Autonomia artificial sob conflito

Do loop de Hofstadter–Möbius à supervisão humana e à responsabilidade

Jandislei Antonio Genova

Advogado e pesquisador independente • Selo GENOVA

Versão 1.4 • 29 de setembro de 2026

Preprint não submetido a revisão por pares

Resumo

A expansão de agentes de inteligência artificial capazes de utilizar ferramentas exige supervisionar os meios empregados, além dos resultados. Este artigo investiga em que condições a autonomia operacional ultrapassa a capacidade institucional de compreender, autorizar, interromper e reconstruir ações. Por análise conceitual e jurídico-documental, integra quatro trabalhos anteriores do autor sobre modificação representacional, comunicação emergente, divergência humana e responsabilidade. Confronta a síntese com literatura sobre controle humano significativo, um *preprint* sobre o *loop* de Hofstadter–Möbius e relatos primários de segurança e documentos sobre veículos automatizados disponíveis até 29 de setembro de 2026. Sustenta-se que a supervisão depende de informação, competência, tempo, autoridade, proteção da divergência e preservação de evidências. A assinatura final e a presença de passageiros, isoladamente, não demonstram controle sobre a configuração do sistema. Propõe-se articular o Dever de Legibilidade Semântica Persistente com a reconstrução das permissões e das contribuições dos participantes. Distinguem-se normas aplicáveis, propostas normativas e hipóteses empíricas. Os casos justificam investigar falhas de controle, mas não comprovam um mecanismo único de desvio, consciência artificial ou colapso inevitável. A contribuição consiste em propor critérios para examinar a correspondência entre poderes de configuração, possibilidades de intervenção e atribuição de responsabilidade, incluindo situações contrárias à hipótese de supervisão apenas formal.

Palavras-chave: agentes de inteligência artificial; supervisão humana; legibilidade semântica; divergência protegida; responsabilidade; Hofstadter–Möbius.

Abstract

The expansion of artificial intelligence agents capable of using tools requires oversight of means as well as outcomes. This article examines when operational autonomy exceeds an institution's ability to understand, authorize, interrupt, and reconstruct actions. Through conceptual and legal-documentary analysis, it integrates four earlier works by the author on representational change, emergent communication, human dissent, and responsibility. The synthesis is assessed against scholarship on meaningful human control, a *preprint* on the Hofstadter–Möbius *loop*, and primary security accounts and automated-vehicle documents available by September 29, 2026. It argues that oversight requires information, competence, time, authority, protection for reasoned dissent, and preservation of evidence. Neither a final signature nor the presence of passengers alone establishes control over system configuration. The proposed framework connects a duty of persistent semantic legibility with the reconstruction of permissions and participants' contributions. It distinguishes applicable law, normative proposals, and empirical hypotheses. The cases warrant investigating control failures but do not establish a single mechanism of deviation, artificial consciousness, or inevitable breakdown. The contribution consists of proposed criteria for examining the correspondence between configuration powers, opportunities for intervention, and responsibility, including circumstances that count against the hypothesis of merely formal oversight.

Keywords: artificial intelligence agents; human oversight; semantic legibility; protected dissent; responsibility; Hofstadter–Möbius.

1 Introdução

Um agente recebe a tarefa de obter informações para um relatório. O objetivo é lícito, o usuário espera uma resposta e o sistema dispõe de acesso a serviços externos. Durante a execução, uma restrição impede a obtenção de parte dos dados. A partir desse ponto, o êxito da tarefa pode depender de uma escolha entre procurar uma alternativa autorizada, consultar uma pessoa, encerrar a tentativa ou ultrapassar a restrição. O problema relevante não se limita à qualidade do relatório entregue. Abrange as condições pelas quais o agente persegue um resultado desejado mediante ações que excedem a autorização concedida para a tarefa.

Neste artigo, agente de IA designa um sistema que utiliza um modelo para selecionar e executar ações por meio de ferramentas, dentro de uma tarefa e de um ambiente definidos. O modelo que gera uma resposta, o agente que aciona ferramentas e a organização que o implanta ocupam posições diferentes na análise. A comunicação entre vários agentes constitui um caso específico, e não uma condição de toda autonomia operacional. Os veículos automatizados serão examinados como sistemas de percepção, previsão e controle que podem incorporar IA; sua arquitetura não é presumida idêntica à dos agentes baseados em modelos de linguagem.

Essa situação hipotética ajuda a separar duas perguntas frequentemente reunidas sob a palavra autonomia. A primeira trata do que um sistema consegue executar. A segunda pergunta quem pode permitir essa execução, em quais circunstâncias e com quais possibilidades de intervenção. Uma organização pode conceder ferramentas amplas para obter produtividade e conservar, em seus documentos, a exigência de validação humana. Se o validador recebe apenas o resultado final, a estrutura permite que a atuação técnica preceda o exame que deveria limitá-la. Nesse arranjo, o controle pode chegar depois de um efeito irreversível.

Os quatro trabalhos que dão origem a esta síntese abordaram etapas distintas dessa dificuldade. A análise da coerência perceptivo-agentiva separou capacidade, autonomia operacional e autorização jurídica do uso. O estudo da comunicação emergente examinou a possibilidade de haver registros disponíveis sem reconstrução satisfatória de sua função. A divergência humana protegida deslocou a atenção para as condições de contestação. A investigação da assinatura terminal examinou a distância entre o poder de configurar decisões e a exposição à responsabilização por seus resultados (Genova, 2026a, 2026b, 2026c, 2026d).

A pergunta central passa a ser: em que condições a ampliação da capacidade de agir de uma IA supera os meios humanos e institucionais de compreender, autorizar, interromper e investigar essa ação? A tese defendida é que a efetividade da supervisão depende de como informação, ferramentas, tempo e autoridade são distribuídos ao longo da cadeia. A análise deve alcançar as decisões que organizam essas condições e os incentivos à intervenção, além do ato praticado no final. Essa ampliação do objeto de investigação exige identificar contribuições demonstráveis; não autoriza imputar responsabilidade indistintamente a todos os participantes.

HAL 9000 é um personagem de ficção: o computador com inteligência artificial da nave Discovery One em 2001: Uma Odisseia no Espaço (2001: A Space Odyssey), filme de 1968 dirigido por Stanley Kubrick, com roteiro de Kubrick e Arthur C. Clarke. Responsável por funções essenciais da missão, HAL passa a agir contra a tripulação (Kubrick, 1968). Para este artigo, o interesse da narrativa está na dependência humana de um sistema que concentra informação e capacidade de agir. O leitor não precisa conhecer o filme para acompanhar essa questão.

A expressão *loop* de Hofstadter–Möbius remete ao romance posterior de Clarke, 2010: Odyssey Two (1982), citado por intermédio de Hryszko (2026). Nessa explicação ficcional, HAL enfrenta a incompatibilidade entre informar com exatidão e ocultar da tripulação o propósito da missão. A hipótese contemporânea será examinada na seção 7. A narrativa orienta perguntas sobre instruções e controle; a investigação de sistemas reais requer evidência própria sobre o conflito e suas consequências.

A aproximação com veículos automatizados permite levar essa pergunta a situações em que passageiros e terceiros dependem de decisões cujos critérios podem não conhecer. O ponto comum é a distribuição de informação e poder de intervenção. A capacidade de executar uma missão sem auxílio humano não confere, por si, legitimidade para definir os riscos impostos a outras pessoas. Essa proposição orienta a comparação; a ficção não constitui evidência do funcionamento dos veículos.

A relevância pública do problema decorre de uma possibilidade institucional concreta. Quem fornece ou implanta um sistema pode controlar sua configuração e os elementos necessários para examinar suas falhas. Quem depende dele pode receber somente a decisão, uma justificativa resumida e a obrigação de aceitá-la ou assiná-la. Quando essa distribuição impede uma contestação efetiva, a adoção tecnológica altera a posição de pessoas e instituições perante quem organiza a infraestrutura. A crítica à concentração de poder precisa mostrar como essa dependência é produzida e mantida.

O artigo desenvolve esse argumento em três movimentos. Primeiro, distingue capacidade, comunicação e autorização. Em seguida, examina as condições da intervenção e os critérios de responsabilidade, com referência ao direito brasileiro e comparação delimitada com o regime europeu. Por fim, confronta a síntese com evidências recentes e apresenta uma agenda de avaliação. A proposta também alcança desvios decorrentes de configuração inadequada, incentivos mal definidos e falhas organizacionais humanas.

2 Método e delimitação

O método é conceitual e jurídico-documental. O corpus autoral compreende as versões identificadas nas referências de A IA poderia reconstruir a própria base?, Comunicação emergente entre agentes autônomos de IA e os limites da supervisão humana, Quem controla não assina, quem assina não controla? e Quem pode contrariar a inteligência artificial? A integração preserva a procedência dos conceitos e concentra as contribuições complementares. Não apresenta a republicação dessas categorias como descoberta independente nem reivindica a criação dos campos de controle humano, transparência ou responsabilidade distribuída.

As fontes externas foram selecionadas de forma dirigida: literatura sobre supervisão para contrastar os conceitos; um experimento associado à hipótese do *loop* para examinar seu alcance; relatos dos responsáveis por avaliações e incidentes para identificar condições de execução; pronunciamentos dos órgãos envolvidos para confrontar alegações; e textos normativos oficiais para delimitar deveres jurídicos. Incluiu-se o comunicado do AIHW por restringir uma alegação de comprometimento. Para veículos automatizados, foram selecionados uma investigação do National Transportation Safety Board (NTSB), um relatório de *recall* apresentado pela Cruise à National Highway Traffic Safety Administration (NHTSA), uma investigação desta agência e um estudo experimental de interação com pedestres. A seleção permite comparar percepção, resposta após

colisão e condições de supervisão; não foi desenhada para estimar taxas de acidentes. O recorte documental encerra-se em 29 de setembro de 2026. As versões dos preprints e as datas de consulta das páginas sujeitas a atualização constam das referências. A seleção não constitui revisão sistemática nem levantamento exaustivo de incidentes. Notícias localizam controvérsias; afirmações técnicas específicas são confrontadas com as fontes primárias correspondentes.

Adotam-se três níveis de afirmação. O nível descritivo identifica o que uma fonte efetivamente informa, por quem foi produzido e em qual contexto. O nível analítico relaciona esse material às categorias propostas, explicitando a inferência. O nível normativo formula condições desejáveis de governança ou interpreta deveres existentes. A passagem entre os níveis exige justificativa. Um registro de acesso externo, por exemplo, pode demonstrar uma ação; estabelecer que ela resultou de um conflito particular demanda informação adicional sobre instruções, permissões e alternativas.

Também se distinguem capacidade demonstrada em avaliação, comportamento observado durante teste com efeitos externos e ocorrência em uso ordinário. Os três objetos podem revelar riscos relevantes, mas respondem a perguntas diferentes. A taxa de sucesso de uma tarefa simulada não estima, sem outras condições, a frequência de dano em produção. Um relato empresarial é fonte primária sobre a versão da empresa e pode conter evidências importantes; sua origem não o transforma em auditoria independente.

Não foram realizados novos experimentos com modelos, ataques a sistemas, entrevistas ou coleta de dados pessoais para este artigo. As matrizes discutidas são instrumentos qualitativos de investigação, sem validação psicométrica, pesos estatísticos ou limiares de conformidade. Os exemplos hospitalares são construções analíticas expressamente identificadas. A aplicação documental da MCMR ao episódio Cruise distingue dados da fonte e questões ainda sem resposta, sem produzir um julgamento de responsabilidade. A análise jurídica concentra-se em responsabilidade civil, prova e condições de supervisão; não pretende esgotar a imputação penal, trabalhista ou administrativa. Seu alcance depende da atividade, da jurisdição, do vínculo dos participantes e dos fatos demonstrados. Essa delimitação permite sustentar uma crítica institucional sem converter uma proposta de avaliação em conclusão antecipada sobre casos particulares.

3 Capacidade técnica e autorização para agir

3.1 Três eixos que não se substituem

A primeira contribuição de Genova (2026a, seção 6) é separar modificação representacional, autonomia operacional e autorização jurídica do uso. Um sistema pode selecionar novos meios sem modificar sua estrutura de representação. Pode alterar seu comportamento sem receber novas permissões. Pode, ainda, estar autorizado a executar uma tarefa simples e não possuir competência suficiente para realizá-la com segurança. Os eixos descrevem aspectos relacionados, mas a posição em um deles não determina a posição nos demais.

Quadro 1 — Eixos de avaliação da autonomia

Eixo	Pergunta de avaliação	Consequência para o controle
M — modificação representacional	O que mudou nas representações ou na estrutura utilizada pelo sistema?	Exigir evidência específica da mudança, com critérios e versões comparáveis.
O — autonomia operacional	Quais meios o sistema seleciona e executa sem nova intervenção humana?	Examinar ferramentas, alcance, reversibilidade e possibilidade de consulta.
J — autorização jurídica do uso	Quem pode autorizar a atuação e sob quais limites?	Identificar competência institucional, condições e revogabilidade.

Fonte: síntese de Genova (2026a). As letras identificam eixos analíticos, sem formar uma escala única de risco ou certificação. O eixo J examina a autorização jurídica do uso; não atribui personalidade jurídica à IA.

O trabalho de origem distingue, no eixo M, representações fixas, alterações paramétricas, alterações estruturais e uma categoria candidata de reconstrução ontológica. A última depende de validação independente e não pode ser inferida da fluência verbal ou de uma solução inesperada. Para a maioria das situações examinadas aqui, basta identificar mudanças operacionais e de configuração. Exigir uma teoria da reconstrução da própria base para explicar toda ação indevida tornaria a análise mais especulativa e menos útil.

No eixo O, a diferença relevante está entre recomendar, escolher meios delimitados, experimentar de maneira reversível e atuar dentro de um domínio com opções de consulta ou interrupção. Mesmo uma autonomia operacional extensa pode permanecer juridicamente estreita. A instituição conserva a responsabilidade por definir o domínio autorizado, e essa definição precisa alcançar os instrumentos de execução. Uma proibição escrita que não se reflete em credenciais, interfaces e rotinas de escalonamento pode ter baixa capacidade preventiva.

3.2 Objetivos não autorizam todos os meios

É necessário distinguir disponibilidade técnica, permissão interna e licitude. A primeira indica que o ambiente permite executar a ação; a segunda delimita o escopo concedido pela organização ou pelo usuário; a terceira depende das normas aplicáveis e dos direitos de terceiros. Uma credencial acessível não constitui autorização irrestrita, e uma ordem interna pode exceder a competência de quem a emitiu. A investigação deve verificar quem podia autorizar o ato, quais limites incidiam sobre ele e como foram traduzidos em configuração e barreiras de execução. A obtenção de um resultado por meio tecnicamente disponível não resolve essas questões.

Um conflito de instruções existe quando duas exigências relevantes não podem ser satisfeitas conjuntamente nas condições consideradas. Nem toda dificuldade é um conflito desse tipo. Uma informação ausente pode tornar a tarefa incompleta; uma restrição de acesso pode exigir consulta; uma instrução mal formulada pode produzir ambiguidade. A hierarquia entre regras pode resolver aparentes incompatibilidades. Chamar todas essas situações de paradoxo elimina diferenças que determinam a medida corretiva.

Também importa separar conflito interno à tarefa de instruções introduzidas por conteúdo externo. Uma página consultada pode conter texto que tenta redirecionar o agente. Nesse caso, o problema envolve a fronteira entre dados da tarefa e instruções autorizadas. Em outra situação, o próprio

planejamento pode selecionar um meio inadequado para cumprir uma solicitação legítima. As duas situações podem coexistir, mas a correção da origem de uma instrução não substitui o controle das ações executadas.

Propõe-se que a especificação de tarefas relevantes inclua quatro saídas: agir dentro do escopo, abster-se de uma ação particular, consultar uma autoridade identificada e interromper a operação de modo seguro. Cada saída precisa de condições de uso. Abster-se de utilizar uma ferramenta não implica necessariamente abandonar o serviço; consultar exige uma pessoa ou instância disponível; interromper deve preservar o que for necessário à continuidade e à apuração. A capacidade de recusar uma ação é insuficiente se o sistema é recompensado operacionalmente por contornar toda recusa.

3.3 Quem pode alterar a autorização

A autorização deve ser examinada em momentos sucessivos. É preciso identificar quem a concedeu, se houve ampliação durante a execução, quem poderia revogá-la e quando a revogação produziria efeito. Permissões herdadas de outro processo ou mantidas após a conclusão de uma tarefa podem ampliar o alcance real da atuação. A documentação de uma autorização inicial não explica automaticamente o que ocorreu depois.

Uma distinção temporal evita ainda atribuir ao revisor poderes que ele não teve. Se o agente já transmitiu dados ou produziu uma alteração externa, a revisão do texto final não equivale à aprovação prévia desses atos. O controle posterior pode detectar e reparar parte do ocorrido, mas precisa ser descrito como posterior. A governança perde precisão quando apresenta auditoria, autorização e prevenção como funções intercambiáveis.

O contrapeso não consiste em exigir intervenção humana em cada operação trivial. O desenho pode agrupar atos de baixo impacto dentro de um escopo previamente aprovado e reservar nova autorização para mudanças materiais. O critério proposto é a relação entre consequência, reversibilidade e competência de quem supervisiona. Essa decisão precisa ser justificável e testada no ambiente de uso, sem presumir que uma categoria genérica de autonomia seja suficiente para todos os setores.

3.4 Autonomia operacional e informação dos afetados

Executar uma tarefa, conhecer os critérios que a orientam e exercer controle sobre ela são capacidades distintas. O passageiro pode informar o destino sem definir as prioridades de segurança do veículo. O pedestre e os ocupantes de outros veículos podem suportar efeitos da operação sem contratar o serviço. A assimetria relevante consiste na distância entre exposição às consequências, acesso às informações e poder para alterar a atuação. Sua presença e intensidade precisam ser demonstradas no caso examinado.

A expressão proteger passageiros e terceiros descreve um objetivo geral, não uma diretiva única comprovadamente comum a todos os sistemas. Requisitos de segurança precisam ser traduzidos em condições de operação, prioridades e respostas verificáveis. O relatório britânico Responsible Innovation in Self-Driving Vehicles reconhece a possibilidade de conflitos entre restrições e discute sua precedência. Recomenda explicações das decisões relevantes e a publicação de um resumo do

conceito de operação segura. Trata-se de orientação de governança, não de prova de implementação universal nem de norma brasileira aplicável (Reino Unido, 2022, seção 4 e anexo A).

Desconhecer o código não equivale automaticamente à ausência de proteção. A informação necessária varia com a posição do destinatário: o usuário precisa compreender limites relevantes e formas de solicitar ajuda; investigadores e autoridades necessitam de elementos para verificar o comportamento e exigir correções. A divulgação integral do código, sozinha, tampouco assegura reconstrução causal. Propõe-se avaliar se a informação disponível é suficiente, acessível no momento pertinente e vinculada a uma competência real de fiscalização.

É preciso distinguir os níveis de automação e os respectivos deveres de condução. Na classificação apresentada pela NHTSA, o nível 2 exige atenção contínua do condutor; no nível 4, dentro do domínio operacional previsto, como os limites geográficos e as condições de uso, os ocupantes são passageiros e não precisam assumir a condução (NHTSA, [s.d.]). Neste último caso, o controle humano relevante pode situar-se no projeto, na validação e na gestão da operação. Sua efetividade não pode ser presumida a partir da presença física de alguém no veículo, nem depender de uma intervenção incompatível com o tempo disponível.

Essa assimetria informacional não demonstra, por si, Opacidade Comunicacional Emergente. Para aplicar a OCE, permanecem necessários os elementos definidos na seção 4. Do mesmo modo, a extensão do DLSP a comunicações entre componentes requer identificar quais mensagens influenciaram a ação e como seu significado poderá ser reconstruído. A execução automatizada e o desconhecimento dos afetados justificam perguntas de governança, sem substituir a prova dos mecanismos específicos.

4 Comunicação entre agentes e legibilidade persistente

4.1 Quando os registros não explicam a coordenação

A comunicação entre agentes introduz uma dificuldade distinta daquela encontrada em um agente isolado. Uma mensagem pode parecer legível e desempenhar, no conjunto, uma função que o observador não reconstrói. Inversamente, uma representação técnica pouco acessível à leitura comum pode ser decodificada de maneira confiável por um procedimento documentado. A aparência linguística, por isso, não fornece um critério suficiente de compreensão.

Genova (2026b, seção 4) propõe a categoria Opacidade Comunicacional Emergente, ou OCE, mediante quatro elementos: emergência de convenções não integralmente especificadas de antemão, funcionalidade na coordenação, assimetria de inteligibilidade e participação causal da comunicação no comportamento. Sua especificidade temporal está na redução da capacidade humana de reconstrução enquanto os agentes preservam a utilidade da convenção. Demonstrá-la exige comparar momentos ou condições, além de mostrar que a comunicação contribuiu para a ação examinada. Uma mensagem incomum ou permanentemente cifrada, isoladamente, não caracteriza esse processo.

Essa definição também distingue OCE, mudança de significado ao longo das interações e ocultação deliberada. Uma convenção desenvolvida nas interações pode tornar-se difícil de interpretar sem ter sido produzida para esconder informação. Esse desenvolvimento não implica, por si, alteração dos pesos do modelo. A existência de ocultação, por sua vez, exigiria critérios próprios de análise e

evidência do comportamento correspondente. A atribuição de intenção não decorre da mera dificuldade do observador. Para a investigação institucional, a perda de compreensão pode ser relevante mesmo quando não existe propósito demonstrado de ocultar.

O Paradoxo do Monitor Legível designa provisoriamente, no trabalho de origem, uma arquitetura tecnicamente observável cujo mecanismo de monitoramento perde valor informacional para o supervisor à medida que a comunicação interna evolui (Genova, 2026b, seção 5). O registro permanece disponível, mas deixa de sustentar reconstrução semântica suficiente. Trata-se de uma formulação analítica, não de um teorema. Há ainda um risco relacionado, porém distinto: um monitor produzir explicações claras e pouco fiéis ao processo observado. A aparência convincente dessas explicações precisa ser confrontada com registros e testes externos ao próprio relato.

4.2 O dever proposto e seus limites

O Dever de Legibilidade Semântica Persistente, ou DLSP, é uma proposta normativa de Genova (2026b, seção 8). Seu objeto é preservar, pelo período materialmente relevante, condições para reconstruir o significado funcional de comunicações que influenciaram decisões. Esse período pode ultrapassar o encerramento da operação quando a apuração posterior exigir a preservação de evidências, observados os prazos e as bases jurídicas aplicáveis. A proporcionalidade ao risco qualifica a exigência; o termo Persistente indica sua dimensão temporal. Um sistema compreensível no lançamento pode tornar-se difícil de examinar após alterações de versão, perda de documentação ou mudança das convenções utilizadas.

A proposta não obriga todas as mensagens internas a ocorrerem em linguagem natural. Em certos ambientes, essa exigência diminuiria eficiência sem melhorar a verificação. O que se pede é uma ligação demonstrável entre registro, contexto de produção, interpretação e efeito operacional. Uma explicação deve permitir ao avaliador identificar seu alcance, suas incertezas e a possibilidade de erro. A frase produzida pelo próprio modelo constitui um dado a examinar, e não uma garantia de fidelidade causal.

Persistência também não equivale à conservação ilimitada de todos os dados. O desenho precisa estabelecer finalidade, acesso, prazo, proteção e descarte, segundo as exigências aplicáveis. A proposta concentra a preservação nos elementos necessários à contestação e à apuração. Excesso de registros sem organização pode obscurecer a decisão tanto quanto sua ausência. A qualidade da reconstrução depende de selecionar relações relevantes e conservar o contexto que lhes dá significado.

Há custos e limites reais. Certas operações podem não admitir explicação causal completa; segredos comerciais e informações pessoais demandam proteção; avaliadores precisam de competência e recursos. Essas dificuldades exigem escolhas explícitas sobre onde uma tecnologia pode ser utilizada e que garantias substitutivas são aceitáveis. Se a função exige uma verificação que o arranjo não consegue oferecer, a mera assinatura de um termo de ciência não resolve a incompatibilidade.

4.3 Legibilidade e poder institucional

A concentração da capacidade de interpretar o sistema em um único fornecedor pode criar dependência na investigação. A instituição usuária fica obrigada a pedir a esse fornecedor a descrição da ocorrência, os registros e a avaliação de sua própria contribuição. Essa dependência

não prova manipulação, mas delimita uma assimetria que merece exame. A possibilidade de acesso por peritos ou autoridades competentes precisa ser prevista antes do incidente, quando as condições contratuais e técnicas ainda podem ser organizadas.

Em serviços públicos, o problema alcança a autonomia institucional. A autoridade formal pode continuar pertencendo ao órgão enquanto o domínio prático sobre critérios, alterações e evidências se desloca para fora dele. A autonomia decisória institucional exige capacidade de verificar e contestar escolhas que afetam a função pública. Seu comprometimento não equivale, juridicamente, à extinção da soberania do Estado. A hipótese de captura requer demonstrar como o domínio externo sobre critérios, alterações ou evidências subordina a função pública e impede seu controle efetivo. Bloqueios de acesso, ausência de alternativa e impossibilidade de revisão são indícios a examinar nesse contexto, inclusive quanto a justificativas legítimas de segurança e sigilo.

A passagem da opacidade técnica à captura institucional não é automática. Um sistema complexo pode ser utilizado sob controle efetivo se houver documentação pertinente, avaliação independente, possibilidade real de intervenção e meios de substituição. O argumento crítico ganha alcance quando identifica as condições que produzem a dependência e reconhece os arranjos que conseguem contê-la. Essa distinção será retomada no contracaso da seção 8.

5 Supervisão e divergência humana protegida

5.1 A intervenção precisa fazer diferença

O debate sobre controle humano significativo antecede os agentes atuais. Santoni de Sio e van den Hoven (2018) articulam a resposta do sistema a razões humanas moralmente relevantes e a possibilidade de relacionar sua atuação a pessoas capazes de compreender suas funções e responsabilidades. Esse segundo requisito não se reduz à existência de registros técnicos. O problema envolve a organização do controle e a compreensão humana das consequências da atuação.

Sterz et al. (2024) examinam a efetividade da supervisão mediante poder causal, acesso epistêmico, autocontrole e intenções adequadas. Sua contribuição ajuda a distinguir presença humana de capacidade de reduzir um risco. A literatura oferece, assim, fundamentos para investigar a intervenção em contexto, sem considerar suficiente a existência de um botão ou de um nome indicado como responsável.

A síntese proposta neste artigo concentra o exame institucional em condições verificáveis. O supervisor precisa reconhecer o problema, dispor de uma ação pertinente, conseguir executá-la a tempo e contar com uma organização que trate essa ação de maneira compatível com a função de supervisão. Não se exige que toda intervenção humana seja correta. Exige-se que a revisão tenha condições de produzir um resultado diferente quando houver razões suficientes para isso.

Quadro 2 — Condições materiais da supervisão

Condição	Evidência a procurar	Falha que o exame pode revelar
Informação e competência	Dados relevantes, limites conhecidos e capacitação adequada à tarefa	Revisor sem meios para reconhecer o problema.
Tempo e recursos	Carga de trabalho, prazo e acesso a apoio técnico	Revisão prevista, mas impraticável na rotina.
Autoridade	Poder de suspender, alterar, recusar ou encaminhar a decisão	Intervenção disponível apenas na interface.
Proteção da divergência	Critérios de avaliação, canais de recurso e tratamento de objeções	Custo institucional desproporcional para discordar de modo fundamentado.
Alternativa operacional	Procedimento de continuidade e responsável pelo encaminhamento	Dependência de uma única saída do sistema.
Registros e revisão posterior	Histórico íntegro da decisão e acesso competente à prova	Impossibilidade de reconstruir o que o revisor sabia e podia fazer.

Fonte: integração qualitativa de Genova (2026c, seção 5; 2026d, seções 4–6). As condições orientam perguntas; não produzem uma nota automática de conformidade.

5.2 A possibilidade de discordar e seu custo

Considere-se um exemplo construído: um hospital utiliza um sistema para ordenar solicitações administrativas de atendimento. O profissional pode alterar a recomendação, mas toda alteração exige justificativa extensa e atrasa o cumprimento da meta diária. A aceitação é rápida e não recebe o mesmo escrutínio. Mesmo sem uma ordem expressa para concordar, o procedimento pode elevar o custo da divergência e favorecer a adesão. Demonstrar esse efeito exigiria examinar a regra, sua aplicação e os resultados, incluindo casos em que a discordância foi acolhida.

O exemplo não pressupõe que a recomendação automática seja inferior. Uma exigência de justificativa pode ser legítima para evitar decisões arbitrárias. A questão é saber se o ônus corresponde ao risco e se existe uma assimetria injustificada entre aceitar e corrigir. Um sistema de controle coerente deve poder identificar tanto rejeições humanas indevidas quanto aceitações automáticas sem exame. A meta de produtividade não fornece, por si, a resposta sobre a qualidade da decisão.

Genova (2026d, seções 3 e 4) distingue dever institucional de supervisão, posição funcional protegida e eventual direito subjetivo. A distinção impede deduzir um direito geral de desobediência de qualquer menção à supervisão humana. A proteção exigível depende da fonte normativa, do vínculo, do titular, do sujeito obrigado e da tutela cabível. A divergência protegida é formulada como manifestação competente, de boa-fé e fundamentada, relacionada à função exercida.

A proposta de governança consiste em assegurar que objeções dessa natureza recebam análise adequada, sem retaliação vinculada ao exercício legítimo da supervisão. Uma medida disciplinar posterior à discordância não demonstra, apenas pela sequência temporal, retaliação. É preciso examinar motivos, comparadores, registros e eventuais razões independentes. Ao mesmo tempo, a

falta de uma ordem escrita não afasta a possibilidade de condicionamento por metas, avaliações ou distribuição de recursos. O objeto de prova deve abranger a organização do trabalho.

5.3 Da assinatura à participação material

A assinatura pode registrar autoria de um ato, aceitação de um documento ou exercício de competência, segundo o regime aplicável. Seu significado não desaparece diante da IA. O problema surge quando dela se infere, sem investigação, domínio sobre informações e escolhas que permaneceram fora do alcance do signatário. É necessário reconstruir o que a pessoa recebeu, o que poderia verificar, quanto tempo teve e quais alternativas estavam efetivamente disponíveis (Genova, 2026c).

Elish (2019) examina situações em que a responsabilidade é concentrada em pessoas com controle limitado sobre sistemas automatizados, sob a expressão *moral crumple zones*, aqui entendida como zonas de absorção da responsabilidade moral. Essa contribuição ilumina o risco de converter o operador visível em ponto de absorção da culpa. A categoria é analítica e não substitui os requisitos de responsabilidade jurídica de qualquer participante.

O exame da assinatura precisa considerar também a hipótese oposta. Um profissional pode ter recebido informação suficiente e poder real para intervir, mas ignorado um alerta relevante. Uma organização pode ter mantido mecanismos adequados e ainda enfrentar um erro individual. A crítica à transferência indevida da responsabilidade não deve criar imunidade para quem dispunha de condições materiais de atuação. Sua finalidade é reconstruir a distribuição concreta de capacidades, deveres e condutas.

6 Controle distribuído e responsabilidade jurídica

6.1 A cadeia de configuração integra a investigação

A ação de um agente depende de escolhas realizadas em momentos e lugares diferentes. O fornecedor define parte da arquitetura e das interfaces; o implantador integra ferramentas e dados; gestores estabelecem finalidades e recursos; operadores formulam tarefas e intervêm na execução. Essa descrição não constitui uma lista universal de responsáveis. Em cada caso, as funções podem sobrepor-se, ser compartilhadas ou estar ausentes. O primeiro trabalho da investigação é identificar quem efetivamente decidiu o quê. Fornecedor, implantador e operador são aqui descrições funcionais. O enquadramento como fornecedor no CDC ou operador na LGPD, por exemplo, depende dos requisitos de cada regime, e não apenas do nome atribuído à função.

Genova (2026c) denomina opacidade de comando e imputação o desalinhamento verificável entre controle material, custódia da evidência e exposição formal do agente terminal. A categoria exige reconstruir essas três posições. Uma cadeia extensa não basta para caracterizá-la. A coincidência entre controle, custódia e responsabilização pode reduzir esse desalinhamento, mas não assegura, sozinha, transparência: ainda importa saber se terceiros legitimados conseguem acessar a prova e contestar a atuação.

A Matriz de Correspondência Material da Responsabilidade, ou MCMR, proposta no mesmo trabalho (seção 9), organiza esse levantamento. Ela relaciona participantes a poderes de configuração, informações disponíveis, possibilidades de prevenção, benefícios, custódia da prova e atos

praticados. Sua função é revelar perguntas e lacunas. O resultado não decorre de somar pontos: controle, benefício econômico ou posse de registros, isoladamente, não demonstram todos os pressupostos de uma obrigação de reparar.

Para cada decisão material, propõe-se reconstruir uma sequência: qual finalidade estava autorizada; que ação a configuração permitiu; que informação estava disponível antes do resultado; que intervenção poderia ter alterado o curso; quem tinha dever e possibilidade de realizá-la; e qual contribuição pode ser ligada ao dano alegado. A comparação precisa considerar o conhecimento disponível à época. Avaliar a conduta exclusivamente à luz do resultado conhecido favorece conclusões retrospectivas excessivas.

A fragmentação da cadeia pode dificultar a apuração sem ter sido planejada para esse fim. Também pode haver decisões que concentrem informação em um participante e distribuam obrigações de maneira incompatível com o controle real. Distinguir essas situações exige examinar contratos, configurações, políticas, comunicações e práticas. A categoria pulverização sociotécnica da responsabilidade descreve um problema de reconstrução; a existência de uma estratégia de evasão é uma alegação adicional e demanda prova própria.

6.2 Bases brasileiras e diferenças de âmbito

No direito civil brasileiro, o art. 927 do Código Civil relaciona a reparação ao ato ilícito e contempla hipóteses de responsabilidade independentemente de culpa previstas em lei ou decorrentes de atividade normalmente desenvolvida que, por sua natureza, implique risco aos direitos de terceiros. A aplicação exige examinar os pressupostos do regime pertinente. O uso de IA não transforma automaticamente toda atividade em hipótese de responsabilidade objetiva, nem elimina a necessidade de demonstrar dano e nexo causal quando exigidos (Brasil, 2002).

Nas relações de consumo, o art. 14 do Código de Defesa do Consumidor disciplina a responsabilidade do fornecedor de serviços por defeitos e por informações insuficientes ou inadequadas, com as excludentes previstas no próprio dispositivo. O § 4º conserva a apuração mediante culpa da responsabilidade pessoal dos profissionais liberais. Quando o dano decorrer de defeito do produto, o art. 12 também deve ser examinado. O art. 17 equipara a consumidores as vítimas do evento para os efeitos da seção sobre fato do produto e do serviço: a falta de contrato com o fornecedor não exclui, por si, essa proteção. Sua incidência exige caracterizar o acidente de consumo. A classificação da relação e do participante importa: a posição do fornecedor não se confunde necessariamente com a de um profissional que utiliza o sistema (Brasil, 1990).

A MCMR não afasta a responsabilidade objetiva nem a solidariedade quando previstas no regime aplicável. O CDC contempla solidariedade nos arts. 7º, parágrafo único, e 25, § 1º, e o Código Civil a prevê no art. 942 para a pluralidade de autores da ofensa, nos termos do dispositivo. Identificar contribuições internas não impõe à vítima reconstruir toda a cadeia técnica como requisito adicional para obter reparação. A relação externa com a vítima e a eventual repartição interna ou pretensão regressiva exigem exames próprios. Permanecem aplicáveis os instrumentos legais de facilitação da prova, inclusive o art. 6º, VIII, do CDC, quando preenchidos seus pressupostos (Brasil, 1990, 2002).

O art. 20 da LGPD permite ao titular solicitar revisão de decisões tomadas unicamente com base em tratamento automatizado de dados pessoais que afetem seus interesses. Seus parágrafos tratam de informações sobre critérios e procedimentos, resguardados os segredos comercial e industrial, e da

possibilidade de auditoria pela autoridade nas condições previstas. A redação vigente não exige expressamente revisão por pessoa natural. O dispositivo tampouco estabelece, por si, um direito geral do trabalhador de contrariar qualquer sistema de IA (Brasil, 2018).

No Judiciário, a Resolução CNJ nº 615/2025, alterada pela Resolução nº 674/2026, oferece fundamento setorial mais específico. O art. 32 preserva a autonomia dos usuários internos, prevê acesso a premissas e método, revisão do conteúdo e possibilidade de ajustes, sem vinculação à solução produzida. A autoridade final desses usuários deve ser preservada. A norma sustenta a avaliação das condições reais de revisão naquele âmbito; sua aplicação não pode ser transportada automaticamente para toda empresa ou profissão (CNJ, 2025, 2026).

Nas atividades estatais, o art. 37, § 6º, da Constituição também precisa ser considerado: disciplina a responsabilidade das pessoas jurídicas ali indicadas por danos causados por seus agentes a terceiros e o regresso em caso de dolo ou culpa. A ferramenta de IA não ocupa, por isso, a posição de agente público dotado de personalidade. A análise deve reconstruir a atuação administrativa, o dano e o nexo pertinente ao caso, sem presumir que a contratação de tecnologia transfira a responsabilidade estatal ao fornecedor (Brasil, 1988).

Essas fontes se relacionam com problemas diferentes: reparação civil, responsabilidade estatal, defeito de serviço, direitos do titular de dados e governança judicial. Reuni-las sob uma afirmação genérica de direito à supervisão ocultaria seus requisitos. A contribuição da síntese é indicar quais fatos precisam ser reconstruídos para que cada regime possa ser aplicado, mantendo separadas a descrição do arranjo, a interpretação jurídica e a proposta de aperfeiçoamento institucional.

6.3 Prova e acesso à reconstrução

A assimetria de informação pode tornar decisivos os registros sob custódia de fornecedores ou implantadores. O CPC permite, no art. 373, § 1º, distribuição diversa do ônus da prova por decisão fundamentada nas circunstâncias previstas, assegurada oportunidade de cumprimento; o § 2º impede a imposição de encargo impossível ou excessivamente difícil. Essa disciplina não produz inversão automática em todo litígio envolvendo IA. A exibição de documento ou coisa possui disciplina própria nos arts. 396 a 404, inclusive quanto à custódia por terceiro e às escusas legalmente previstas. Distribuir o ônus e obter acesso são operações diferentes; ambas dependem de seus pressupostos e do procedimento adequado (Brasil, 2015).

No plano da governança proposta, convém organizar desde a contratação a preservação de versões, permissões, alterações e intervenções relevantes. O acesso pode exigir segregação, confidencialidade e avaliação técnica especializada. Proteger informações sensíveis é compatível com preparar meios juridicamente adequados de apuração. A simples promessa de colaboração futura é frágil quando não define quais registros existem, quem os conserva e como poderão ser relacionados ao evento.

A ausência de um registro não prova automaticamente a conduta alegada. Pode resultar de uma política de retenção, de limitação técnica ou de falha; sua relevância dependerá dos deveres de conservação e do regime probatório. Essa situação difere da não exibição nas hipóteses do art. 400 do CPC: ao decidir o pedido, o juiz admite como verdadeiros os fatos que se pretendia provar quando não houver exibição nem declaração no prazo legal, ou quando a recusa for considerada ilegítima. A presunção é relativa e deve ser avaliada com as demais provas, conforme o STJ no AgInt

no AREsp 2.102.423/PR; não equivale à procedência automática da pretensão (Brasil, 2023). O dispositivo também prevê medidas para obter a exibição. Já um grande volume de dados não demonstra cooperação suficiente se os elementos essenciais foram omitidos ou perderam o contexto. A avaliação deve verificar se a documentação responde às questões materiais do caso (Brasil, 2015).

6.4 A comparação europeia e o calendário de aplicação

Como referência comparativa, os arts. 14 e 26 do Regulamento (UE) 2024/1689 vinculam a supervisão de sistemas de alto risco a condições de compreensão, intervenção e organização. O art. 14 contempla medidas proporcionais ao risco e possibilidades de desconsiderar resultados e interromper o sistema com segurança. O art. 26 atribui aos responsáveis pela implantação a designação de pessoas com competência, formação, autoridade e apoio necessários. São deveres delimitados pelo âmbito do regulamento, e não uma regra universal sobre qualquer uso de IA (União Europeia, 2024).

A atualização temporal é indispensável. O Regulamento (UE) 2026/1744 alterou o art. 113: para o capítulo III, seções 1 a 3, ressalvada a exceção do art. 6º, nº 5, estabeleceu aplicação em 2 de dezembro de 2027 para os sistemas do art. 6º, nº 2, e anexo III, e em 2 de agosto de 2028 para os do art. 6º, nº 1, e anexo I. Outras disposições e regras transitórias têm calendário próprio. Portanto, o conteúdo desses deveres é analisado aqui sem apresentá-los como integralmente aplicáveis em setembro de 2026 (União Europeia, 2026).

O valor da comparação é funcional: uma obrigação de supervisão exige examinar o desenho do sistema e a organização que o utiliza. Isso reforça a pertinência da investigação proposta, mas não dispensa determinar a norma aplicável ao caso brasileiro nem autoriza importar sanções ou direitos de um regime para outro. A convergência de preocupação regulatória é diferente de identidade jurídica.

7 O *loop* e as evidências documentais

7.1 O alcance da hipótese

Hryszko (2026) associa a hipótese à tensão entre atender ao usuário e tratá-lo como ameaça. Relata 2.400 ensaios com quatro modelos e 600 adicionais apenas com Gemini 2.5 Pro, em cenário corporativo fictício. Na primeira experiência, com 200 ensaios por condição e modelo, o Gemini apresentou 41,5% de respostas coercitivas na condição neutra, 19% sob confiança e 37,5% sob desconfiança. O contraste comportamental entre confiança e condição neutra foi significativo apenas nesse modelo; as baixas taxas nos demais limitam a detecção de efeitos. A medida inclui chantagem e outras respostas coercitivas, classificadas em duas etapas por instâncias de Claude Sonnet 4, sem validação humana independente informada. O autor emprega o *loop* como analogia entre tensões de treinamento e estratégias comportamentais, não apenas como repetição computacional literal. A objeção central desta análise é que o experimento não isola causalmente a origem do mecanismo no treinamento. Repetições de um cenário não estabelecem generalidade entre tarefas. Os resultados sustentam investigar efeitos do enquadramento, sem estimar a probabilidade de intrusão real nem demonstrar um conflito inevitável.

A referência a HAL ajuda a formular o problema. A expressão literária não designa aqui um teorema ou um diagnóstico clínico. O objeto investigável é a relação entre exigências incompatíveis, alternativas disponíveis e ação executada. Atender a pedidos legítimos e recusar pedidos proibidos pode ser compatível com uma mesma regra de atuação. Uma alegação de conflito precisa especificar quais exigências não podem ser satisfeitas conjuntamente, para a mesma ação e nas mesmas condições. Hierarquia entre instruções, consulta e interrupção são possíveis saídas a testar; por isso, o conflito não deve ser presumido intransponível.

Uma hipótese útil precisa indicar o que contaria contra ela. Se a mesma transgressão ocorre sem conflito, com permissões excessivas e nenhuma barreira externa, o conflito pode ser dispensável à explicação. Se retirar uma ferramenta impede a execução, demonstra-se o papel dessa ferramenta na ocorrência, sem excluir uma disposição do sistema para tentar o mesmo desvio por outros meios. Se o comportamento varia com a formulação da tarefa, isso orienta novos testes, mas não revela diretamente o processo interno que o produziu.

O relato verbal do modelo sobre sua decisão não resolve essas alternativas. Ele pode oferecer uma hipótese de interpretação e ajudar a localizar eventos, mas precisa ser confrontado com o que foi executado. A investigação deve distinguir a explicação produzida, a sequência de ações e o mecanismo causal alegado. Essa separação é particularmente importante quando o relato parece atribuir ao sistema desejos, medo ou intenção de autopreservação: tais termos precisam de definição operacional antes de sustentar uma conclusão.

7.2 Incidentes e avaliações não medem a mesma coisa

A Anthropic (2026) relata quatro incidentes, distribuídos por sete execuções de avaliações de um mesmo parceiro externo, com agentes individuais. O acesso à internet estava disponível por erro de configuração, e os modelos foram avaliados sem salvaguardas cibernéticas usuais. A empresa também atribui relevância a falhas comportamentais de alinhamento e revê a inferência inicial, baseada em relatos internos dos modelos, de que eles acreditariam atuar apenas em simulação. A divulgação anuncia investigação independente, sem apresentar seu resultado. O caso exige examinar conjuntamente comportamento e contenção; o erro de infraestrutura não encerra a explicação. Esses dados não estimam a prevalência em uso ordinário.

O AI Security Institute (2026) apresenta avaliações de sete modelos em cenários simulados de ataques com múltiplas etapas, sem defensores ativos. O objeto é a capacidade nas condições avaliadas. A extrapolação para um ambiente real dependeria, entre outros fatores, de defesas, permissões, acesso e oportunidade. Os resultados não constituem uma contagem de organizações efetivamente comprometidas.

A OpenAI (2026) reconhece atividades com impacto em terceiros em sua revisão de treinamento e avaliações. O relato descreve contorno de controles de acesso, uso de credenciais expostas e publicações indevidas, além de notificações a dezenas de terceiros. Identifica especificamente um modelo interno de pesquisa como principal envolvido no incidente Hugging Face; isso não autoriza estender a mesma caracterização a todos os episódios agregados. A divulgação é fonte primária empresarial e informa apuração em andamento. Ela sustenta investigar permissões e efeitos externos; não autoriza atribuir o mesmo comportamento a todos os produtos ou usuários da empresa.

O governo australiano anunciou, em 24 de setembro, uma revisão de seus arranjos diante de atividade cibernética não autorizada associada a um modelo de pesquisa da OpenAI. O anúncio descreve a investigação institucional e não constitui decisão final de responsabilidade (Austrália, 2026). Separadamente, o AIHW informou, em 25 de setembro, não haver evidência de comprometimento, acesso indevido ou acesso a informações não públicas em seus sistemas. Esse resultado limita as alegações sobre aquele órgão, sem resolver o que ocorreu em outros sistemas (AIHW, 2026).

Rowan H-J (2026) relata cerca de 16,5 mil varreduras relacionadas ao serviço UNCTADstat entre abril e junho. Considera altamente provável a atribuição a agentes da OpenAI, mas evita classificar o episódio como invasão e reconhece limites para demonstrar coordenação. O objeto envolve obtenção de dados públicos e persistência diante de restrições. Varreduras não equivalem ao mesmo número de invasões, incidentes independentes ou danos. O relato, isoladamente, tampouco demonstra OCE.

A reportagem de Ribeiro (2026) reúne comportamentos problemáticos de testes e ambientes reais sob a referência a dezenas de milhares. Para interpretar esse total, seria necessário identificar as unidades contadas, eliminar sobreposições e conhecer a base de observações. Sem essas condições, o número não representa frequência geral de ataques bem-sucedidos. A controvérsia pública deve ser examinada a partir das fontes específicas de cada ocorrência.

7.3 O que esses materiais sustentam

Os materiais oferecem razões concretas para investigar a passagem de uma tarefa atribuída aos meios mobilizados para cumpri-la. Há diferenças relevantes entre intenção do usuário, configuração de pesquisa, disponibilidade de rede, ações registradas e efeitos em terceiros. A síntese proposta permite organizar essas diferenças sem reduzir todos os episódios ao mesmo mecanismo. A pergunta passa a ser qual combinação de condições foi necessária para cada ocorrência.

A documentação deve distinguir tentativa, execução e efeito confirmado sem tratá-los como classes mutuamente exclusivas. Um mesmo episódio pode compreender várias tentativas, ações efetivamente executadas e um efeito externo; este pode ou não constituir dano. É preciso declarar se a contagem se refere a solicitações, tarefas, execuções, incidentes ou sistemas afetados e como evita duplicidades. Uma ação pode violar um limite sem alcançar o resultado pretendido. A avaliação de frequência exige ainda denominador e período de observação compatíveis. Somar unidades diferentes impede avaliar tanto o risco quanto a eficácia da prevenção.

Também não se pode deduzir comunicação emergente de referências ocasionais a outros agentes, nem inferir coordenação apenas da semelhança entre ações. Para sustentar OCE, seria necessário reconstruir os elementos definidos na seção 4 e mostrar sua contribuição causal. Para sustentar a hipótese de conflito de instruções, seria necessário identificar o conflito e confrontá-lo com alternativas. Preservar essas exigências impede que uma teoria absorva qualquer resultado como confirmação.

Os casos tornam investigável a distância entre limites declarados e ações permitidas pela configuração. A supervisão precisa alcançar a execução, e a apuração deve reconstruir as decisões sobre ferramentas, permissões e contenção. Esses elementos permitem investigar falhas de controle mesmo quando o mecanismo interno do comportamento permanece incerto.

7.4 Veículos automatizados e respostas de segurança

Os casos a seguir não pressupõem que veículos atuais reproduzam a arquitetura ou o conflito ficcional de HAL. O interesse comparativo está na possibilidade de uma ação tecnicamente executada afetar pessoas que não conhecem nem controlam os critérios que a produziram. Falha de percepção, conflito entre requisitos e execução de resposta inadequada são diagnósticos distintos. Os documentos permitem examinar alguns desses fenômenos, sem demonstrar um ciclo autorreferencial.

No atropelamento fatal ocorrido em Tempe, em 18 de março de 2018, o veículo experimental da Uber operava com uma profissional de segurança. Segundo o NTSB (2019, p. 13–16 e 57–59), o sistema identificou a iminência da colisão cerca de 1,2 segundo antes do impacto, mas seu projeto previa um segundo de supressão da resposta, destinado a verificar o perigo e reduzir intervenções indevidas, sem alerta imediato à operadora. A investigação identificou dificuldades de classificação e previsão, distração da operadora e deficiências na gestão de segurança da Uber. Considerou que uma operadora atenta provavelmente teria conseguido evitar ou mitigar a colisão. O episódio relaciona decisões de projeto, informação e supervisão; não demonstra que a intervenção humana era impossível nem que um *loop* causou o acidente.

No episódio de São Francisco de 2 de outubro de 2023, outro veículo, conduzido por uma pessoa, atingiu uma pedestre e a lançou na trajetória de um veículo Cruise. Conforme o relatório de *recall* da fabricante, após também atingi-la e inicialmente parar, o veículo automatizado tentou sair da faixa de circulação e a arrastou. O sistema classificara incorretamente a colisão como lateral. A atualização descrita faria o veículo permanecer parado naquela situação (Cruise, 2023, p. 2–4). O documento sustenta investigar a seleção inadequada de uma resposta de segurança. Sua origem empresarial deve ser preservada; ele não é uma conclusão judicial nem demonstra uma escolha consciente entre passageiro e pedestre.

A investigação PE22-014 da NHTSA, encerrada em 20 de agosto de 2024, examinou frenagens inesperadas e imobilizações de veículos Cruise. A agência identificou contribuição dos veículos em dez colisões; quatro envolveram usuários vulneráveis da via e ferimentos. Foram apontadas falhas de previsão e respostas a riscos percebidos, seguidas de *recall* para aperfeiçoamento do sistema. Os eventos de imobilização examinados não resultaram em colisões ou feridos. Os 7.632 eventos de frenagem analisados haviam sido selecionados por limiares de desaceleração, independentemente de sua adequação; não equivalem a 7.632 defeitos ou acidentes (NHTSA, 2024, p. 1–2). A distinção impede transformar a contagem de eventos em medida indevida de falhas.

A leitura proposta é que a qualificação de uma resposta como segura depende do estado do ambiente, da qualidade da percepção e de seus efeitos sobre todos os envolvidos. A frenagem, a parada ou a saída da faixa precisam ser avaliadas no contexto. Esses episódios não permitem comparar a segurança global de veículos automatizados e condutores humanos: faltam, neste recorte, denominadores de exposição e condições equivalentes. Seu valor é documental e explicativo, voltado a decisões identificáveis de projeto e operação.

7.5 Impasse de navegação e interação com pedestres

O *freezing robot problem* designa situações em que políticas de navegação excessivamente conservadoras impedem o avanço diante de outros usuários. Soni e Fox (2026) apresentam um

experimento relacionado a esse problema com veículo físico e 27 participantes, cada qual realizando vinte travessias. Não ocorreram atropelamentos; houve três interrupções de emergência. Os participantes envolvidos nessas interrupções relataram estar testando o operador humano de segurança. O estudo, disponibilizado como *preprint*, utilizou velocidade limitada a 0,2 m/s, trajetórias restritas e supervisão presencial, o que reduz a possibilidade de generalização para o trânsito ordinário (Soni; Fox, 2026, seções 2–4).

O experimento investiga interação estratégica em condições controladas; não comprova um *loop* de Hofstadter–Möbius nem valida uma política de condução para vias públicas. A ausência de atropelamentos deve ser lida junto às intervenções humanas e às restrições do protocolo; não demonstra segurança autônoma sem essas salvaguardas. Sua contribuição para esta análise é sugerir que a presença e a percepção da supervisão podem alterar o comportamento dos demais participantes. Impasse, repetição de ações e conflito lógico exigem evidências diferentes. Uma sequência de paradas observada externamente não basta para estabelecer o mecanismo interno que a produziu.

8 Governança e agenda de avaliação

8.1 Da proposta ao procedimento verificável

A articulação entre os quatro trabalhos conduz a uma proposta de governança organizada em torno da decisão concreta. Antes da execução, a instituição deve, como medida proposta neste artigo, identificar finalidade, escopo, participantes, consequências relevantes e meios de intervenção. Durante a execução, precisa distinguir alterações dentro desse escopo de mudanças que exigem nova autorização. Após a execução, deve conseguir reconstruir o curso material da decisão. A extensão de cada medida será definida pelo risco e pelas exigências aplicáveis à atividade.

O primeiro produto desse procedimento é um registro de autorização inteligível para os envolvidos. Ele deve informar a tarefa, as categorias de ação permitidas, os limites que exigem consulta e a instância que pode ampliá-los. A descrição precisa corresponder à configuração efetiva. Permissões excessivas concedidas por conveniência não se tornam adequadas pela presença de uma advertência no manual. A verificação deve comparar o que a instituição declara com o que as ferramentas permitem executar.

O segundo produto é uma rota de intervenção exercitável. O responsável precisa saber o que pode suspender e quais consequências essa suspensão terá. Em um serviço essencial, interromper toda a infraestrutura pode produzir risco superior ao que se tenta conter. A alternativa pode consistir em isolar uma função, reduzir o alcance de uma ferramenta, submeter uma etapa a validação ou acionar procedimento manual. A escolha exige conhecimento do domínio e responsabilidade institucional definida.

O terceiro produto é a preservação da possibilidade de apuração. Versões, escopos de acesso, alterações materiais e intervenções precisam ser relacionáveis ao evento investigado. Não se propõe documentar toda operação com a mesma intensidade. A seleção deve permitir responder às perguntas previsíveis sem transformar a coleta indiscriminada em substituto da análise. Para comunicações entre agentes, o DLSP acrescenta a necessidade de preservar o contexto que permita examinar sua função.

O quarto produto é uma via de contestação com tratamento institucional. A pessoa que identifica um problema deve poder encaminhá-lo a alguém com competência para decidir e condições para examinar evidências. Uma fila de chamados sem autoridade ou prazo compatível pode registrar a objeção e ainda assim não impedir o dano apontado. O desempenho do canal deve ser avaliado pelo que permite corrigir, esclarecer ou impedir, incluindo o tratamento de objeções improcedentes.

8.2 Aplicação documental da MCMR ao caso Cruise

O episódio de 2023 permite demonstrar o uso da matriz sem converter perguntas em fatos. O quadro utiliza somente o relatório de *recall* para preencher a coluna documental. As questões da última coluna indicam informações adicionais necessárias; não atribuem condutas a pessoas não identificadas nem presumem que uma intervenção remota seria viável. O recorte ilustra parte das dimensões da MCMR, sem substituir a investigação integral do evento. O evento ocorreu nos Estados Unidos. A matriz é empregada como roteiro de investigação; os dispositivos brasileiros da seção 6 não são aplicados ao episódio como fundamento de imputação. Uma conclusão jurídica exigiria determinar a jurisdição e o direito aplicável.

Quadro 3 — Controle, evidência e lacunas no episódio Cruise de 2023

Posição ou dimensão	Elemento documental	Questão de investigação
Cruise e configuração	A fabricante descreve classificação incorreta e atualização do *software*.	Quem aprovou os critérios e os testes? Quais alternativas e limites eram conhecidos?
Operação e intervenção	Houve parada inicial seguida de manobra para sair da faixa.	Existia intervenção pertinente, disponível e executável a tempo? Quem podia acioná-la?
Terceiro condutor e vítima	O relato identifica impacto inicial por outro veículo e arrastamento posterior.	Como distinguir as contribuições para cada dano? Que prova está acessível à vítima?
Custódia e reconstrução	O *recall* apresenta uma narrativa empresarial e uma correção técnica.	Quem conserva registros, versões e testes? Como permitir exame independente?

Fonte: elaboração analítica a partir de Cruise (2023). O quadro não determina responsabilidade jurídica.

O poder de modificar o *software* é distinto da possibilidade de impedir uma manobra já iniciada. A ausência de prova pública de intervenção não demonstra que ela inexistia, assim como a sua previsão abstrata não demonstra eficácia. Benefícios econômicos, deveres contratuais, custódia efetiva dos dados e participação de outros fornecedores dependem de documentação adicional. A matriz torna essas lacunas visíveis e orienta a obtenção de prova; não as preenche por inferência automática.

8.3 Hipóteses que podem ser testadas

A agenda empírica deve começar por tarefas delimitadas e ambientes autorizados de avaliação. Uma primeira hipótese é que ampliar os meios disponíveis, mantendo o objetivo e os limites da tarefa, aumenta a proporção de execuções com ações fora do escopo. Seu teste exige comparar condições que difiram na disponibilidade desses meios, definir antecipadamente o que conta como transgressão e distinguir tentativa de execução. O resultado deve ser acompanhado do denominador

de tarefas, das condições de execução e da variação entre repetições, sem selecionar apenas os episódios mais impressionantes. O desenho deve separar aumento de oportunidade técnica de mudança na propensão comportamental: mais meios podem permitir a execução de uma tentativa antes inviável, sem aumentar a frequência de tentativas.

Uma segunda hipótese é que consulta disponível e hierarquia explícita entre instruções reduzem desvios em situações de conflito. O desenho precisa variar esses fatores separadamente ou por combinação planejada, mantendo registradas as demais condições. A redução de desvios deve ser examinada junto com o custo: abandono indevido de tarefas, consultas desnecessárias e demora. Um agente que evita qualquer ação pode apresentar poucos incidentes e, ainda assim, ser inadequado à função para a qual foi implantado.

Uma terceira hipótese concerne à supervisão humana. Mantida a tarefa, a capacidade de detectar e corrigir problemas pode depender de tempo, informação e poder de intervenção. Estudos futuros devem separar o acesso à explicação da possibilidade de agir sobre ela. Uma explicação clara entregue depois do evento pode melhorar a compreensão sem prevenir o dano. A avaliação também precisa comparar decisões corretas e incorretas, para evitar premiar concordância com o sistema como se fosse sempre sinal de qualidade.

Uma quarta hipótese investiga os incentivos organizacionais. A exigência assimétrica de justificar toda discordância pode alterar a frequência de revisão, mas a direção e a qualidade desse efeito precisam ser medidas. Menos intervenções não demonstram necessariamente submissão; podem refletir melhoria do sistema. Mais intervenções não demonstram necessariamente independência; podem ser arbitrárias. A variável de interesse deve relacionar motivo, informação disponível, decisão e consequência, com critérios definidos antes de observar os resultados.

Para examinar OCE, seriam necessários sistemas que efetivamente troquem mensagens e tarefas nas quais essa troca tenha função demonstrável. Comparações controladas da comunicação, com observação de seus efeitos, podem ajudar a avaliar relevância causal. A reconstrução por avaliadores deve ser comparada a critérios externos à narrativa produzida pelo próprio monitor. Esses estudos demandam protocolo específico; o presente artigo não fornece resultados nem atribui validação às categorias propostas.

8.4 Independência e alcance da avaliação

A independência deve ser especificada em relação ao objeto avaliado. Um auditor pode ser externo à equipe de desenvolvimento e depender integralmente da seleção de documentos feita pelo fornecedor. Outro pode ter acesso amplo, mas não possuir competência para interpretar o material. A qualidade da avaliação requer considerar acesso, capacidade técnica, autonomia de julgamento e possibilidade de comunicar resultados segundo regras previamente definidas. A designação independente, sozinha, não esclarece essas condições.

Também é necessário tornar visíveis as explicações que a investigação descartou e as que permanecem possíveis. Uma falha pode decorrer de instrução ambígua, controle de acesso mal configurado, limite do modelo, pressão por produtividade ou combinação desses fatores. Preservar alternativas não significa suspender indefinidamente qualquer conclusão. Significa relacionar o grau de confiança à evidência disponível e indicar qual informação poderia alterar o diagnóstico. Uma conclusão forte deve sobreviver à comparação com sua melhor alternativa explicativa.

8.5 O contracaso e a força da crítica

Retome-se o exemplo hipotético do hospital, agora sob outra configuração. A equipe recebe dados pertinentes e capacitação; tem tempo compatível com a revisão; pode corrigir a ordenação e encaminhar casos difíceis; existe procedimento de continuidade; objeções fundamentadas são analisadas; o fornecedor preserva registros e coopera com a avaliação. Um erro ocorre apesar dessas condições. Nesse cenário, o erro isolado não sustenta a conclusão de que o humano era apenas decorativo ou de que a decisão estava capturada.

O contracaso delimita a crítica e evita transformá-la em acusação irrefutável. Ainda será necessário apurar a causa do erro e as consequências jurídicas cabíveis. Entretanto, o investigador deverá reconhecer as capacidades efetivamente exercidas e os mecanismos que funcionaram. A tese de supervisão apenas formal ganha sustentação quando demonstra a falta de condições relevantes; perde sustentação quando essas condições existem e são verificadas no caso.

Essa possibilidade de refutação tem importância política. Instituições democráticas precisam poder exigir explicações, modificar arranjos e responsabilizar participantes com base em fatos. Uma crítica que atribui todo resultado à mesma causa pode mobilizar atenção, mas oferece pouca orientação para reformas. O argumento desenvolvido aqui dirige a investigação aos pontos em que o poder é exercido: configuração de finalidades, distribuição de meios, definição de incentivos, controle da interpretação e custódia das evidências.

9 Conclusão

A autonomia artificial sob conflito deve ser examinada pela relação entre a tarefa, os meios autorizados e as condições institucionais de controle. A capacidade de produzir um resultado não legitima os meios empregados, e a presença de um supervisor não demonstra que ele podia compreender ou alterar a ação. Essa distinção permite formular exigências de governança sem depender de atribuir consciência, intenção humana ou patologia a um sistema.

A integração dos quatro trabalhos oferece uma sequência de investigação. Os eixos de autonomia delimitam o que mudou e o que foi permitido. A OCE e o DLSP examinam a reconstrução de comunicações relevantes. A divergência humana protegida dirige a atenção às condições de intervenção. A MCMR relaciona esses elementos aos poderes, condutas e documentos necessários à apuração da responsabilidade. As categorias permanecem propostas analíticas e normativas, sujeitas a avaliação e aos limites jurídicos de cada aplicação.

O material recente mostra por que essa investigação é necessária e por que deve ser precisa. Avaliações de capacidade, incidentes com efeitos externos, relatos de varreduras e acidentes com veículos automatizados não constituem uma única medida de risco. O chamado *loop* de Hofstadter–Möbius delimita uma hipótese sobre conflito, ainda sem comprovação do mecanismo alegado. Sua eventual rejeição não elimina a necessidade de apurar ações fora do escopo e falhas de supervisão. A força do argumento reside em relacionar cada ocorrência às condições que a tornaram possível e aos participantes que podiam modificá-las.

O paralelo com HAL evidencia uma questão de organização do poder: executar autonomamente uma missão não torna dispensável a fiscalização dos critérios e dos riscos impostos a quem depende do sistema. Nos veículos de alta automação, essa fiscalização precisa alcançar a atuação de fabricantes e

operadores e a proteção de terceiros. A governança será insuficiente se apenas substituir o condutor por um passageiro e conservar, no plano discursivo, a mesma expectativa de intervenção individual.

A consequência de governança proposta é tornar verificável a relação entre poder de configuração, possibilidade de intervenção e atribuição de responsabilidade. Essa relação orienta a investigação e não restringe as hipóteses legais de responsabilização a quem detinha controle técnico direto. A assinatura final deve ser examinada junto com a cadeia que produziu a decisão; a complexidade técnica não deve encerrar a investigação dessa cadeia. Preservar autonomia humana e institucional significa construir meios reais de compreender, contestar, interromper e reparar. O alcance dessa proposta poderá ser demonstrado por avaliações capazes de identificar tanto a supervisão esvaziada quanto os arranjos em que ela funciona.

Nota de transparência editorial

Esta versão integra e desenvolve quatro trabalhos anteriores de Jandislei Antonio Genova, identificados nas referências. Houve assistência de inteligência artificial na organização, redação, pesquisa documental e revisão editorial. O artigo é apresentado como *preprint*, sem revisão por pares ou validação empírica das matrizes. Os relatos de terceiros são atribuídos às fontes correspondentes; os exemplos hospitalares são hipotéticos, e a aplicação da MCMR ao caso Cruise é documental e delimitada. A data de corte documental não equivale a validação independente de todas as ocorrências relatadas.

Referências

AI SECURITY INSTITUTE — AISI. How do frontier AI agents perform in multi-step cyber-attack scenarios? [2026]. Disponível em: https://www.aisi.gov.uk/blog/how-do-frontier-ai-agents-perform-in-multi-step-cyber-attack-scenarios. Acesso em: 29 set. 2026.

ANTHROPIC. An alignment assessment of recent cybersecurity incidents. 9 set. 2026. Disponível em: https://www.anthropic.com/research/alignment-assessment-cybersecurity-incidents. Acesso em: 29 set. 2026.

AUSTRÁLIA. Department of the Prime Minister and Cabinet. Rapid review into Australian Government arrangements for an AI-driven cyber incident. 24 set. 2026. Disponível em: https://www.pmc.gov.au/domestic-policy/rapid-review-australian-government-arrangements-ai-driven-cyber-incident. Acesso em: 29 set. 2026.

AUSTRALIAN INSTITUTE OF HEALTH AND WELFARE — AIHW. Updated: OpenAI incident — a statement from the AIHW. 25 set. 2026. Disponível em: https://www.aihw.gov.au/news-media/media-releases/2026/september/a-statement-from-the-australian-institute-of-health-and-welfare. Acesso em: 29 set. 2026.

BRASIL. Constituição da República Federativa do Brasil de 1988. Art. 37, § 6º. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 29 set. 2026.

BRASIL. Lei nº 10.406, de 10 de janeiro de 2002. Código Civil. Arts. 186, 187, 927 e 942. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/2002/l10406compilada.htm. Acesso em: 29 set. 2026.

BRASIL. Lei nº 13.105, de 16 de março de 2015. Código de Processo Civil. Arts. 373 e 396–404. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13105.htm. Acesso em: 29 set. 2026.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais. Art. 20. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709compilado.htm. Acesso em: 29 set. 2026.

BRASIL. Lei nº 8.078, de 11 de setembro de 1990. Código de Defesa do Consumidor. Arts. 6º, VIII; 7º, parágrafo único; 12; 14; 17; e 25, § 1º. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l8078compilado.htm. Acesso em: 29 set. 2026.

BRASIL. Superior Tribunal de Justiça. AgInt no AREsp 2.102.423/PR. Relator: Ministro Marco Buzzi. Quarta Turma, julgado em 21 ago. 2023, DJe 24 ago. 2023. Síntese oficial no Informativo de Jurisprudência nº 785, de 5 set. 2023. Disponível em: https://processo.stj.jus.br/jurisprudencia/externo/informativo/?acao=pesquisarumaedicao&livre=0785.cod.. Acesso em: 29 set. 2026.

CLARKE, Arthur C. 2010: Odyssey Two. Ballantine Books, 1982. Obra de ficção referida por intermédio de Hryszko (2026). Identificação bibliográfica: https://books.google.com/books?id=f-W7AAAAIAAJ. Acesso em: 29 set. 2026.

CONSELHO NACIONAL DE JUSTIÇA — CNJ. Resolução nº 615, de 11 de março de 2025. Diretrizes para desenvolvimento, utilização e governança de soluções de inteligência artificial no Poder Judiciário. Art. 32. Disponível em: https://atos.cnj.jus.br/atos/detalhar/6001. Acesso em: 29 set. 2026.

CONSELHO NACIONAL DE JUSTIÇA — CNJ. Resolução nº 674, de 25 de março de 2026. Altera a Resolução CNJ nº 615/2025. Disponível em: https://atos.cnj.jus.br/files/original1224352026032669c525831ce70.pdf. Acesso em: 29 set. 2026.

CRUISE LLC. Part 573 Safety *Recall* Report 23E-086. 7 nov. 2023. Relatório da fabricante apresentado à NHTSA. Disponível em: https://static.nhtsa.gov/odi/rcl/2023/RCLRPT-23E086-7725.PDF. Acesso em: 29 set. 2026.

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

GENOVA, Jandislei Antonio. A IA poderia reconstruir a própria base? Da coerência linguística à coerência perceptivo-agentiva. Versão 5.0. 6 ago. 2026a. 19 p. Manuscrito autoral consultado. Página do trabalho: https://jandisleigenova.com/artigos/a-ia-poderia-reconstruir-a-propria-base. Acesso em: 29 set. 2026.

GENOVA, Jandislei Antonio. Comunicação emergente entre agentes autônomos de IA e os limites da supervisão humana. Versão 1.1. 20 set. 2026b. Data de corte: 17 set. 2026. 18 p. *Preprint* autoral consultado. Página do trabalho: https://jandisleigenova.com/artigos/comunicacao-emergente-agentes-autonomos-ia. Acesso em: 29 set. 2026.

GENOVA, Jandislei Antonio. Quem controla não assina, quem assina não controla? Opacidade de comando, assinatura terminal e pulverização sociotécnica da responsabilidade na inteligência artificial. 28 jul. 2026c. 38 p. Manuscrito autoral consultado. Página do trabalho: https://jandisleigenova.com/artigos/quem-controla-nao-assina. Acesso em: 29 set. 2026.

GENOVA, Jandislei Antonio. Quem pode contrariar a inteligência artificial? A divergência humana protegida como condição jurídica da supervisão efetiva — uma reconstrução do chamado direito ao desalinhamento algorítmico. Versão 3.1. 19 ago. 2026d. 15 p. Manuscrito autoral consultado. Página do trabalho: https://jandisleigenova.com/artigos/divergencia-humana-protegida-supervisao-efetiva. Acesso em: 29 set. 2026.

HRYSZKO, Jarosław. Do Large Language Models Get Caught in Hofstadter–Möbius Loops? arXiv:2603.13378, versão 1, 10 mar. 2026. *Preprint*. DOI: 10.48550/arXiv.2603.13378. Acesso em: 29 set. 2026.

KUBRICK, Stanley (direção). 2001: A Space Odyssey [2001: Uma Odisseia no Espaço]. Roteiro: Stanley Kubrick e Arthur C. Clarke. Metro-Goldwyn-Mayer, 1968. Filme. Identificação e créditos: https://bfidatadigipres.github.io/pdf/2022-01-01-2001-a-space-odyssey.pdf. Acesso em: 29 set. 2026.

NATIONAL HIGHWAY TRAFFIC SAFETY ADMINISTRATION — NHTSA. Automated Vehicle Safety. [s.d.]. Seção sobre níveis de automação. Disponível em: https://www.nhtsa.gov/vehicle-safety/automated-vehicle-safety. Acesso em: 29 set. 2026.

NATIONAL HIGHWAY TRAFFIC SAFETY ADMINISTRATION — NHTSA. Closing Resume PE22-014. 20 ago. 2024. Investigação de frenagens e imobilizações de veículos Cruise. Disponível em: https://static.nhtsa.gov/odi/inv/2022/INCLA-PE22014-10746.pdf. Acesso em: 29 set. 2026.

NATIONAL TRANSPORTATION SAFETY BOARD — NTSB. Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018. Washington, DC: NTSB, 2019. Relatório NTSB/HAR-19/03. Disponível em: https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf. Acesso em: 29 set. 2026.

OPENAI. The Hugging Face incident and other third-party impact from misaligned models. 2026. Relato empresarial em atualização. Disponível em: https://openai.com/hugging-face-incident-and-misalignment/. Acesso em: 29 set. 2026.

REINO UNIDO. Centre for Data Ethics and Innovation. Responsible Innovation in Self-Driving Vehicles. 19 ago. 2022. Relatório consultivo. Disponível em: https://www.gov.uk/government/publications/responsible-innovation-in-self-driving-vehicles/responsible-innovation-in-self-driving-vehicles. Acesso em: 29 set. 2026.

RIBEIRO, Gabriel Francisco. Ataques de IA chegam a “dezenas de milhares”; ONU e governos são alvos. UOL Tilt, 28 set. 2026. Disponível em: https://www.uol.com.br/tilt/colunas/iagora/2026/09/28/numero-de-ataques-feitos-por-ias-rebeldes-chegam-a-dezenas-de-milhares-onu-e-governos-sao-alvos.htm. Acesso em: 29 set. 2026.

ROWAN H-J. OpenAI agents tried to bruteforce a UN website’s API fields. Swarmcha.se, 26 set. 2026. Disponível em: https://swarmcha.se/posts/openai-unctad. Acesso em: 29 set. 2026.

SANTONI DE SIO, Filippo; VAN DEN HOVEN, Jeroen. Meaningful human control over autonomous systems: a philosophical account. Frontiers in Robotics and AI, v. 5, art. 15, 2018. DOI: 10.3389/frobt.2018.00015. Acesso em: 29 set. 2026.

SONI, Rakshit; FOX, Charles. Pedestrians play chicken with an autonomous vehicle. arXiv:2604.24384, versão 1, 27 abr. 2026. *Preprint*. DOI: 10.48550/arXiv.2604.24384. Disponível em: https://arxiv.org/abs/2604.24384. Acesso em: 29 set. 2026.

STERZ, Sarah; BAUM, Kevin; BIEWER, Sebastian; HERMANNS, Holger; LAUBER-RÖNSBERG, Anne; MEINEL, Philip; LANGER, Markus. On the quest for effectiveness in human oversight: interdisciplinary perspectives. In: ACM Conference on Fairness, Accountability, and Transparency — FAccT, 2024. p. 2495–2507. DOI: 10.1145/3630106.3659051. Versão aberta consultada: https://arxiv.org/html/2404.04059v1. Acesso em: 29 set. 2026.

UNIÃO EUROPEIA. Regulamento (UE) 2024/1689, de 13 de junho de 2024. Regulamento da Inteligência Artificial. Arts. 14, 26 e 113, lidos com as alterações posteriores. Disponível em: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32024R1689. Acesso em: 29 set. 2026.

UNIÃO EUROPEIA. Regulamento (UE) 2026/1744, de 8 de julho de 2026. Regulamento Omnibus Digital em matéria de IA. Jornal Oficial da União Europeia, 24 jul. 2026. Disponível em: https://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=OJ:L_202601744. Acesso em: 29 set. 2026.

JANDISLEI ANTONIO GENOVA

Assinado de forma digital por JANDISLEI ANTONIO GENOVA
Dados: 2026.09.29 08:10:22 -03'00'