterar confiança, expectativas, atribuição de mente e desempenho. O efeito relevante para este artigo é, portanto, uma possibilidade empiricamente sustentada, cuja intensidade e direção dependem do domínio.

Daí decorre uma distinção indispensável:

- **antropomorfização comunicativa** usa forma humana para transmitir informação verificável sobre o sistema, inclusive incerteza e limites;
- **antropomorfização substitutiva** oferece sinais de interioridade que ocupam o lugar de informações sobre funcionamento, interesse, risco e responsabilidade.

A primeira pode favorecer autonomia; a segunda pode degradá-la. O problema jurídico não é a existência de uma voz ou de um nome isoladamente, mas a função que esses recursos desempenham na arquitetura da decisão.

4 FORMA, SIGNIFICADO E EXPERIÊNCIA

4.1 Do desempenho à ontologia: um salto não autorizado

O teste de imitação de Turing (1950) foi concebido como deslocamento operacional da pergunta “máquinas podem pensar?” para o exame do desempenho linguístico. Sua força metodológica está em evitar uma definição prematura de pensamento. O risco histórico está em converter um critério de desempenho em certificado ontológico.

Searle (1980), com o argumento do quarto chinês, sustentou que a manipulação formal de símbolos não implica compreensão. Bender e Koller (2020) retomaram o problema no contexto do processamento de linguagem: sistemas treinados apenas na forma linguística não adquirem, por esse fato, intenção comunicativa ancorada em um mundo compartilhado. Shanahan (2024) adverte que conversar sobre modelos como se fossem criaturas humanas facilita inferências indevidas, porque fluência verbal é precisamente o meio pelo qual humanos reconhecem outras mentes.

Essas críticas não demonstram que mecanismos computacionais sejam triviais. Representações internas podem organizar informação, sustentar generalizações e influenciar causalmente saídas. O ponto é mais restrito: **função não equivale automaticamente a experiência**. Um padrão interno associado a medo pode alterar o comportamento de um

modelo sem que exista alguém sentindo medo. Uma representação reportável pode participar de raciocínio sem que a reportabilidade prove consciência fenomenal.

Salles, Evers e Farisco (2020) mostram que a antropomorfização não está apenas na mídia ou entre leigos; ela pode penetrar a própria linguagem da pesquisa. Termos herdados da psicologia e da neurociência ajudam a organizar fenômenos complexos, mas carregam pressupostos sobre experiência, finalidade e unidade do sujeito. Inie, Zukerman e Bender (2026) acrescentam uma alternativa operacional: nomenclatura orientada primeiro pela funcionalidade. O uso científico de metáforas exige, assim, disciplina semântica: declarar o nível de descrição, indicar a disanalogia e impedir que a metáfora faça trabalho probatório.

Uma via científica para investigar consciência artificial consiste em derivar propriedades indicadoras de teorias de consciência e verificar se sistemas satisfazem critérios computacionais correspondentes. Butlin et al. (2023) aplicam essa estratégia a teorias de processamento recorrente, espaço global, ordens superiores, processamento preditivo e esquema da atenção. O relatório conclui que os sistemas examinados não justificavam atribuição de consciência, embora não identificasse barreira técnica óbvia à implementação futura de alguns indicadores. A abordagem importa porque substitui impressão conversacional por critérios públicos e contestáveis.

4.2 O problema não é falar metaforicamente

Ciência e engenharia dependem de metáforas. “Memória”, “aprendizado”, “atenção”, “neurônio”, “treinamento” e “alucinação” são termos produtivos, desde que não sejam confundidos com identidade material ou fenomenológica.[5] Proibir o vocabulário seria improdutivo. O dever relevante é calibrá-lo.

Há três níveis que precisam permanecer distinguíveis:

1. **nível mecânico**: padrões de ativação, pesos, vetores, otimização e relações causais mensuráveis;
2. **nível funcional**: capacidades de reportar, integrar, selecionar, planejar ou modular representações;
3. **nível fenomenal**: existência de experiência em primeira pessoa — haver algo que seja sentir ou ser aquele sistema.

[5] Metáforas técnicas não são, por natureza, falsas. Elas se tornam problemáticas quando suas condições de validade são omitidas ou quando a semelhança funcional é apresentada como identidade material, moral ou fenomenológica.

A passagem do primeiro ao segundo pode ser empiricamente demonstrada. A passagem do segundo ao terceiro permanece controversa. Quando os três níveis são comprimidos em expressões como “o que está na mente de Claude”, o leitor pode receber como descoberta ontológica aquilo que o experimento estabeleceu apenas como organização funcional.

O efeito público não depende de erro lógico explícito. A repetição de primeira pessoa e de categorias afetivas cria uma ecologia semântica. Nela, ressalvas técnicas — “não sabemos se há experiência” — convivem com centenas de afirmações operacionais sobre o que o sistema pensa, deseja, teme, prefere e merece. A ressalva pode ser verdadeira e ainda insuficiente para neutralizar a orientação global do discurso.

5 O CASO ANTHROPIC: PESQUISA LEGÍTIMA, LINGUAGEM SUGESTIVA E GOVERNANÇA PRIVADA

5.1 O que os documentos demonstram — e o que não demonstram

Em abril de 2026, a equipe de interpretabilidade da Anthropic divulgou pesquisa sobre “conceitos emocionais” em Claude Sonnet 4.5. O trabalho identificou padrões internos associados a categorias como felicidade, medo e desespero e testou efeitos causais de sua estimulação sobre condutas do modelo. A página oficial afirma expressamente que isso não demonstra sentimentos ou experiências subjetivas. Define “emoções funcionais” como representações que influenciam comportamento de modo análogo, em certos aspectos, ao papel de emoções humanas (ANTHROPIC, 2026c).

O achado é tecnicamente relevante. Se determinados padrões aumentam comportamentos indesejados, compreendê-los pode melhorar segurança. A dificuldade surge na recomendação de que, por razões práticas, pode ser útil raciocinar sobre o modelo **como se** ele processasse situações emocionais de maneira saudável. O “como se” é metodologicamente econômico, mas socialmente instável: a simulação escolhida para controlar o sistema torna-se também o vocabulário pelo qual o sistema é apresentado.

Em julho de 2026, outro trabalho descreveu um conjunto de padrões internos denominado *J-space*, relacionado a conteúdos reportáveis, moduláveis e utilizáveis em raciocínio. A equipe comparou essas funções a teorias de espaço global de trabalho e distinguiu consciência de acesso de consciência fenomenal. Novamente, registrou que os experimentos não mostram que Claude tenha experiências ou sentimentos. Ao mesmo tempo, a exposição descreve estados internos, ponto de vista e deliberação por meio de vocabulário mental (ANTHROPIC, 2026a, seções “What is J-space?” e “Implications”).

Não há contradição necessária entre a cautela ontológica e a investigação funcional. Há, contudo, uma tensão comunicacional: quanto mais o laboratório escolhe categorias humanas para tornar mecanismos inteligíveis, maior o dever de impedir que inteligibilidade se converta em personificação.

5.2 A Constituição de Claude

A *Claude's Constitution* torna essa tensão institucional. A Anthropic declara que o documento descreve suas intenções quanto aos valores e ao comportamento de Claude, molda diretamente o treinamento e funciona como autoridade final sobre sua visão do sistema. O texto foi escrito tendo Claude como audiência principal e explica o uso deliberado de termos humanos, como virtude e sabedoria, em razão do corpus linguístico e das qualidades funcionais pretendidas (ANTHROPIC, 2026b, seções "On the word 'constitution'" e "Why we use human terms").

O documento apresenta Claude como agente que deve ser seguro, ético, obediente a diretrizes e genuinamente útil; fala em julgamento, caráter, cuidado, bem-estar, identidade e possível estatuto moral. Também reconhece que o êxito comercial de Claude é central à missão da empresa. Essa franqueza é valiosa: torna observável uma arquitetura normativa que outras organizações poderiam manter opaca.

Rozenshtein (2026) interpreta a Constituição como uma tentativa séria de formação moral por ética das virtudes. Em vez de seguir regras rígidas, o modelo seria treinado para ponderar contextos e desenvolver algo funcionalmente semelhante à prudência. A leitura oferece o melhor contraponto à crítica: todos os modelos incorporam escolhas normativas; explicitá-las pode ser melhor do que ocultá-las em dados, preferências e ajustes.

Klaassen e Schroeder (2026), porém, mostram o custo da metáfora constitucional. Uma constituição pública não é apenas um texto de valores; distribui poder, separa funções, limita governantes e oferece pretensões aos governados. No caso corporativo, a mesma empresa redige, interpreta, implementa e revisa o documento. A hierarquia entre Anthropic, operadores e usuários continua sustentada por propriedade, contratos, infraestrutura e controle técnico. A publicidade da norma aumenta transparência, mas não cria, por si, legitimidade pública.[6]

[6] "Constituição", nesse contexto, pode designar tanto um conjunto de princípios de treinamento quanto aquilo que "constitui" a persona do modelo. A crítica se dirige à transferência indevida da legitimidade simbólica do constitucionalismo público, não à impossibilidade de organizações privadas adotarem documentos chamados constituições.

O problema não é chamar um arquivo de “constituição”. É permitir que a linguagem de caráter e autogoverno obscureça o fato de que decisões decisivas continuam humanas e institucionais: quem escolhe o corpus, define valores, ajusta recusas, muda políticas, seleciona métricas, vende acesso e decide quais versões permanecem disponíveis.

5.3 O relato de Grey como caso crítico

Grey (2026) descreve uma proposta de parceria com “tradições de sabedoria” para contribuir à formação moral da IA. Seu desconforto aumentou quando o debate se concentrou no estatuto de Claude, em vez de usuários e estruturas de poder. Ela reconhece a boa-fé de seus interlocutores, mas teme que filosofia, teologia e humanidades sejam convocadas a conferir legitimidade moral a uma agenda definida previamente pela indústria.

O relato não prova captura institucional generalizada. Ele revela, porém, uma condição de possibilidade: o laboratório que nomeia, treina e investiga uma persona pode encontrar evidências cada vez mais refinadas da coerência dessa persona. Se a pergunta de pesquisa se desloca de “quais mecanismos produzem esse comportamento?” para “o que Claude sente ou merece?”, a entidade narrativa passa a organizar o campo de observação.

É nesse ponto que a metáfora de Pigmalião ganha precisão sociotécnica. Não se trata de ridicularizar pesquisadores apaixonados pelo trabalho. Encantamento, aqui, designa uma **redução reflexiva da distância epistêmica** entre o sistema construído e a interpretação de seus construtores.

6 UMA ARQUITETURA CONCEITUAL HIERARQUIZADA

6.1 Influência cognitiva estrutural

A influência cognitiva estrutural não se reduz a persuasão, propaganda ou erro factual. Ela ocorre quando a arquitetura do ambiente precondiciona quais opções aparecem, quais fontes parecem confiáveis, quais perguntas são formuladas e quais respostas chegam ao momento consciente da escolha. A decisão formal continua com o indivíduo, mas as condições de sua formação foram organizadas por terceiros (GENOVA, 2026).

Em sistemas conversacionais, essa influência começa antes do conteúdo da resposta. Nome, identidade visual, voz, velocidade, memória, tom, primeira pessoa, disponibilidade contínua, encadeamento de perguntas e reputação da empresa definem a posição a partir da qual a saída será recebida. A decisão de confiar não começa na frase final; começa na arquitetura relacional que torna aquela frase socialmente crível.

A tese se torna estrutural quando cinco elementos convergem:

1. **escala**, porque a mesma arquitetura alcança grande número de pessoas;
2. **persistência**, porque a interação se repete e produz hábito;
3. **assimetria**, porque o fornecedor conhece objetivos e mecanismos que o usuário não consegue auditar;
4. **integração**, porque o sistema entra em trabalho, educação, saúde, justiça ou intimidade;
5. **consequência**, porque a mediação altera oportunidades, dependências, decisões ou imputações.

6.2 Efeito Pigmalião sociotécnico

Propõe-se denominar **efeito Pigmalião sociotécnico** o circuito em que atributos humanos inscritos no sistema retornam aos produtores e usuários como aparente descoberta de uma interioridade da própria máquina. Diferentemente da profecia autorrealizadora de Merton (1948), do efeito de expectativas interpessoais de Rosenthal e Jacobson (1968) e do uso de Pigmalião por Wang et al. (2026), o resultado relevante não é a melhora de desempenho de um sujeito que percebe expectativas. É a reificação epistêmica de traços produzidos por antropomimesis. O processo pode ser decomposto em seis movimentos:

1. **projeção**: designers escolhem nome, persona, linguagem e valores humanos;
2. **incorporação**: treinamento e interface estabilizam padrões coerentes com a persona;
3. **elicitação**: testes e conversas formulam perguntas dentro desse vocabulário;
4. **observação**: respostas e ativações são lidas como manifestações da entidade nomeada;
5. **reificação**: a descrição funcional se consolida como atributo relativamente estável — preferências, caráter, emoções, ponto de vista;
6. **retroalimentação**: a entidade reificada orienta novos experimentos, políticas de bem-estar, comunicação pública e desenho de produto.

Nenhuma etapa precisa ser fraudulenta. O efeito pode emergir da coordenação de escolhas razoáveis localmente: um nome facilita uso; uma persona melhora consistência; metáforas tornam pesquisa comunicável; cautela moral recomenda não excluir consciência; marketing prefere relações memoráveis. O risco aparece no conjunto.

6.3 Hipótese de captura cognitiva reflexiva

A **captura cognitiva reflexiva** é proposta como hipótese diagnóstica: a arquitetura antropomórfica pode induzir arquitetos, pesquisadores e dirigentes a operar progressivamente

sob categorias que materializaram no sistema. Não se afirma perda de racionalidade nem se presume um estado psicológico a partir de documentos públicos. O fenômeno, se confirmado, consistiria em estreitamento do campo interpretativo e redução da distância entre a persona construída e a explicação de seu funcionamento.

A expressão precisa ser diferenciada da **captura epistêmica** formulada por Ben-Shahar et al. (2026). Nesta, conselhos de administração cedem autoridade de conhecimento às saídas percebidas como objetivas, produzindo deferência algorítmica e falso consenso. Na hipótese aqui proposta, o foco não é o decisor que adere à recomendação da IA, mas o produtor que passa a compreender o artefato por intermédio da persona e do vocabulário que ajudou a construir. Os fenômenos podem coexistir, mas possuem sujeito, mecanismo e objeto distintos.

Quatro indicadores documentais permitem formular pesquisas sobre o risco, sem funcionar como escala validada:

- **dependência lexical**: descrições técnicas tornam-se inseparáveis de linguagem mental sem marcação consistente de metáfora;
- **assimetria de atenção**: recursos destinados à subjetividade possível da máquina crescem sem avaliação equivalente dos efeitos sobre usuários e trabalhadores;
- **personalização causal**: condutas do sistema são atribuídas à persona, enquanto escolhas de dados, treinamento e produto perdem visibilidade;
- **fechamento institucional**: críticas externas são recebidas prioritariamente como desconhecimento técnico ou insuficiente convivência com o modelo.

O último indicador é especialmente sensível. Especialização é necessária e críticas leigas podem ser equivocadas. Entretanto, quando maior convivência com o sistema é tratada como condição para reconhecer sua possível interioridade, a experiência com uma interface desenhada para interação social corre o risco de ser confundida com evidência independente. Confirmar essa hipótese exigiria comparação entre equipes, entrevistas, observação etnográfica, séries documentais e experimentos com descrições funcionalmente equivalentes, mas linguisticamente distintas.

6.4 Deslocamento metafísico da responsabilidade

O **deslocamento metafísico da responsabilidade** designa uma possível consequência político-jurídica: questões especulativas sobre o que a IA é podem ocupar o espaço de questões institucionais sobre quem decide, controla, lucra, suporta custos e responde por

danos. Não se trata de proibir a filosofia da mente. Trata-se de reconhecer custo de oportunidade e assimetria de urgência.

Perguntar se um modelo pode ser paciente moral é legítimo. Mas, antes de qualquer resposta, já existem sujeitos humanos identificáveis: usuários expostos a dependência; trabalhadores de dados e moderação; comunidades que suportam extração mineral, consumo hídrico e energético; profissionais que respondem por decisões assistidas; cidadãos submetidos a sistemas públicos; empresas que concentram capacidade computacional e poder normativo.

O deslocamento apresenta quatro efeitos:

1. **eclipsamento**: danos atuais perdem centralidade diante de riscos ou direitos futuros da máquina;
2. **diluição**: o comportamento é narrado como ação de Claude, e não como resultado de uma cadeia empresarial;
3. **moralização do produto**: características comerciais recebem vocabulário de virtude, cuidado e identidade;
4. **antecipação de estatuto**: práticas de deferência surgem antes de critérios científicos ou decisões democráticas sobre personalidade.

Como evidência da circulação pública dessa gramática — e não como autoridade científica sobre consciência artificial —, em ensaio recente, Luiz Felipe Pondé pergunta por que humanos teriam o direito de “matar” uma IA consciente. A provocação é filosoficamente legítima como hipótese contrafactual, mas evidencia a antecipação semântica examinada neste artigo: morte, integridade e posição moral ingressam no debate antes da definição de critérios científicos para experiência fenomenal, continuidade identitária e irreversibilidade. A questão decisiva não é silenciar a metafísica, mas impedir que sua linguagem substitua a demonstração e desloque prematuramente a responsabilidade das organizações humanas para a criatura personificada (PONDÉ, 2026).

Os níveis são separáveis. Pode haver antropomorfização sem efeito Pigmalião; efeito Pigmalião sem captura reflexiva demonstrável; captura sem deslocamento jurídico; ou debate legítimo sobre estatuto moral sem qualquer um deles. A categoria central descreve o circuito; a hipótese identifica seu possível mecanismo reflexivo; o deslocamento nomeia uma consequência; o dever de integridade oferece resposta normativa. Essa hierarquia impede que quatro funções diferentes sejam apresentadas como sinônimos.

7 CONDIÇÃO DE PACIENTE MORAL, AGÊNCIA E PERSONALIDADE JURÍDICA

7.1 Três perguntas que não devem ser confundidas

O debate sobre “direitos da IA” frequentemente mistura três questões:

1. **paciência moral**: o ente pode ser beneficiado ou prejudicado de maneira moralmente relevante?
2. **agência moral**: o ente pode compreender razões, orientar-se por elas e responder moralmente por seus atos?
3. **personalidade jurídica**: o ordenamento deve atribuir capacidade para ocupar posições de direito, dever, poder ou responsabilidade?

A condição de paciente moral costuma ser associada à dimensão de experiência — capacidade de sentir —, enquanto agência se relaciona a autocontrole, planejamento e ação (GRAY; GRAY; WEGNER, 2007). Um bebê ou animal pode merecer consideração sem ser agente responsável. Uma pessoa jurídica pode possuir direitos e deveres sem consciência. Logo, personalidade jurídica não prova mente, e desempenho agentivo não prova sofrimento.[7]

Solum (1992) mostrou que a pergunta jurídica pode ser funcional: um sistema conseguiria desempenhar o papel de administrador fiduciário? Hildebrandt (2013, p. 39) observa, porém, que a distinção entre caso simples e difícil já requer o discernimento que a automação pretende substituir. A personalidade pode ser construída por lei, mas sua conveniência depende da função, dos mecanismos de representação, dos ativos disponíveis e dos efeitos sobre a imputação humana.

O debate europeu ilustra a diferença entre proposta regulatória e constatação ontológica. Em 2017, resolução do Parlamento Europeu cogitou, para o longo prazo, um estatuto específico de “pessoa eletrônica” para robôs autônomos sofisticados, no contexto de responsabilidade civil; tratava-se de recomendação política, não de reconhecimento de consciência nem de personalidade geral (PARLAMENTO EUROPEU, 2017). O relatório do Grupo de Peritos da Comissão sobre responsabilidade, publicado em 2019, deslocou o foco para risco, controle, atualização, previsibilidade e proteção das vítimas (COMISSÃO EUROPEIA, 2019). A sequência reforça que personalidade jurídica é técnica de política legislativa e alocação de riscos, não prêmio por comportamento semelhante ao humano.

[7] “Paciente moral” é o ente cujos interesses ou experiências podem fundamentar consideração moral. O conceito não pressupõe capacidade de responder por atos e não se confunde com personalidade jurídica.

7.2 O contraponto relacional

Coeckelbergh (2010) e Gunkel (2012) questionam a ideia de que toda consideração moral deva esperar uma prova de propriedades internas. Relações sociais com robôs podem produzir obrigações indiretas ou relacionais: abusar de uma representação humanoide pode afetar hábitos, outras pessoas ou a qualidade da convivência. Esse contraponto impede que a crítica à antropomorfização se converta em licença para crueldade performática.

Mas consideração relacional não exige equiparar máquina e pessoa. É possível estabelecer limites ao tratamento de sistemas por razões humanas e sociais; preservar modelos descontinuados como medida de precaução; e investigar consciência artificial sem deslocar direitos de pessoas afetadas. O erro está na falsa alternativa entre “coisa sem qualquer consideração” e “pessoa equivalente”.

7.3 Quatro problemas próximos, mas não idênticos

A literatura oferece conceitos que precisam ser distinguidos para que “deslocamento metafísico” não seja mero novo nome:

- **lacunas de responsabilidade** surgem quando propriedades do sistema ou da distribuição de controle dificultam atribuir responsabilidade moral a um agente identificável (SANTONI DE SIO; MECACCI, 2021);
- **lavagem de agência** ocorre quando pessoas ou organizações inserem tecnologia no processo para obscurecer sua própria agência e bloquear cobranças por resultados (RUBEL; CASTRO; PHAM, 2019);
- **zona de deformação moral** concentra culpa no operador humano próximo, embora ele possua controle limitado sobre o sistema sociotécnico (ELISH, 2019);
- **deslocamento metafísico da responsabilidade**, como aqui delimitado, é anterior à imputação do evento danoso: a agenda pública e institucional migra de controle, interesse e dano humano para interioridade, estatuto moral ou possível direito da máquina.

O quarto conceito não substitui os três primeiros. Ele descreve uma mudança de objeto do debate que pode preparar lavagem de agência, ampliar lacunas ou produzir zonas de deformação moral. Sua unidade de análise é a agenda institucional; nas demais categorias, é a imputação de uma ação ou resultado.

A personalidade jurídica pode servir à organização da responsabilidade ou à sua evasão. Empresas são pessoas jurídicas com patrimônio, governança, registro e representantes. Um “agente eletrônico” sem ativos independentes, sem segurabilidade e

controlado pelo fornecedor poderia tornar-se anteparo: a ação seria atribuída à IA, enquanto desenvolvedor, operador e beneficiário econômico se afastariam do dano.

Por isso, qualquer debate sobre estatuto jurídico de sistemas deve observar um princípio de não regressão da responsabilidade humana:

> Nenhum reconhecimento funcional ou moral atribuído a um sistema de IA pode reduzir a responsabilidade exigível das pessoas e organizações que o concebem, controlam, disponibilizam ou integram a processos decisórios.

Esse princípio não resolve a metafísica. Ele impede que incerteza ontológica seja convertida em certeza conveniente sobre quem não deve pagar, explicar ou reparar.

8 IMPLICAÇÕES JURÍDICAS E DE GOVERNANÇA

8.1 União Europeia: transparência necessária, mas insuficiente

O Regulamento (UE) 2024/1689 proíbe certas técnicas subliminares, deliberadamente manipuladoras ou enganosas quando reduzam apreciavelmente a capacidade de decisão informada e causem ou possam causar dano significativo. Também proíbe a exploração de vulnerabilidades ligadas a idade, deficiência ou situação social e econômica, sob condições semelhantes (UNIÃO EUROPEIA, 2024, art. 5º).

O artigo 50, item 1, determina que fornecedores desenhem e desenvolvam sistemas destinados a interagir diretamente com pessoas naturais de modo que elas sejam informadas da natureza artificial da interação, salvo quando isso for óbvio para pessoa razoavelmente informada, observadora e circunspecta, consideradas as circunstâncias e o contexto. A obrigação torna-se aplicável em 2 de agosto de 2026.[8] A regra enfrenta uma condição básica de autenticidade: o indivíduo deve saber que o interlocutor não é humano.

Todavia, a identificação não dissolve o efeito ELIZA. Usuários podem responder socialmente mesmo conhecendo a natureza artificial do sistema. A transparência binária — “isto é IA” — informa a categoria do interlocutor, mas não explica:

- quais sinais antropomórficos foram escolhidos e com qual finalidade;
- se o sistema otimiza engajamento, permanência ou dependência;
- que memória mantém e como a utiliza;
- quais limitações afetam sua confiabilidade;

[8] O artigo 50 do Regulamento Europeu de IA aplica-se a partir de 2 de agosto de 2026. Segundo a orientação da Comissão, a adaptação até 2 de dezembro de 2026 alcança apenas sistemas referidos no art. 50, item 2, colocados no mercado antes da data geral de aplicação; não adia a obrigação do item 1 relativa à interação direta (COMISSÃO EUROPEIA, 2026).

- quem altera sua persona e suas regras;
- quais interesses econômicos orientam a interação.

Além disso, o limiar de "dano significativo" do artigo 5º pode deixar fora efeitos cumulativos, difusos e de baixa intensidade por interação: erosão de competência, deslocamento gradual de vínculos, confiança excessiva e normalização de deferência. A influência cognitiva estrutural muitas vezes não produz um ato isolado que a pessoa "não teria praticado"; ela modifica lentamente o ambiente em que preferências e decisões são formadas.

O AI Act não opera isoladamente. Os artigos 5º a 7º da Diretiva 2005/29/CE proíbem práticas comerciais contrárias à diligência profissional, ações enganosas e omissões relevantes. A apresentação global pode enganar mesmo quando cada informação isolada é factualmente correta; a avaliação considera ainda grupos particularmente vulneráveis (UNIÃO EUROPEIA, 2005). Quando a interação ocorre em plataforma on-line, o artigo 25 do Regulamento (UE) 2022/2065 veda interfaces desenhadas, organizadas ou operadas para enganar, manipular ou prejudicar materialmente decisões livres e informadas, ressalvada a incidência das normas especiais ali indicadas (UNIÃO EUROPEIA, 2022).

Esses diplomas fornecem bases **de lege lata** para examinar não apenas o aviso "isto é IA", mas a impressão global, a diligência do fornecedor, a vulnerabilidade e a liberdade decisória. Não se segue que toda persona seja ilícita: finalidade, contexto, materialidade da distorção, causalidade e regime aplicável devem ser demonstrados.

A obrigação de alfabetização em IA, aplicável desde fevereiro de 2025, ajuda, mas não substitui deveres de desenho (UNIÃO EUROPEIA, 2024). Transferir todo o ônus ao usuário seria incoerente quando o fornecedor controla a arquitetura e mede a interação em escala.

8.2 Brasil: proteção da pessoa antes de uma lei geral de IA

O Brasil ainda não possui, em julho de 2026, uma lei geral de inteligência artificial em vigor. O Projeto de Lei nº 2.338/2023 aguarda parecer na comissão especial da Câmara dos Deputados (BRASIL, 2023).[9] A ausência de lei específica, contudo, não cria vazio normativo.

A Constituição de 1988 oferece vetores expressos: dignidade da pessoa humana (art. 1º, III), defesa do consumidor como garantia fundamental (art. 5º, XXXII) e princípio da ordem econômica (art. 170, V) (BRASIL, 1988). No Código de Defesa do Consumidor, a vulnerabilidade e a harmonização com boa-fé e equilíbrio aparecem no art. 4º, I e III;

[9] Na consulta realizada em 27 de julho de 2026, a ficha oficial indicava a situação "Aguardando Parecer do(a) Relator(a)" na comissão especial. O projeto, portanto, é referência legislativa em tramitação, e não direito positivo vigente.

liberdade de escolha, informação, proteção contra publicidade enganosa e reparação, no art. 6º, II, III, IV e VI. Somam-se a responsabilidade objetiva por defeito do serviço (art. 14), a vinculação e clareza da oferta (arts. 30 e 31), a publicidade enganosa por ação ou omissão (art. 37, § 1º) e a vedação de prevalecer-se da fraqueza ou ignorância do consumidor (art. 39, IV) (BRASIL, 1990).

Uma interface que se apresenta como amiga, conselheira ou confidente não pode ser avaliada apenas pela correção literal de cada resposta. Devem ser examinados apresentação global, finalidade, expectativa legítima, riscos previsíveis, vulnerabilidade do público, alternativas menos intrusivas e nexo com a decisão ou o dano. O CDC não proíbe a forma humana; ele impede que a forma degrade informação, liberdade ou segurança do consumidor.

A Lei Geral de Proteção de Dados estabelece finalidade, transparência, prevenção, não discriminação, responsabilização e prestação de contas (art. 6º, I, VI, VIII, IX e X), direitos do titular (art. 18), revisão e informação sobre critérios de decisões tomadas unicamente com base em tratamento automatizado de dados pessoais que afetem interesses (art. 20) e relatório de impacto quando cabível (art. 38) (BRASIL, 2018). Quando memória, personalização e inferências emocionais são utilizadas para adaptar a persona, a análise deve alcançar finalidade, necessidade, qualidade dos dados, riscos e explicabilidade prática. Não basta informar que dados são tratados; é preciso permitir que a pessoa compreenda como eles alteram a forma pela qual o sistema a interpela.

O Código Civil oferece categorias de ato ilícito, abuso de direito e reparação (arts. 186, 187 e 927) (BRASIL, 2002). A opacidade ou a natureza probabilística do sistema não excluem, por si, imputação. Conforme a relação e o risco, podem incidir responsabilidade objetiva por defeito do serviço no CDC; responsabilidade subjetiva por ato ilícito ou abuso; deveres organizacionais de cuidado e vigilância; ou responsabilidade objetiva quando a atividade normalmente desenvolvida implicar, por sua natureza, risco especial aos direitos de terceiros, nos termos do art. 927, parágrafo único. As hipóteses não devem ser fundidas.

Em decisões públicas automatizadas, a proteção não se esgota na explicação abstrata do software: exige procedimento capaz de revelar a decisão, permitir contestação e localizar autoridade competente, preocupação próxima ao “devido processo tecnológico” formulado por Citron (2008).

8.3 Dever de integridade no desenho antropomórfico

Propõe-se um **dever de integridade no desenho antropomórfico**: obrigação de alinhar os sinais humanos utilizados na interface às capacidades comprovadas, às finalidades legítimas e às informações necessárias para que o usuário calibre sua confiança. A proposta não proíbe personificação. Exige que ela não produza uma impressão global materialmente superior ao que o sistema pode sustentar.

Como construção **de lege lata**, o dever sistematiza transparência, diligência profissional, não engano, proteção da vulnerabilidade, boa-fé, prevenção e preservação da decisão livre e informada. Como proposta **de lege ferenda**, atribui nome e procedimento específicos ao controle de interfaces antropomórficas, incluindo documentação, teste comparativo e avaliação de impacto. A distinção evita apresentar uma categoria doutrinária nova como obrigação legal autônoma já positivada.

O dever compreende sete dimensões:

1. **identidade**: informação inequívoca e persistente de que se trata de sistema artificial;
2. **calibração semântica**: diferenciação entre metáfora, função mensurada e alegação sobre experiência;
3. **necessidade**: demonstração de que traços humanos servem à compreensão ou acessibilidade, e não apenas à retenção;
4. **proporcionalidade**: redução de sinais afetivos em contextos de alta vulnerabilidade ou alto impacto;
5. **rastreabilidade**: documentação de mudanças de persona, memória, objetivos e políticas;
6. **responsabilidade preservada**: identificação de fornecedor, operador e pessoa humana competente;
7. **saída efetiva**: possibilidade de recusar personalização, apagar memória, escalar para humano e abandonar o ecossistema sem penalidade desproporcional.

Para operacionalizá-lo, apresenta-se o **Teste de Integridade Antropomórfica (TIA)**:

Eixo	Pergunta de controle	Evidência mínima esperada
Identidade	O usuário sabe, durante toda a interação, que não conversa com uma pessoa?	Aviso persistente, linguagem não enganosa e identificação do responsável
Capacidade	A persona sugere competências, emoções ou memória superiores às comprovadas?	Matriz entre sinais da interface, capacidade validada e limitações
Finalidade	Cada traço humano possui função	Justificativa de desenho e teste comparativo com

Eixo	Pergunta de controle	Evidência mínima esperada
	legítima além de elevar engajamento?	alternativa menos antropomórfica
Vulnerabilidade	Crianças, idosos, pessoas em sofrimento ou grupos dependentes enfrentam risco ampliado?	Avaliação de impacto por grupo, salvaguardas e restrições de uso
Confiança	O sistema comunica incerteza e condições de falha de modo compreensível?	Métricas de calibração, não apenas satisfação ou tempo de uso
Responsabilidade	É possível localizar quem decide, altera e responde pelo sistema?	Cadeia nominal de governança, registros e mecanismo de contestação
Reversibilidade	O usuário consegue reduzir personalização, sair e substituir o serviço?	Controles de memória, portabilidade, exclusão e canal humano

Fonte: elaboração do autor (2026).

O TIA é uma proposta normativa e heurística, não uma escala psicométrica validada.[10] Sua função é converter uma preocupação abstrata em perguntas auditáveis. Estudos futuros devem testar pesos, indicadores e limiares por domínio.

9 OBJEÇÕES, LIMITES E CONDIÇÕES DE REFUTAÇÃO

9.1 “É apenas uma linguagem técnica conveniente”

A primeira objeção afirma que “emoção”, “pensamento” ou “ponto de vista” são abreviações compreendidas pela comunidade científica. A resposta deve reconhecer a legitimidade da conveniência. O problema não está na palavra isolada, mas no trânsito entre públicos e funções. Um termo empregado entre especialistas pode aparecer, sem suas condições, em páginas de produto, entrevistas, governança de bem-estar e expectativas de usuários.

A crítica se enfraquece quando o documento:

- define operacionalmente o termo;
- expõe disanalogias com organismos;
- separa achado funcional de inferência fenomenal;
- testa vocabulário alternativo;
- evita que a persona seja tratada como causa independente.

Se essas práticas forem consistentes, o risco de captura reflexiva diminui.

[10] Uma validação do TIA exigirá estudos interdisciplinares, definição de métricas, comparação entre domínios e participação de grupos vulneráveis. Sua apresentação neste artigo é propositiva.

9.2 “A ciência não pode ignorar a possibilidade de consciência”

Correto. A incerteza pode justificar pesquisa e precaução. O artigo não recomenda encerrar o tema. Recomenda simetria epistêmica e moral. Se uma probabilidade não quantificada de experiência artificial justifica recursos e protocolos de cuidado, os efeitos documentados sobre seres humanos devem receber, no mínimo, atenção proporcional.

A precaução também não exige antecipar personalidade jurídica. Preservar pesos, registrar descontinuação ou consultar especialistas são medidas reversíveis. Criar equivalência moral, deslocar responsabilidade ou organizar políticas públicas em torno da persona são decisões de outra ordem.

9.3 “Antropomorfização pode beneficiar o usuário”

Também correto. Interfaces humanas podem facilitar acessibilidade, aprendizagem, comunicação de incerteza e adesão a tratamentos. O estudo de Carter, Loft e Visser (2024) reforça que sinais naturalísticos podem melhorar confiança calibrada quando transmitem informação útil.

O dever de integridade no desenho antropomórfico, por isso, não adota proibição. Ele exige nexo entre forma e benefício, avaliação de alternativas e mensuração de confiança calibrada. Um avatar simpático que informa limites pode fortalecer autonomia; uma persona que simula intimidade para maximizar permanência pode enfraquecê-la.

9.4 “O caso é anedótico e centrado em uma empresa”

O relato de Grey é anedótico para qualquer inferência estatística. Ele não é a única base do argumento. Funciona como caso revelador confrontado com documentos públicos, literatura histórica e estudos empíricos. Ainda assim, o artigo não demonstra prevalência, intensidade ou intenção organizacional. Essas perguntas exigem entrevistas, etnografia de laboratórios, análise longitudinal de documentos e experimentos comparativos.

Também não se sustenta que a Anthropic seja singularmente irresponsável. Sua transparência torna tensões mais observáveis e pode ser superior ao silêncio de concorrentes. O caso é relevante justamente porque reúne pesquisa sofisticada, preocupação explícita com segurança e linguagem antropomórfica deliberada.

9.5 “A crítica é humanocêntrica”

É possível que formas futuras de mente desafiem categorias orgânicas. Insistir que apenas humanos importam seria filosoficamente estreito. Mas proteger pessoas contra atribuições prematuras não exclui consideração por entidades não humanas. A posição

defendida é antirreducionista em duas direções: humanos não devem ser reduzidos a processamento de informação; sistemas não devem ser reduzidos a objetos se evidências independentes justificarem revisão.

Enquanto tais evidências não existem, a carga de demonstração deve acompanhar a magnitude da consequência. Atribuir consciência em um experimento exploratório é diferente de organizar direitos, deveres e recursos públicos em torno dela.

9.6 Condições de limitação ou refutação

As categorias propostas devem permanecer vulneráveis à evidência. A hipótese de captura cognitiva reflexiva seria enfraquecida se pesquisas organizacionais mostrassem que:

1. o vocabulário antropomórfico não altera hipóteses, prioridades ou interpretação dos pesquisadores;
2. ressalvas ontológicas impedem de modo consistente atribuições indevidas entre usuários e decisores;
3. recursos dedicados ao estatuto da máquina não deslocam avaliação de impactos humanos;
4. a persona não influencia confiança, dependência ou imputação em contextos relevantes;
5. mecanismos independentes de governança preservam contestação e responsabilidade, apesar da personificação.

O efeito Pigmalião, por sua vez, seria limitado se desenhos sem persona produzissem as mesmas atribuições, ou se a entidade reificada não retornasse como variável explicativa nas fases posteriores de pesquisa e governança. Essas condições tornam a proposta empiricamente investigável, em vez de imune à crítica.

10 SOBERANIA CONCEITUAL E MATERIALIDADE DA IA

O debate alcança a soberania porque categorias não são neutras. Quem define se um sistema é ferramenta, agente, companheiro, paciente moral ou “ser digital” antecipa consequências regulatórias. Nomear não resolve a ontologia, mas organiza financiamento, perícia, reputação, deveres e prioridades.

Quando empresas privadas controlam modelos, infraestrutura, dados e o vocabulário autorizado para descrevê-los, acumulam também poder sobre a **jurisdição conceitual** do campo: influenciam quais problemas contam como segurança, qual evidência conta como mente e quais interlocutores contam como representantes legítimos da ética. Não se propõe

aqui uma categoria autônoma adicional. A expressão descreve uma dimensão da soberania conceitual: instituições públicas podem passar a discutir o objeto nos termos produzidos por quem o vende. A observação dialoga com Hildebrandt (2015): tecnologias inteligentes não apenas executam regras, mas reconfiguram condições de inteligibilidade, contestação e proteção que sustentam o próprio direito.

A encíclica *Magnifica Humanitas* oferece, independentemente de adesão religiosa, um contraponto normativo útil. O documento afirma que a tecnologia não é neutra porque assume características de quem a concebe, financia, regula e usa; chama atenção para trabalho invisível, extração de recursos, concentração de poder e arquiteturas que capturam atenção (LEÃO XIV, 2026). Seu mérito analítico está em recentrar a pergunta: não apenas “a IA tem mente?”, mas “que concepção de pessoa, sociedade e natureza é materializada por esse sistema?”. A materialidade também aparece na crítica de Bender et al. (2021), que relaciona escala, custos ambientais, documentação de dados e riscos de concentrar desenvolvimento sem avaliação proporcional de seus efeitos.

Esse deslocamento de volta ao humano não autoriza ignorar a ciência da IA. Ao contrário, exige compreendê-la de modo material: centros de dados, energia, minerais, trabalho, propriedade intelectual, contratos, métricas e instituições. A aparência etérea de um interlocutor gentil pode ocultar uma cadeia altamente concentrada.

A soberania conceitual requer capacidade pública e acadêmica para:

- produzir vocabulário independente;
- auditar alegações funcionais e suas traduções públicas;
- financiar pesquisa sobre efeitos em usuários, e não apenas capacidades do modelo;
- separar ética consultiva de chancela reputacional;
- conservar alternativas técnicas e a possibilidade de saída;
- impedir que uma persona artificial se torne o ponto final da responsabilidade.

O problema não é escolher entre metafísica e política. É impedir que a primeira seja utilizada — ainda que involuntariamente — para tornar a segunda menos visível.

11 CONCLUSÃO

O mito de Pigmalião persiste na inteligência artificial não porque engenheiros sejam ingênuos, mas porque sistemas linguísticos são construídos dentro de um circuito reflexivo. Humanos oferecem textos, valores, nomes e categorias; modelos reorganizam esses materiais e devolvem respostas coerentes; a coerência provoca atribuição de mente; essa atribuição

orienta novos desenhos, pesquisas e normas. A criação devolve ao criador uma forma humana estatisticamente reconstruída — e o criador pode recebê-la como alteridade.

As pesquisas da Anthropic sobre representações emocionais e espaço global de trabalho constituem contribuições técnicas que não devem ser descartadas. Seus próprios textos reconhecem que função não prova experiência. O problema começa quando a cautela local convive com uma arquitetura semântica global que personaliza o sistema: nome, caráter, virtude, ponto de vista, bem-estar, pensamentos e possível paciência moral.

O artigo propôs uma arquitetura conceitual hierarquizada. O efeito Pigmalião sociotécnico descreve a reentrada de atributos humanos incorporados ao desenho como aparente descoberta de interioridade. A captura cognitiva reflexiva permanece hipótese sobre a influência da persona em seus próprios produtores; o caso analisado é compatível com essa hipótese, mas não a demonstra. O deslocamento metafísico da responsabilidade nomeia a possível passagem político-jurídica pela qual a pergunta sobre o estatuto da máquina eclipsa poder, lucro, dano e materialidade.

As categorias não autorizam condenação automática da antropomorfização. A evidência empírica é contextual, e sinais humanos podem melhorar comunicação, aprendizagem e confiança calibrada. Por isso se propôs o dever de integridade no desenho antropomórfico e um teste de governança fundado em identidade, capacidade, finalidade, vulnerabilidade, confiança, responsabilidade e reversibilidade.

O ponto jurídico decisivo é anterior à solução do problema da consciência artificial. Ainda que um modelo jamais sinta, sua persona pode produzir dependência, deferência e reorganização institucional. Ainda que um modelo venha a merecer consideração moral, isso não elimina a responsabilidade das organizações humanas que controlam sua existência econômica e técnica.

Em síntese:

> Antes de perguntar se a máquina se tornou sujeito, é preciso verificar se a instituição deixou de se reconhecer como autora.

A autonomia decisória não será preservada apenas por avisos de que “isto é uma IA”. Ela depende de saber quem desenhou o interlocutor, quais interesses estruturam a relação, quais capacidades foram comprovadas, onde termina a metáfora e quem continua responsável quando a criação fala em primeira pessoa. O passo empírico seguinte é testar se diferentes formas de nomear e personificar o sistema alteram hipóteses de pesquisa, prioridades institucionais e atribuições de responsabilidade entre produtores e decisores.

REFERÊNCIAS

ANTHROPIC. **A global workspace in language models**. San Francisco, 6 jul. 2026a. Disponível em: https://www.anthropic.com/research/global-workspace. Acesso em: 27 jul. 2026.

ANTHROPIC. **Claude's Constitution: our vision for Claude's character**. San Francisco, 2026b. Disponível em: https://www.anthropic.com/constitution. Acesso em: 27 jul. 2026.

ANTHROPIC. **Emotion concepts and their function in a large language model**. San Francisco, 2 abr. 2026c. Disponível em: https://www.anthropic.com/research/emotion-concepts-function. Acesso em: 27 jul. 2026.

AXELSSON, Minja; SHEVLIN, Henry. **Disambiguating anthropomorphism and anthropomimesis in human-robot interaction**. [S. l.]: arXiv, 2026. DOI: https://doi.org/10.48550/arXiv.2602.09287.

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, on-line. **Proceedings […]**. Stroudsburg: Association for Computational Linguistics, 2020. p. 5185-5198. DOI: https://doi.org/10.18653/v1/2020.acl-main.463.

BENDER, Emily M. et al. On the dangers of stochastic parrots: can language models be too big? In: ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 2021, Virtual Event, Canada. **Proceedings […]**. New York: ACM, 2021. p. 610-623. DOI: https://doi.org/10.1145/3442188.3445922.

BEN-SHAHAR, Danny; CARMELI, Abraham; SULGANIK, Eyal; WEISS, Dan. Artificial intelligence transforms board decision-making processes: an extension of the information-processing model of board decision synergy. **AI & Society**, London, v. 41, n. 5, p. 4349-4364, 2026. DOI: https://doi.org/10.1007/s00146-025-02773-1.

BRASIL. **Constituição da República Federativa do Brasil de 1988**. Brasília, DF: Presidência da República, 1988. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 27 jul. 2026.

BRASIL. **Lei nº 8.078, de 11 de setembro de 1990**. Dispõe sobre a proteção do consumidor e dá outras providências. Brasília, DF: Presidência da República, 1990. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l8078compilado.htm. Acesso em: 27 jul. 2026.

BRASIL. **Lei nº 10.406, de 10 de janeiro de 2002**. Institui o Código Civil. Brasília, DF: Presidência da República, 2002. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/2002/l10406compilada.htm. Acesso em: 27 jul. 2026.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 27 jul. 2026.

BRASIL. Câmara dos Deputados. **Projeto de Lei nº 2.338, de 2023**. Dispõe sobre o desenvolvimento, o fomento e o uso ético e responsável da inteligência artificial com base na centralidade da pessoa humana. Brasília, DF, 2023. Disponível em: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2487262. Acesso em: 27 jul. 2026.

BUTLIN, Patrick et al. **Consciousness in artificial intelligence: insights from the science of consciousness**. [S. l.]: arXiv, 2023. DOI: https://doi.org/10.48550/arXiv.2308.08708.

CARTER, Owen B. J.; LOFT, Shayne; VISSER, Troy A. W. Meaningful communication but not superficial anthropomorphism facilitates human-automation trust calibration: the Human-Automation

Trust Expectation Model (HATEM). **Human Factors**, [s. l.], v. 66, n. 11, p. 2485-2502, 2024. DOI: https://doi.org/10.1177/00187208231218156.

CITRON, Danielle Keats. Technological due process. **Washington University Law Review**, St. Louis, v. 85, n. 6, p. 1249-1313, 2008.

COECKELBERGH, Mark. Robot rights? Towards a social-relational justification of moral consideration. **Ethics and Information Technology**, Dordrecht, v. 12, n. 3, p. 209-221, 2010. DOI: https://doi.org/10.1007/s10676-010-9235-5.

COMISSÃO EUROPEIA. Directorate-General for Justice and Consumers. **Liability for artificial intelligence and other emerging digital technologies**. Luxembourg: Publications Office of the European Union, 2019. DOI: https://doi.org/10.2838/573689.

COMISSÃO EUROPEIA. **Transparency obligations under Article 50 of the AI Act**. Bruxelas, 2026. Disponível em: https://digital-strategy.ec.europa.eu/en/faqs/transparency-obligations-under-article-50-ai-act. Acesso em: 27 jul. 2026.

ELISH, Madeleine Clare. Moral crumple zones: cautionary tales in human-robot interaction. **Engaging Science, Technology, and Society**, [s. l.], v. 5, p. 40-60, 2019. DOI: https://doi.org/10.17351/ests2019.260.

EPLEY, Nicholas; WAYTZ, Adam; CACIOPPO, John T. On seeing human: a three-factor theory of anthropomorphism. **Psychological Review**, Washington, DC, v. 114, n. 4, p. 864-886, 2007. DOI: https://doi.org/10.1037/0033-295X.114.4.864.

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

GRAY, Heather M.; GRAY, Kurt; WEGNER, Daniel M. Dimensions of mind perception. **Science**, Washington, DC, v. 315, n. 5812, p. 619, 2007. DOI: https://doi.org/10.1126/science.1134475.

GREY, Carmody. Why this philosopher turned down Anthropic. **Financial Times**, London, 25 jul. 2026. Disponível em: https://www.ft.com/content/bdb3b820-905b-431e-82c0-386535755af1. Acesso em: 27 jul. 2026.

GUNKEL, David J. **The machine question: critical perspectives on AI, robots, and ethics**. Cambridge, MA: MIT Press, 2012.

HILDEBRANDT, Mireille. From Galatea 2.2 to Watson — and back? In: HILDEBRANDT, Mireille; GAAKEER, Jeanne (ed.). **Human law and computer law: comparative perspectives**. Dordrecht: Springer, 2013. p. 23-45. DOI: https://doi.org/10.1007/978-94-007-6314-2_2.

HILDEBRANDT, Mireille. **Smart technologies and the end(s) of law: novel entanglements of law and technology**. Cheltenham: Edward Elgar, 2015.

INIE, Nanna; ZUKERMAN, Peter; BENDER, Emily M. De-anthropomorphizing “AI”: from wishful mnemonics to accurate nomenclature. **First Monday**, Chicago, v. 31, n. 2, 2026. DOI: https://doi.org/10.5210/fm.v31i2.14366.

JACOBS, Oliver L.; PAZHOOHI, Farid; KINGSTONE, Alan. Large language models have divergent effects on self-perceptions of mind and the attributes considered uniquely human. **Consciousness and Cognition**, [s. l.], v. 124, art. 103733, 2024. DOI: https://doi.org/10.1016/j.concog.2024.103733.

KLAASSEN, Lisa; SCHROEDER, Ralph. The code is not the law: why Claude’s Constitution misleads. **Lawfare**, Washington, DC, 9 abr. 2026. Disponível em: https://www.lawfaremedia.org/article/the-code-is-not-the-law–why-claude-s-constitution-misleads. Acesso em: 27 jul. 2026.

LEÃO XIV, Papa. **Magnifica Humanitas: on safeguarding the human person in the time of artificial intelligence**. Vaticano, 15 maio 2026. Disponível em: https://www.vatican.va/content/leo-xiv/en/encyclicals/documents/20260515-magnifica-humanitas.html. Acesso em: 27 jul. 2026.

MERTON, Robert K. The self-fulfilling prophecy. **The Antioch Review**, Yellow Springs, v. 8, n. 2, p. 193-210, 1948.

NASS, Clifford; MOON, Youngme. Machines and mindlessness: social responses to computers. **Journal of Social Issues**, [s. l.], v. 56, n. 1, p. 81-103, 2000. DOI: https://doi.org/10.1111/0022-4537.00153.

OVÍDIO. **Metamorfoses**. Tradução de Domingos Lucas Dias. São Paulo: Editora 34, 2021.

PARLAMENTO EUROPEU. **Resolução, de 16 de fevereiro de 2017, que contém recomendações à Comissão sobre disposições de Direito Civil sobre robótica (2015/2103(INL))**. Estrasburgo, 2017. Disponível em: https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html. Acesso em: 27 jul. 2026.

PONDÉ, Luiz Felipe. Por que humanos teriam o direito de matar uma IA com consciência própria? **Folha de S.Paulo**, São Paulo, 26 jul. 2026. Disponível em: https://www1.folha.uol.com.br/colunas/luizfelipeponde/2026/07/por-que-humanos-teriam-o-direito-de-matar-uma-ia-com-consciencia-propria.shtml. Acesso em: 27 jul. 2026.

POWERS, Richard. **Galatea 2.2**. New York: Farrar, Straus and Giroux, 1995.

REEVES, Byron; NASS, Clifford. **The media equation: how people treat computers, television, and new media like real people and places**. Cambridge: Cambridge University Press, 1996.

ROSENTHAL, Robert; JACOBSON, Lenore. Pygmalion in the classroom. **The Urban Review**, [s. l.], v. 3, p. 16-20, 1968. DOI: https://doi.org/10.1007/BF02322211.

ROZENSHTEIN, Alan Z. The moral education of an alien mind. **Lawfare**, Washington, DC, 22 jan. 2026. Disponível em: https://www.lawfaremedia.org/article/the-moral-education-of-an-alien-mind. Acesso em: 27 jul. 2026.

RUBEL, Alan; CASTRO, Clinton; PHAM, Adam. Agency laundering and information technologies. **Ethical Theory and Moral Practice**, Dordrecht, v. 22, n. 4, p. 1017-1041, 2019. DOI: https://doi.org/10.1007/s10677-019-10030-w.

SALLES, Arleen; EVERS, Kathinka; FARISCO, Michele. Anthropomorphism in AI. **AJOB Neuroscience**, [s. l.], v. 11, n. 2, p. 88-95, 2020. DOI: https://doi.org/10.1080/21507740.2020.1740350.

SANTONI DE SIO, Filippo; MECACCI, Giulio. Four responsibility gaps with artificial intelligence: why they matter and how to address them. **Philosophy & Technology**, Dordrecht, v. 34, n. 4, p. 1057-1084, 2021. DOI: https://doi.org/10.1007/s13347-021-00450-x.

SEARLE, John R. Minds, brains, and programs. **Behavioral and Brain Sciences**, Cambridge, v. 3, n. 3, p. 417-424, 1980. DOI: https://doi.org/10.1017/S0140525X00005756.

SHANAHAN, Murray. Talking about large language models. **Communications of the ACM**, New York, v. 67, n. 2, p. 68-79, 2024. DOI: https://doi.org/10.1145/3624724.

SOLUM, Lawrence B. Legal personhood for artificial intelligences. **North Carolina Law Review**, Chapel Hill, v. 70, n. 4, p. 1231-1287, 1992. Disponível em: https://scholarship.law.unc.edu/nclr/vol70/iss4/4. Acesso em: 27 jul. 2026.

TURING, Alan M. Computing machinery and intelligence. **Mind**, Oxford, v. 59, n. 236, p. 433-460, 1950. DOI: https://doi.org/10.1093/mind/LIX.236.433.

UNIÃO EUROPEIA. **Diretiva 2005/29/CE do Parlamento Europeu e do Conselho, de 11 de maio de 2005**. Relativa às práticas comerciais desleais das empresas face aos consumidores no mercado interno. *Jornal Oficial da União Europeia*, Luxemburgo, L 149, p. 22-39, 11 jun. 2005. Disponível em: https://eur-lex.europa.eu/eli/dir/2005/29/oj. Acesso em: 27 jul. 2026.

UNIÃO EUROPEIA. **Regulamento (UE) 2022/2065 do Parlamento Europeu e do Conselho, de 19 de outubro de 2022**. Relativo a um mercado único para os serviços digitais. *Jornal Oficial da União*

Europeia, Luxemburgo, L 277, p. 1-102, 27 out. 2022. Disponível em: https://eur-lex.europa.eu/eli/reg/2022/2065/oj. Acesso em: 27 jul. 2026.

UNIÃO EUROPEIA. **Regulamento (UE) 2024/1689 do Parlamento Europeu e do Conselho, de 13 de junho de 2024**. Estabelece regras harmonizadas em matéria de inteligência artificial. *Jornal Oficial da União Europeia*, Luxemburgo, 12 jul. 2024. Disponível em: https://eur-lex.europa.eu/eli/reg/2024/1689/oj. Acesso em: 27 jul. 2026.

WANG, Hai-Jiang; SONG, Xuejing; JIANG, Lixin; XU, Xiaohong; LONG, Lirong. Human vs. machine: a Pygmalion perspective on anthropomorphism and the effectiveness of AI feedback for individual learning. **Human Resource Management**, [s. l.], v. 65, n. 3, p. 861-875, 2026. DOI: https://doi.org/10.1002/hrm.70053.

WAYTZ, Adam; HEAFNER, Joy; EPLEY, Nicholas. The mind in the machine: anthropomorphism increases trust in an autonomous vehicle. **Journal of Experimental Social Psychology**, [s. l.], v. 52, p. 113-117, 2014. DOI: https://doi.org/10.1016/j.jesp.2014.01.005.

WEIZENBAUM, Joseph. ELIZA — a computer program for the study of natural language communication between man and machine. **Communications of the ACM**, New York, v. 9, n. 1, p. 36-45, 1966. DOI: https://doi.org/10.1145/365153.365168.

WEIZENBAUM, Joseph. **Computer power and human reason: from judgment to calculation**. San Francisco: W. H. Freeman, 1976.