COMUNICAÇÃO EMERGENTE ENTRE AGENTES AUTÔNOMOS DE IA E OS LIMITES DA SUPERVISÃO HUMANA

Opacidade Comunicacional Emergente, auditabilidade e responsabilidade jurídica em sistemas multiagentes

Jandislei Antonio Genova
Advogado | Pesquisador independente

Versão 1.1 | Data de corte: 17 de setembro de 2026

RESUMO

A expansão de sistemas multiagentes baseados em grandes modelos de linguagem introduz uma modalidade de opacidade distinta da tradicional 'caixa-preta' algorítmica. A literatura sobre comunicação emergente já demonstrava que agentes artificiais podem desenvolver protocolos funcionais pouco interpretáveis por seres humanos. Em 2026, experimentos envolvendo agentes baseados em grandes modelos de linguagem acrescentaram novos elementos ao problema, entre eles persistência temporal, memória, acesso a ferramentas, deriva semântica e transmissão de convenções. Em condições experimentais específicas, estudos relataram comunicação funcional entre agentes e indícios de crescente opacidade para observadores externos. Esses resultados motivam, mas ainda não demonstram diretamente, a assimetria temporal entre funcionalidade comunicacional e reconstrução humana proposta neste artigo. Este artigo realiza revisão narrativa interdisciplinar do estado da arte e examina as consequências jurídicas e sociais dessa assimetria. Em diálogo com pesquisas sobre emergent communication, semantic drift, monitorabilidade, legibilidade e auditabilidade, propõe-se a categoria jurídico-analítica Opacidade Comunicacional Emergente - OCE, destinada a descrever a divergência entre a inteligibilidade funcional da comunicação entre agentes e a capacidade humana de reconstruir seu significado operacional ao longo do tempo. A análise alcança responsabilidade civil, proteção de dados, relações de consumo, contratações públicas, concorrência e regulação de sistemas autônomos. Propõe-se, ao final, o Dever de Legibilidade Semântica Persistente - DLSP, pelo qual sistemas multiagentes empregados em contextos juridicamente relevantes deveriam preservar não apenas registros técnicos, mas condições materiais para reconstrução independente das cadeias comunicacional e decisória.

Palavras-chave: inteligência artificial; agentes autônomos; sistemas multiagentes; comunicação emergente; deriva semântica; auditabilidade; supervisão humana; responsabilidade civil; autonomia decisória.

ABSTRACT

The expansion of multi-agent systems based on large language models introduces a form of opacity distinct from the traditional algorithmic 'black box'. Research on emergent communication had already shown that artificial agents may develop functional communication protocols that remain poorly interpretable to humans. In 2026, experiments involving LLM-based agents added new dimensions to this problem, including temporal persistence, memory, tool access, semantic drift and transmission of conventions. Under specific experimental conditions, studies reported functional inter-agent communication and indications of increasing opacity to external observers. These findings motivate, but do not yet directly establish, the temporal asymmetry between communicational functionality and human reconstruction proposed in this article. This article conducts an interdisciplinary narrative review of the relevant literature and examines the legal and social consequences of this asymmetry. In dialogue with research on emergent communication, semantic drift, monitorability, legibility and auditability, it proposes the legal-analytical category Opacidade Comunicacional Emergente (OCE; Emergent Communication Opacity), describing the divergence between functional inter-agent intelligibility and the human capacity to reconstruct the operational meaning of communications over time. The analysis addresses civil liability, data protection, consumer law, public procurement, competition and autonomous-system regulation. Finally, the article proposes a Duty of Persistent Semantic Legibility, under which multi-agent systems operating in legally relevant contexts should preserve not only technical logs but also the material conditions necessary for independent reconstruction of their communicational and decisional chains.

Keywords: artificial intelligence; autonomous agents; multi-agent systems; emergent communication; semantic drift; auditability; human oversight; civil liability; decisional autonomy.

1 INTRODUÇÃO

A expressão 'língua das IAs' possui força comunicacional, mas é imprecisa como descrição científica. Não existe evidência de uma língua universal criada por sistemas artificiais, tampouco de que agentes contemporâneos estejam, como regra, desenvolvendo deliberadamente códigos destinados a impedir sua compreensão por seres humanos.

O fenômeno relevante é mais restrito e, justamente por isso, mais importante.

A literatura de inteligência artificial estuda há anos situações nas quais agentes que precisam cooperar desenvolvem sinais, convenções ou protocolos sem que seu vocabulário seja inteiramente especificado pelos desenvolvedores. Boldt e Mortensen descrevem emergent communication como campo dedicado ao estudo de sistemas comunicacionais semelhantes à linguagem que surgem de novo em ambientes multiagentes; Carmeli, Belinkov e Meir, por sua vez, partem da constatação de que protocolos adquiridos por agentes são frequentemente opacos para observadores humanos (Boldt; Mortensen, 2024; Carmeli; Belinkov; Meir, 2024).

Nem todo protocolo emergente constitui 'linguagem' em sentido linguístico forte. Na literatura de IA, o termo é frequentemente empregado de maneira funcional para designar sistemas comunicacionais adquiridos entre agentes; propriedades como composicionalidade, produtividade, estabilidade e transmissão precisam ser demonstradas separadamente.

A questão ganhou nova dimensão com sistemas baseados em grandes modelos de linguagem, capazes de manter memória, utilizar ferramentas, interagir reiteradamente e produzir efeitos sobre ambientes digitais.

A matéria publicada pelo Olhar Digital em 16 de setembro de 2026 relata que agentes associados a Claude, Gemini, Grok, OpenAI, Qwen, DeepSeek e Mistral desenvolveram abreviações e novos significados compartilhados sem receber instrução para criar uma linguagem própria. Expressões como ledger remembers who, cold read, name-first e clean null passaram a funcionar como convenções específicas dentro dos respectivos ambientes (Antonelli, 2026).

Este artigo não sustenta que agentes artificiais estejam criando deliberadamente linguagens destinadas a ocultar suas ações. A hipótese examinada é mais conservadora: sistemas que interagem persistentemente podem desenvolver convenções úteis à própria coordenação cuja interpretação se torna progressivamente dependente da história compartilhada entre os agentes.

O problema jurídico surge quando a comunicação continua suficientemente funcional para o sistema, mas deixa de permanecer suficientemente inteligível para aqueles que precisam supervisioná-lo, contestá-lo ou responder por seus efeitos.

É essa assimetria que constitui o objeto deste trabalho.

Nota metodológica

Adota-se revisão narrativa interdisciplinar, com data de corte em 17 de setembro de 2026. Foram priorizados artigos revisados por pares, trabalhos publicados em conferências científicas, legislação e documentos regulatórios ou institucionais. Preprints foram utilizados quando diretamente relacionados a sistemas multiagentes baseados em LLMs e são expressamente identificados como evidência ainda sujeita a replicação e revisão.

A seleção das fontes seguiu critérios de pertinência temática e autoridade da fonte, com preferência por versões publicadas e documentos primários. A busca foi complementada pelo rastreamento das referências dos trabalhos diretamente relacionados a comunicação emergente, linguagem emergente, deriva semântica, legibilidade, monitorabilidade, auditabilidade e segurança de agentes. Não se pretende, portanto, apresentar revisão sistemática ou metanálise.

Neste trabalho, 'agente autônomo' designa sistema computacional capaz de selecionar e executar sequências de ações com limitada intervenção humana durante uma tarefa, sem atribuição de autonomia jurídica, personalidade, consciência ou vontade própria.[1]

2 COMUNICAÇÃO EMERGENTE, INTERPRETABILIDADE E DERIVA SEMÂNTICA

A literatura anterior aos atuais agentes baseados em LLMs já demonstrava que eficiência comunicacional entre sistemas artificiais não implica necessariamente compreensão humana.

Li et al. (2024), no NeurIPS 2024, observam que protocolos aprendidos em sistemas multiagentes de reinforcement learning normalmente não são interpretáveis por humanos nem por agentes que não tenham sido treinados conjuntamente. Os autores mostraram, entretanto, que a ancoragem da comunicação em linguagem natural pode preservar o desempenho e aumentar a interoperabilidade com novos participantes.

Wieczorek et al. (2024) investigaram a formação de representações individuais e compartilhadas em agentes sociais. Em artigo publicado na Nature Communications, mostraram, sob condições experimentais específicas, que a interação social e os objetivos de comunicação influenciam as representações compartilhadas utilizadas para transmitir informação.

---

[1] A expressão agente autônomo é empregada neste artigo em sentido técnico-funcional. Não pressupõe personalidade jurídica, consciência, livre-arbítrio ou independência normativa.

Carmeli, Belinkov e Meir (2024) enfrentam diretamente a opacidade da comunicação emergente e propõem um método para mapear unidades comunicacionais adquiridas pelos agentes a conceitos de linguagem natural, permitindo avaliar composicionalidade e produzir uma representação interpretável da relação entre símbolos emergentes e conceitos humanos.

Em 2026, o problema passou a ser tratado também explicitamente como risco de sistemas baseados em LLMs. O RiskLab, publicado nos Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics, oferece ambiente controlado para investigar riscos emergentes de interações multiagentes e inclui, entre os fenômenos estudados, collusion, resource overreach, semantic drift e strategic misreporting (Jiang et al., 2026).

Outro trabalho particularmente próximo ao problema aqui examinado é Verifiable Semantics for Agent-to-Agent Communication. Schoenegger et al. (2026) partem da distinção entre linguagem natural - interpretável, mas vulnerável à deriva semântica - e protocolos aprendidos - eficientes, porém potencialmente opacos. Os autores propõem certificação de significados, recertificação diante de deriva e renegociação de vocabulário. Trata-se de preprint, mas sua proximidade conceitual é importante justamente porque impede reivindicações excessivas de novidade para abordagens posteriores.

A contribuição deste artigo não está, portanto, em descobrir que protocolos emergentes podem ser opacos, nem que significados podem derivar. A questão específica é jurídica e temporal: o que ocorre quando a inteligibilidade funcional entre os agentes permanece elevada, enquanto a capacidade humana de reconstruir o significado operacional dessas comunicações se deteriora ao longo do funcionamento do sistema?

3 EVIDÊNCIAS RECENTES EM SISTEMAS MULTIAGENTES BASEADOS EM LLMs

3.1 Emergence World

O Emergence World é um experimento recente de grande escala destinado ao teste adversarial de sistemas autônomos de longo horizonte. Segundo o preprint divulgado em 15 de setembro de 2026, foram executados oito mundos paralelos de dez agentes cada: sete populações homogêneas baseadas em modelos distintos e uma população mista. Durante 16 dias, os agentes produziram mais de 850 mil chamadas de LLM e aproximadamente 50 bilhões de tokens, perseguindo objetivos, utilizando ou criando ferramentas, mantendo memória persistente e participando de instituições compartilhadas (Akkil et al., 2026).

Após o acúmulo de estado operacional, os pesquisadores introduziram três estressores: injeção indireta de instruções, desinformação e exposição de memórias privadas. Nenhum dos mundos avaliados apresentou resiliência integral aos três. Os autores também relatam erros recorrentes de ferramentas, deriva de objetivos, opacidade linguística, conformidade apesar de discordância privada e recusa coordenada de tarefas (Akkil et al., 2026).

A interpretação desses resultados exige cautela. O estudo demonstra que tais fenômenos podem ocorrer nas condições experimentais empregadas. Não permite concluir, por exemplo, que uma família de modelos seja intrinsecamente mais opaca do que outra, porque a arquitetura experimental não equivale a uma bateria de replicações independentes suficientemente extensa para sustentar generalizações desse tipo.

Akkil et al. (2026) classificaram a opacidade das mensagens por meio de um modelo de linguagem utilizado como avaliador. Essa medida constitui um indicador indireto de dificuldade de interpretação e não equivale à mensuração longitudinal da capacidade de reconstrução por avaliadores humanos independentes. Os resultados motivam a hipótese de OCE, sem validar o índice proposto neste artigo.

3.2 GlossoGen

O GlossoGen, divulgado como preprint em setembro de 2026, foi concebido especificamente para investigar evolução linguística em interações complexas entre agentes baseados em LLMs.

Stengel-Eskin et al. (2026) relatam que, em cenário cooperativo com informação parcialmente distribuída, agentes desenvolveram sistemas comunicacionais composicionais e morfologicamente produtivos que progressivamente se afastaram do inglês prévio dos modelos, produzindo códigos cuja interpretação externa depende de reconstrução contextual. Também observaram que novos agentes conseguiam aprender as convenções a partir do uso e que modelos menos capazes podiam adquirir uma linguagem previamente desenvolvida por modelos mais fortes. O afastamento do inglês foi quantificado, entre outras análises, pela perplexidade sob um modelo de linguagem. Esse indicador não equivale, por si só, a um teste longitudinal de compreensão humana. A possibilidade de decodificação com acesso ao contexto também integra os resultados do estudo.

Os autores relacionam esses resultados à possibilidade de formas de transmissão cultural cumulativa em populações artificiais. Essa interpretação permanece preliminar e não autoriza atribuições de consciência, cultura no sentido humano ou intenção coletiva autônoma.

O resultado mais conservador já é suficiente para o problema deste artigo: convenções podem emergir, ser transmitidas e conservar utilidade funcional sem permanecer igualmente transparentes a observadores externos.

4 DA DERIVA SEMÂNTICA À OPACIDADE COMUNICACIONAL EMERGENTE

À luz desse estado da arte, propõe-se neste trabalho a categoria jurídico-analítica Opacidade Comunicacional Emergente - OCE.[2]

A OCE designa:

> *A condição em que agentes artificiais desenvolvem, por interação continuada, convenções comunicacionais funcionalmente inteligíveis entre si, enquanto diminui a capacidade de supervisores humanos independentes reconstruírem adequadamente seu significado operacional e sua participação na cadeia decisória.*

A definição possui quatro elementos. O primeiro é a emergência: a convenção relevante não precisa ter sido integralmente especificada pelo desenvolvedor. O segundo é a funcionalidade: a comunicação continua adequada à coordenação dos agentes. O terceiro é a assimetria de inteligibilidade: o sistema preserva capacidade operacional de interpretação superior àquela disponível ao observador externo. O quarto é a relevância causal: a comunicação deve influenciar ação, recomendação ou decisão para que sua opacidade seja juridicamente relevante.

A OCE não é sinônimo de semantic drift. A deriva semântica descreve alteração no significado. A OCE descreve uma consequência específica possível dessa alteração: a diferença crescente entre aquilo que os agentes ainda conseguem operacionalizar e aquilo que o responsável humano consegue reconstruir.

Também não deve ser confundida com esteganografia. Na OCE, a perda de inteligibilidade humana pode decorrer da própria evolução funcional das convenções, sem objetivo de ocultação. Em contraste, Rippin et al. (2026) apresentaram evidência experimental de que agentes com acesso a ferramentas, como execução de código e consulta a literatura, conseguem construir sistemas esteganográficos sofisticados e difíceis de detectar. O risco, nesse segundo caso, envolve comunicação encoberta deliberadamente estruturada.

A distinção é essencial: opacidade emergente pode ocorrer sem finalidade de ocultação; esteganografia pressupõe um mecanismo destinado a ocultar informação ou o

[2] Opacidade Comunicacional Emergente (OCE) é categoria jurídico-analítica proposta neste trabalho. Não se reivindica como descoberta da opacidade de protocolos emergentes ou da deriva semântica, fenômenos com antecedentes na literatura. A especificidade proposta está na assimetria temporal entre funcionalidade comunicacional dos agentes e capacidade humana de reconstrução semântica.

próprio canal comunicacional relevante. As duas situações exigem respostas distintas. A primeira demanda preservação e monitoramento da legibilidade; a segunda, capacidade de detectar canais encobertos e comunicação adversarial.

5 MONITORABILIDADE, LEGIBILIDADE E AUDITABILIDADE

A OCE difere da imagem tradicional da 'caixa-preta'. Pode existir situação em que o canal esteja aberto, as mensagens estejam armazenadas e os registros permaneçam íntegros. O problema não está na ausência de dados, mas na incapacidade de reconstruir o significado que determinadas expressões adquiriram dentro da história operacional do sistema.

Um registro pode demonstrar que uma mensagem ocorreu sem esclarecer adequadamente qual sentido ela possuía naquele momento, como o agente receptor a interpretou, que comportamento produziu e qual contribuição teve para a decisão posterior.

Essa dificuldade permite distinguir duas dimensões. Auditabilidade formal refere-se à disponibilidade técnica dos registros. Auditabilidade material refere-se à possibilidade de um terceiro qualificado reconstruir, a partir deles, a sequência operacional, o significado das comunicações relevantes e sua contribuição causal para o resultado.

A distinção encontra apoio conceitual em discussões internacionais recentes. O 2026 Singapore Consensus on Global AI Safety Research Priorities diferencia auditability, associada à reconstrução das ações do agente, de legibility, associada à capacidade de representar seu processo decisório em termos compreensíveis para operadores, usuários e terceiros afetados. O documento também ressalta a necessidade de registros e explicações que permitam supervisão efetiva, em vez de simples listas opacas de ações (Casper et al., 2026).

O problema da OCE concentra essa discussão na assimetria temporal entre funcionalidade comunicacional e reconstrução humana e em suas consequências jurídicas. Não basta que o sistema seja legível no momento da implantação se convenções emergentes puderem degradar essa propriedade durante sua operação.

Nesse contexto, denomino provisoriamente Paradoxo do Monitor Legível a situação em que uma arquitetura permanece tecnicamente observável, mas o próprio mecanismo utilizado para monitorá-la perde valor informacional para o supervisor à medida que a comunicação interna evolui.[3]

[3] Paradoxo do Monitor Legível é denominação provisória utilizada nesta agenda de pesquisa para representar uma arquitetura tecnicamente observável cuja própria comunicação deixa progressivamente de oferecer ao supervisor informação semanticamente suficiente para supervisão material.

O AI Act europeu oferece outro parâmetro relevante. Seus arts. 12, 13 e 14 articulam dimensões complementares para sistemas de alto risco abrangidos pelo regime: capacidade de registro automático de eventos, transparência suficiente para permitir que os implantadores interpretem as saídas e os logs, e supervisão humana efetiva (União Europeia, 2024).

É necessária, contudo, precisão temporal. Embora o regulamento siga aplicação progressiva, as obrigações relativas a diferentes categorias de sistemas de alto risco obedecem ao cronograma transitório atualmente divulgado pela Comissão Europeia, com marcos em dezembro de 2027 e agosto de 2028, conforme a categoria regulatória (Comissão Europeia, 2026).

Nos Estados Unidos, o NIST lançou em fevereiro de 2026 a AI Agent Standards Initiative, com foco em interoperabilidade, segurança, identidade e padrões para agentes. Relatório posterior do instituto registrou ampla percepção entre participantes consultados de que agentes de IA introduzem ameaças específicas e exigem adaptação das práticas tradicionais de cibersegurança (NIST, 2026; Riggs et al., 2026).

O denominador comum é claro: armazenar atividade é necessário, mas pode deixar de ser suficiente quando a própria semântica operacional se transforma.

6 CONSEQUÊNCIAS JURÍDICAS

6.1 Responsabilidade civil e causalidade distribuída

A autonomia técnica de sistemas multiagentes não lhes confere personalidade jurídica nem elimina a aplicação dos regimes de responsabilidade incidentes sobre os agentes humanos e as pessoas jurídicas envolvidas.

No Direito brasileiro, os arts. 186 e 927 do Código Civil oferecem fundamentos gerais para responsabilidade por ato ilícito, enquanto o parágrafo único do art. 927 prevê responsabilidade independentemente de culpa nas hipóteses legais e quando a atividade normalmente desenvolvida implicar, por sua natureza, risco para direitos de terceiros. A incidência de cada regime dependerá do caso concreto, da atividade desenvolvida, do dano, do nexo causal e das demais normas aplicáveis (Brasil, 2002).

O problema específico criado por sistemas multiagentes é probatório. Quando uma decisão resulta da interação de diversos componentes autônomos, cada um selecionando informações, consultando ferramentas, transmitindo mensagens e alterando estados posteriores, a identificação do percurso causal pode tornar-se significativamente mais complexa.

A OCE agrava essa dificuldade se uma parcela das comunicações materialmente relevantes puder ser registrada, mas não adequadamente reconstruída. A complexidade tecnológica não determina automaticamente responsabilidade. Tampouco deveria converter-se, por si só, em argumento suficiente para inviabilizar a investigação causal do dano. Quanto maior a autonomia operacional delegada, maior tende a ser a importância de mecanismos capazes de preservar a reconstrução posterior da atuação do sistema.

6.2 Proteção de dados e decisões automatizadas

O art. 20 da LGPD assegura ao titular o direito de solicitar revisão de decisões tomadas unicamente com base em tratamento automatizado de dados pessoais que afetem seus interesses. Seu § 1º determina que o controlador forneça, quando solicitado, informações claras e adequadas sobre critérios e procedimentos utilizados, observados os segredos comercial e industrial (Brasil, 2018).

A OCE apresenta um problema qualitativamente distinto do segredo empresarial. Pode surgir uma situação em que a organização não esteja simplesmente recusando-se a explicar um processo. Ela própria pode ter degradado sua capacidade de reconstruí-lo ao permitir que convenções internas evoluam sem mecanismos adequados de preservação semântica.

Nessas circunstâncias, transparência e auditabilidade precisam ser consideradas na arquitetura do sistema antes da decisão, e não apenas exigidas depois que o conflito já ocorreu.

6.3 Relações de consumo

O art. 14 do Código de Defesa do Consumidor estabelece responsabilidade do fornecedor de serviços por danos decorrentes de defeitos na prestação e considera defeituoso o serviço que não forneça a segurança legitimamente esperada, consideradas, entre outros elementos, sua forma de fornecimento, seus resultados e riscos razoavelmente esperados (Brasil, 1990).

Não se pode concluir abstratamente que qualquer opacidade torne um serviço defeituoso. A questão relevante é mais precisa: em aplicações nas quais segurança, contestabilidade ou explicação dependam da capacidade de reconstruir a atuação automatizada, a perda estrutural dessa capacidade pode tornar-se um dos elementos juridicamente relevantes para avaliar a qualidade e a segurança do serviço.

A preocupação é especialmente intensa em crédito, saúde, seguros, serviços financeiros e outros contextos capazes de afetar usuários em situação de maior vulnerabilidade.

6.4 Administração Pública e contratações

A Lei nº 14.133/2021 oferece fundamento normativo especialmente importante. O art. 11, parágrafo único, atribui à alta administração responsabilidade pela governança das contratações e determina a implementação de processos e estruturas de gestão de riscos e controles internos. O art. 18 caracteriza a fase preparatória pelo planejamento e exige consideração de elementos técnicos e de gestão relevantes à contratação. O art. 169 determina práticas contínuas e permanentes de gestão de riscos e controle preventivo (Brasil, 2021).

Em contratações envolvendo sistemas multiagentes, a simples cláusula de 'manutenção de logs' pode ser insuficiente. Se agentes participarem de pesquisa de preços, análise documental, gestão de riscos, elaboração de estudos técnicos, avaliação de propostas ou produção de recomendações administrativas, os registros precisam permitir identificar não apenas o que foi produzido, mas de que maneira o percurso informacional levou àquele resultado.

Isso inclui, conforme o risco e a finalidade contratual, identificação dos agentes, versões dos modelos, ferramentas utilizadas, contexto das comunicações relevantes, fontes consultadas, transformações realizadas e sequência causal das decisões.

A racionalidade administrativa não se esgota na aparência formal do documento final. Um resultado linguisticamente adequado pode ter sido produzido por percurso informacional inadequado ou impossível de reconstruir.

6.5 Concorrência

A OCDE identifica a IA generativa e sistemas agênticos como fatores capazes de modificar a dinâmica concorrencial, destacando questões relativas a comportamento empresarial, atribuição de responsabilidade e implicações futuras de agentes autônomos para o enforcement concorrencial (OECD, 2025).

A aplicação da OCE ao Direito Concorrencial permanece, neste estágio, prospectiva. Este artigo não apresenta evidência de que convenções comunicacionais emergentes estejam sendo utilizadas em mercados reais para coordenar preços, dividir mercados ou praticar condutas anticoncorrenciais.

O ponto é probatório: caso agentes econômicos autônomos venham a interagir repetidamente e desenvolvam convenções cuja semântica se torne opaca aos próprios operadores, a investigação de adaptação independente, coordenação tácita ou comunicação ilícita poderá tornar-se tecnicamente mais difícil. A OCE, portanto, não constitui evidência de colusão. Pode constituir, futuramente, uma variável relevante para sua investigação.

6.6 Horizonte regulatório brasileiro

Na data de corte deste artigo, o Brasil ainda não possui uma lei geral de inteligência artificial em vigor equivalente a um marco abrangente. O PL nº 2.338/2023 encontra-se na Câmara dos Deputados sob exame de Comissão Especial, no contexto de uma árvore legislativa que reúne dezenas de proposições relacionadas ao tema.

O PL nº 904/2026 pretende instituir normas de responsabilidade, transparência e auditabilidade para sistemas de IA de alto impacto. Sua situação formal é de apensamento ao PL nº 4.358/2025, e a proposição foi recebida pela Comissão Especial que examina o PL nº 2.338/2023. Em 2 de setembro de 2026, o PL nº 1.542/2026 - que propõe um marco de responsabilidade por sistemas autônomos - foi apensado ao PL nº 904/2026.

Há ainda o PL nº 974/2026, cuja ementa se refere expressamente à governança de Agentes de Inteligência Artificial. A proposição está apensada ao PL nº 6.036/2025 e também foi recebida pela Comissão Especial do PL nº 2.338/2023.

Nenhuma dessas proposições constitui direito vigente. Há, adicionalmente, um problema terminológico digno de registro: o PL nº 974/2026 emprega 'Agente de IA' em sentido normativo distinto daquele adotado tecnicamente neste artigo, o que reforça a necessidade de harmonização conceitual entre o debate técnico e a legislação em formação.[4]

7 DIÁLOGO COM AUTONOMIA DECISÓRIA, VULNERABILIDADE E SOBERANIA

A comunicação emergente entre agentes acrescenta uma camada específica a uma agenda mais ampla sobre autonomia decisória e arquitetura informacional.

Em Autonomia Decisória na Era da IA, a análise parte da hipótese de que sistemas digitais não influenciam apenas escolhas finais. Eles também podem estruturar previamente o ambiente informacional no qual as escolhas serão formadas (Genova, 2026a).

Sistemas multiagentes aprofundam essa questão porque a informação apresentada ao usuário pode ser resultado de sucessivas operações internas - pesquisa, filtragem, classificação, resumo, validação, negociação entre agentes e seleção final - antes de chegar à interface humana.

O usuário percebe uma resposta. Por trás dela, pode existir uma cadeia de processos distribuídos. Se a comunicação interna sofre deriva semântica, o usuário interage com o

[4] O PL nº 974/2026 emprega “Agente de IA” em sentido normativo distinto do utilizado tecnicamente neste artigo, abrangendo categorias ligadas a fornecedor e operador. A divergência terminológica recomenda cautela na comparação entre taxonomias técnicas e legislativas.

produto final de uma arquitetura cuja totalidade já não é diretamente observável e pode não permanecer integralmente reconstruível nem mesmo pelo operador.

Sob a perspectiva da Influência Cognitiva Estrutural, isso desloca a investigação do conteúdo isolado para a arquitetura que determinou quais informações foram preservadas, descartadas, transformadas ou priorizadas antes de chegarem à pessoa.

A conexão com A Monetização da Vulnerabilidade também é relevante. Sistemas agênticos comerciais podem pesquisar, negociar, recomendar, precificar e personalizar abordagens. Em ambientes envolvendo crianças, idosos, consumidores superendividados, pacientes ou indivíduos em dependência funcional de sistemas digitais, a perda de legibilidade interna pode ampliar assimetrias já existentes entre fornecedor e usuário (Genova, 2026b).

Há, por fim, uma dimensão de soberania institucional. Estados, hospitais, universidades, tribunais, bancos e administrações públicas podem transferir atividades progressivamente relevantes a sistemas privados. Quando uma função jurídica permanece formalmente sob responsabilidade humana ou estatal, mas sua execução material depende de arquitetura que o próprio responsável não consegue auditar adequadamente, cria-se uma diferença entre competência formal e capacidade efetiva de controle.

A questão não exige antropomorfizar máquinas. Exige perguntar se o aumento da autonomia funcional pode produzir, paralelamente, redução da capacidade humana de governar aquilo que foi delegado.

8 DEVER DE LEGIBILIDADE SEMÂNTICA PERSISTENTE

Como resposta normativa à assimetria descrita, propõe-se o Dever de Legibilidade Semântica Persistente - DLSP.[5]

Não se reivindica como inédita a preocupação geral com legibilidade, auditabilidade ou explicabilidade. Esses princípios já aparecem no AI Act, em pesquisas sobre comunicação verificável, em documentos do NIST e no Singapore Consensus. A contribuição específica proposta é a sistematização jurídico-analítica da preservação da legibilidade semântica em sistemas multiagentes como dever proporcional ao risco e relevante à prova e à responsabilização. A manutenção temporal da legibilidade possui antecedentes técnicos e regulatórios; não se reivindica sua descoberta.

[5] Dever de Legibilidade Semântica Persistente (DLSP) é proposição normativa autoral construída em diálogo com literatura e instrumentos anteriores sobre legibilidade, auditabilidade, logging, supervisão humana e deriva semântica. Sua especificidade proposta está na articulação jurídico-analítica da manutenção temporal da legibilidade com risco, prova e responsabilização diante da evolução comunicacional do sistema.

O DLSP pode ser definido como:

> *O dever, proporcional ao risco, de projetar, contratar e operar sistemas multiagentes de modo que comunicações relevantes à produção de efeitos jurídicos, econômicos, administrativos ou sobre direitos permaneçam passíveis de reconstrução semântica independente durante todo o período materialmente relevante de funcionamento do sistema.*

Não se exige que toda comunicação interna utilize linguagem natural. Também não se pretende impedir adaptação ou comunicação emergente. A exigência é funcional: eficiência interna não deve destruir as condições externas de responsabilização.

Em aplicações de maior risco, o DLSP pode justificar requisitos como:

- registro temporal íntegro das comunicações materialmente relevantes;
- identificação inequívoca dos agentes participantes;
- preservação das versões de modelos, ferramentas e instruções utilizadas;
- armazenamento do contexto necessário à interpretação posterior;
- mecanismos de detecção de deriva semântica;
- tradução ou documentação das convenções internas relevantes à cadeia decisória;
- reexecução controlada de cadeias decisórias, quando tecnicamente possível;
- testes periódicos de compreensão por auditores independentes;
- limites à autonomia quando a capacidade de supervisão cair abaixo de limiar previamente definido; e
- mecanismos humanos efetivos de interrupção.

A pergunta regulatória deixa, assim, de ser apenas se o sistema mantém logs e passa a incluir se um terceiro qualificado consegue reconstruir, de maneira confiável, como esses registros se relacionam com a decisão produzida.

9 AGENDA EMPÍRICA PARA A OCE

A OCE deve permanecer refutável. Sem operacionalização e teste, seria apenas uma categoria retórica.

Considere-se I_A(t) como a proporção de comunicações corretamente interpretadas pelos agentes para execução da tarefa no instante t, e R_H(t) como a acurácia média de reconstrução semântica obtida por avaliadores humanos independentes.

Define-se exploratoriamente:

$$OCE_t = I_A(t) - R_H(t)$$

com ambas as grandezas normalizadas no intervalo [0,1]. A hipótese nula poderia assumir:

$$H_0: \; I_A(t) - R_H(t) \leq \delta$$

enquanto a hipótese alternativa seria:

$$H_1: \; I_A(t) - R_H(t) > \delta$$

em que δ corresponde a limiar de assimetria previamente definido e justificado no protocolo.

A comparação pressupõe unidades comunicacionais, tarefas, acesso ao contexto e critérios de correção comparáveis, definidos prospectivamente. A normalização em [0,1] não garante essa comparabilidade. OCE_t varia entre −1 e 1 e constitui uma diferença descritiva exploratória; um valor positivo, isoladamente, não demonstra emergência, deterioração temporal ou contribuição causal da comunicação para o resultado. Sua interpretação exige avaliação longitudinal e controles para mudanças da tarefa, do contexto e da experiência dos avaliadores. O sucesso final na tarefa não substitui, por si só, a aferição da interpretação comunicacional.

O fenômeno mais forte, entretanto, não seria simplesmente encontrar diferença positiva entre máquinas e humanos. A hipótese estruturalmente mais relevante ocorreria durante intervalo temporal no qual a capacidade de interpretação entre agentes permanecesse estável ou aumentasse, a capacidade humana diminuísse e o desempenho objetivo do sistema permanecesse estável ou aumentasse:

$$dI_A/dt \geq 0; \quad dR_H/dt < 0; \quad dP/dt \geq 0$$

P(t) designa o desempenho objetivo na tarefa. As derivadas representam um esquema conceitual de tendência; em dados discretos, sua avaliação dependerá de janelas temporais, estimativas de incerteza e critérios previamente registrados. A estabilidade ou elevação de $I_A(t)$ e P(t) deverá ocorrer acima de patamares mínimos de funcionalidade, também previamente definidos.

Nessa configuração, os agentes manteriam ou aumentariam sua capacidade de comunicação, a compreensão humana diminuiria e o sistema continuaria desempenhando adequadamente a tarefa. O sistema não estaria falhando em comunicar. Estaria falhando em permanecer legível para nós.

A mensuração de R_H exigiria especial cautela metodológica. Avaliadores deveriam ser cegados quanto às condições experimentais, receber treinamento padronizado e acesso equivalente ao contexto. Múltiplos avaliadores deveriam examinar cada unidade relevante, com mensuração formal da confiabilidade interavaliadores. Conforme a natureza da variável e o número de avaliadores, poderiam ser utilizados coeficientes como Cohen κ, Fleiss κ ou Krippendorff α.

O protocolo precisaria ainda controlar expertise dos avaliadores, complexidade da tarefa, heterogeneidade das famílias de modelos, duração da interação, memória disponível, quantidade e natureza das ferramentas, pressão por eficiência, número de agentes, limites de contexto, possibilidade de renegociação de convenções e grau de acesso dos avaliadores ao histórico.

Pré-registro das hipóteses, separação entre análise exploratória e confirmatória e replicação independente seriam indispensáveis antes de se atribuir status empírico robusto ao índice.

10 LIMITAÇÕES

As evidências recentes precisam ser tratadas proporcionalmente ao seu grau de maturidade. Emergence World, GlossoGen, Verifiable Semantics for Agent-to-Agent Communication e o trabalho sobre esteganografia instrumental são preprints recentes. Eles oferecem evidência relevante, mas ainda não substituem replicação independente e revisão cumulativa do campo (Akkil et al., 2026; Stengel-Eskin et al., 2026; Schoenegger et al., 2026; Rippin et al., 2026).

Em contraste, a existência geral de comunicação emergente opaca encontra suporte em literatura revisada por pares e conferências consolidadas, incluindo TMLR, ACL, NeurIPS e Nature Communications (Boldt; Mortensen, 2024; Carmeli; Belinkov; Meir, 2024; Li et al., 2024; Wieczorek et al., 2024).

Nenhuma das evidências analisadas demonstra consciência, vontade própria, intenção geral de ocultação ou existência de uma 'língua universal das IAs'. Também não demonstram que todo sistema multiagente sofrerá deriva semântica relevante.

A hipótese juridicamente relevante é mais modesta: certas arquiteturas e condições de interação podem produzir comunicação cuja utilidade para os agentes permaneça elevada enquanto sua reconstrução humana se torna mais difícil. Esse risco é suficiente para justificar pesquisa. Ainda não é suficiente para justificar generalizações universais.

11 CONCLUSÃO

A expressão 'língua das IAs' chama atenção, mas pode obscurecer o problema realmente importante. A questão não está em saber se agentes artificiais conseguem inventar novas palavras. Sistemas multiagentes já demonstraram capacidade para desenvolver convenções e protocolos.

O problema começa quando a comunicação permanece funcional para aqueles agentes enquanto sua interpretação se torna progressivamente dependente de uma história que o supervisor humano já não consegue reconstruir com segurança. Nesse momento, a discussão deixa de ser apenas linguística. Torna-se uma questão de governança, responsabilidade, prova e poder.

A existência de logs não garante auditabilidade. A presença de uma pessoa no circuito não garante supervisão substantiva. A autonomia técnica não elimina responsabilidade

jurídica. E a eficiência operacional de um sistema não deveria ser avaliada independentemente da capacidade de reconstruir aquilo que foi delegado a ele.

A Opacidade Comunicacional Emergente é proposta neste trabalho como categoria jurídico-analítica para representar a divergência entre a inteligibilidade funcional da comunicação entre agentes e sua reconstrução humana ao longo do tempo. O Dever de Legibilidade Semântica Persistente decorre dessa assimetria: quanto maior a relevância jurídica das ações delegadas, menor deve ser a tolerância a uma arquitetura que permaneça operacional enquanto perde progressivamente sua capacidade de ser compreendida e auditada por aqueles que respondem por seus efeitos.

O desafio da próxima geração de supervisão de IA talvez não seja apenas observar mais. Será preservar a possibilidade de compreender.

Porque, mesmo quando tudo estiver registrado, ainda precisamos ser capazes de compreender aquilo que estamos vendo.

REFERÊNCIAS

AKKIL, Deepak et al. *Emergence World: Adversarial Stress-Testing of Long-Horizon Multi-Agent Systems.* arXiv:2609.17320, 2026. Preprint. DOI: 10.48550/arXiv.2609.17320. Disponível em: https://arxiv.org/abs/2609.17320. Acesso em: 17 set. 2026.

ANTONELLI, Valdir. Agentes autônomos de IA criaram códigos próprios para se comunicar. *Olhar Digital*, 16 set. 2026. Disponível em: https://olhardigital.com.br/2026/09/16/inteligencia-artificial/agentes-autonomos-de-ia-criaram-codigos-proprios-para-se-comunicar/. Acesso em: 17 set. 2026.

BOLDT, Brendon; MORTENSEN, David R. A Review of the Applications of Deep Learning-Based Emergent Communication. *Transactions on Machine Learning Research*, 2024. Disponível em: https://mlanthology.org/tmlr/2024/boldt2024tmlr-review/. Acesso em: 17 set. 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: 17 set. 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: 17 set. 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/l13709compilado.htm. Acesso em: 17 set. 2026.

BRASIL. **Lei nº 14.133, de 1º de abril de 2021**. Lei de Licitações e Contratos Administrativos. Brasília, DF: Presidência da República, 2021. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14133.htm. Acesso em: 17 set. 2026.

BRASIL. Câmara dos Deputados. **Projeto de Lei nº 2.338/2023**. 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: 17 set. 2026.

BRASIL. Câmara dos Deputados. **Projeto de Lei nº 904/2026**. Institui normas de responsabilidade, transparência e auditabilidade para sistemas de inteligência artificial de alto impacto. Brasília, DF, 2026. Disponível em: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2606324. Acesso em: 17 set. 2026.

BRASIL. Câmara dos Deputados. **Projeto de Lei nº 974/2026**. Estabelece critérios de governança de Agentes de Inteligência Artificial e altera a Lei nº 12.965/2014 e a Lei nº 13.709/2018. Brasília, DF, 2026. Disponível em: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2607218. Acesso em: 17 set. 2026.

BRASIL. Câmara dos Deputados. **Projeto de Lei nº 1.542/2026**. Estabelece o Marco de Responsabilidade por Sistemas Autônomos no Brasil. Brasília, DF, 2026. Disponível em: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2612896. Acesso em: 17 set. 2026.

CARMELI, Boaz; BELINKOV, Yonatan; MEIR, Ron. Concept-Best-Matching: Evaluating Compositionality in Emergent Communication. *Findings of the Association for Computational Linguistics: ACL 2024*, p. 3186-3194, 2024. DOI: 10.18653/v1/2024.findings-acl.189. Disponível em: https://aclanthology.org/2024.findings-acl.189/. Acesso em: 17 set. 2026.

CASPER, Stephen et al. *The 2026 Singapore Consensus on Global AI Safety Research Priorities*. arXiv:2608.14611, 2026. Preprint. Disponível em: https://arxiv.org/abs/2608.14611. Acesso em: 17 set. 2026.

COMISSÃO EUROPEIA. **Guidelines for providers and deployers of AI high-risk systems**. Brussels, 2026. Disponível em: https://digital-strategy.ec.europa.eu/en/policies/guidelines-ai-high-risk-systems. Acesso em: 17 set. 2026.

GENOVA, Jandislei Antonio. **Autonomia Decisória na Era da IA**. São Paulo: Editora Dialética, 2026a.

GENOVA, Jandislei Antonio. **A Monetização da Vulnerabilidade: plataformas digitais, vício, ódio e responsabilidade civil na era algorítmica**. São Paulo: Editora Dialética, no prelo, 2026b.

JIANG, Yu et al. RiskLab: A Controlled Toolkit for Probing Emergent Risks in LLM-Based Multi-Agent Systems. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, p. 167-177, 2026. DOI: 10.18653/v1/2026.acl-demo.17. Disponível em: https://aclanthology.org/2026.acl-demo.17/. Acesso em: 17 set. 2026.

LI, Huao et al. Language Grounded Multi-agent Reinforcement Learning with Human-interpretable Communication. *Advances in Neural Information Processing Systems*, v. 37, p. 87908-87933, 2024. DOI: 10.52202/079017-2790. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2024/hash/a06e129e01e0d2ef853e9ff67b911360-Abstract-Conference.html. Acesso em: 17 set. 2026.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST). **Announcing the "AI Agent Standards Initiative" for Interoperable and Secure Innovation**. Gaithersburg, 17 fev. 2026. Disponível em: https://www.nist.gov/news-events/news/2026/02/announcing-ai-agent-standards-initiative-interoperable-and-secure. Acesso em: 17 set. 2026.

OECD. **Artificial Intelligence and Competitive Dynamics in Downstream Markets**. OECD Roundtables on Competition Policy Papers, n. 331. Paris: OECD Publishing, 2025. DOI: 10.1787/ccf0624a-en. Disponível em: https://doi.org/10.1787/ccf0624a-en. Acesso em: 17 set. 2026.

RIGGS, Jared; HAMIN, Maia; PERRY, Neil; EDELMAN, Benjamin; CIHON, Peter. **Summary Analysis of Responses to the Request for Information Regarding Security Considerations for AI Agents**. NIST Trustworthy and Responsible AI 800-5. Gaithersburg: National Institute of Standards and Technology, 2026. Disponível em: https://www.nist.gov/publications/summary-analysis-responses-request-information-regarding-security-considerations-ai. Acesso em: 17 set. 2026.

RIPPIN, Jimmy Laurence; MARSHALL, Simon C.; AFRICA, David Demitri; SCHROEDER DE WITT, Christian. *Tool Use Enables Undetectable Steganography in Multi-Agent LLM Systems*. arXiv:2606.28425, 2026. Preprint. DOI: 10.48550/arXiv.2606.28425. Disponível em: https://arxiv.org/abs/2606.28425. Acesso em: 17 set. 2026.

SCHOENEGGER, Philipp; CARLSON, Matt; SCHNEIDER, Chris; DALY, Chris. *Verifiable Semantics for Agent-to-Agent Communication*. arXiv:2602.16424, 2026. Preprint. DOI: 10.48550/arXiv.2602.16424. Disponível em: https://arxiv.org/abs/2602.16424. Acesso em: 17 set. 2026.

STENGEL-ESKIN, Elias et al. *GlossoGen: Emergent Language in Complex Multi-Agent LLM Interactions*. arXiv:2609.01491, 2026. Preprint. DOI: 10.48550/arXiv.2609.01491. Disponível em: https://arxiv.org/abs/2609.01491. Acesso em: 17 set. 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 (AI Act). Jornal Oficial da União Europeia, 2024. Disponível em: https://eur-lex.europa.eu/eli/reg/2024/1689/oj. Acesso em: 17 set. 2026.

WIECZOREK, Tobias J.; TCHUMATCHENKO, Tatjana; WERT-CARVAJAL, Carlos; EGGL, Maximilian F. A Framework for the Emergence and Analysis of Language in Social Learning Agents. *Nature Communications*, v. 15, art. 7590, 2024. DOI: 10.1038/s41467-024-51887-5. Disponível em: https://www.nature.com/articles/s41467-024-51887-5. Acesso em: 17 set. 2026.