ReGeo ISSN: 2177-3246

A INSUFICIÊNCIA DA SUPERVISÃO HUMANA ISOLADA E A GOVERNANÇA ALGORÍTMICA NAS POLÍTICAS PÚBLICAS DO PODER JUDICIÁRIO

THE INADEQUACY OF HUMAN SUPERVISION ALONE AND ALGORITHMIC GOVERNANCE IN THE JUDICIARY'S PUBLIC POLICIES

LA INSUFICIENCIA DE LA SUPERVISIÓN HUMANA AISLADA Y LA GOBERNANZA ALGORÍTMICA EN LAS POLÍTICAS PÚBLICAS DEL PODER JUDICIAL

https://doi.org/10.56238/revgeov17n9-176

Data de submissão: 28/08/2026 **Data de publicação:** 28/09/2026

Andrey Mendes Santos
Doutorando em Direito
Instituição: Centro Universitário de Brasília (CEUB)
E-mail: andreyjcs@gmail.com
Orcid: https://orcid.org/0000-0002-4788-0653
Lattes: http://lattes.cnpq.br/1516844653649284

RESUMO

A incorporação de inteligência artificial generativa ao trabalho judicial exige avaliar as condições em que a revisão humana consegue preservar a fundamentação e o contraditório. Este artigo investiga por que a presença de um revisor, isoladamente, não constitui garantia suficiente de controle e quais consequências decorrem dessa limitação para as políticas públicas judiciárias. Desenvolve-se pesquisa jurídico-dogmática de abordagem qualitativa e sociotécnica, com revisão narrativa da literatura, análise documental e exame ilustrativo de decisões estrangeiras sobre referências jurisprudenciais falsas. O referencial articula a concepção de processo como procedimento em contraditório, de Elio Fazzalari, à literatura sobre confiança em automação, consciência situacional e governança de inteligência artificial. A análise distingue os deveres previstos na Resolução CNJ nº 615/2025 das condições organizacionais necessárias à sua efetividade. Sustenta-se que a revisão pode se tornar meramente formal quando faltam acesso às fontes, tempo, competência e autoridade para rejeitar resultados automatizados. As evidências examinadas não demonstram a ineficácia universal da supervisão nem permitem estimar a frequência desses problemas no Judiciário brasileiro. Fundamentam, contudo, uma proposta de supervisão ampliada que combina rastreabilidade documental, recuperação de fontes verificáveis, avaliação proporcional ao risco e alocação institucional de tempo para conferência. A contribuição consiste em traduzir a crítica à supervisão isolada em requisitos de implementação e indicadores de avaliação, subordinando os ganhos de produtividade à possibilidade de reconstruir, contestar e revisar as razões da decisão judicial.

Palavras-chave: Inteligência Artificial Generativa. Supervisão Humana. Governança Algorítmica. Contraditório. Políticas Públicas Judiciárias.

ABSTRACT

Integrating generative artificial intelligence into judicial work requires assessing the conditions under which human review can preserve reasoned adjudication and adversarial participation. This article examines why the mere presence of a reviewer is insufficient to ensure effective control and explores the implications for judicial public policy. It adopts a qualitative, sociotechnical approach to doctrinal legal research, combining a narrative literature review, documentary analysis, and an illustrative

examination of foreign judicial decisions concerning fabricated legal authorities. The theoretical framework connects Elio Fazzalari's conception of proceedings structured by adversarial participation with research on trust in automation, situation awareness, and AI governance. The analysis distinguishes the duties established by Brazilian National Council of Justice Resolution No. 615/2025 from the organizational conditions required for their effectiveness. It argues that review may become merely formal when reviewers lack access to sources, time, expertise, or authority to reject automated outputs. The evidence does not establish that human oversight is universally ineffective or allow the prevalence of these problems within the Brazilian judiciary to be estimated. It nevertheless supports a proposal for amplified oversight combining documentary traceability, retrieval of verifiable sources, risk-proportionate evaluation, and institutional allocation of verification time. The contribution translates the critique of isolated oversight into implementation requirements and evaluation indicators, making productivity gains conditional on the ability to reconstruct, challenge, and review the reasons supporting judicial decisions.

Keywords: Generative Artificial Intelligence. Human Oversight. Algorithmic Governance. Adversarial Participation. Judicial Public Policy.

RESUMEN

La incorporación de inteligencia artificial generativa al trabajo judicial exige evaluar las condiciones en que la revisión humana consigue preservar la fundamentación y la contradicción. Este artículo investiga por qué la presencia de un revisor, aisladamente, no constituye garantía suficiente de control y qué consecuencias se derivan de esa limitación para las políticas públicas judiciales. Se desarrolla investigación jurídico-dogmática de abordaje cualitativo y sociotécnico, con revisión narrativa de la literatura, análisis documental y examen ilustrativo de decisiones extranjeras sobre referencias jurisprudenciales falsas. El referencial articula la concepción del proceso como procedimiento en contradicción, de Elio Fazzalari, a la literatura sobre confianza en automatización, conciencia situacional y gobernanza de inteligencia artificial. El análisis distingue los deberes previstos en la Resolución CNJ nº 615/2025 de las condiciones organizacionales necesarias para su efectividad. Se sostiene que la revisión puede tornarse meramente formal cuando faltan acceso a las fuentes, tiempo, competencia y autoridad para rechazar resultados automatizados. Las evidencias examinadas no demuestran la ineficacia universal de la supervisión ni permiten estimar la frecuencia de esos problemas en el Poder Judicial brasileño. Fundamentan, sin embargo, una propuesta de supervisión ampliada que combina trazabilidad documental, recuperación de fuentes verificables, evaluación proporcional al riesgo y asignación institucional de tiempo para constatación. La contribución consiste en traducir la crítica a la supervisión aislada en requisitos de implementación e indicadores de evaluación, subordinando las ganancias de productividad a la posibilidad de reconstruir, cuestionar y revisar las razones de la decisión judicial.

Palabras clave: Inteligencia Artificial Generativa. Supervisión Humana. Gobernanza Algorítmica. Contradicción. Políticas Públicas Judiciales.

1 INTRODUÇÃO

A incorporação de inteligência artificial generativa ao Poder Judiciário suscita um problema que não se esgota na capacidade de produzir textos mais rapidamente. Quando uma ferramenta resume os autos, seleciona precedentes ou sugere uma minuta, ela pode influenciar as informações a partir das quais o julgador compreende a controvérsia. O ganho operacional precisa, portanto, ser examinado em conjunto com as condições de formação e controle da decisão. Uma síntese eficiente que omite uma alegação decisiva pode reduzir o tempo de leitura e, simultaneamente, aumentar o risco de uma fundamentação inadequada.

O interesse público na duração razoável do processo não autoriza tratar a qualidade da prestação jurisdicional como variável residual. A Constituição exige motivação das decisões e assegura devido processo legal, contraditório e ampla defesa. Esses parâmetros vinculam o desenho dos instrumentos de apoio ao julgamento, embora não imponham uma tecnologia específica. O problema jurídico surge quando os meios utilizados para organizar a informação dificultam o cumprimento das garantias que deveriam viabilizar.

A supervisão humana é uma resposta necessária a esse problema, mas sua previsão abstrata deixa perguntas em aberto. Quem revisa o resultado dispõe de acesso ao documento original? Consegue identificar uma fonte inventada ou uma omissão relevante? Recebe tempo para fazer essa conferência? Pode rejeitar a sugestão sem sofrer penalização organizacional? Tais questões deslocam o debate da simples presença de uma pessoa para as condições concretas em que ela exerce controle.

Formula-se, assim, a seguinte pergunta de pesquisa: quais condições tornam insuficiente a supervisão humana isolada de inteligência artificial generativa no apoio à decisão judicial e como essas condições devem orientar as políticas públicas de governança do Poder Judiciário? A hipótese é que a revisão meramente nominal não oferece garantia suficiente quando a fluência do texto automatizado se combina com baixa rastreabilidade, restrição de tempo e incentivos predominantemente quantitativos. A insuficiência afirmada é condicional e institucional: não equivale à impossibilidade de supervisão eficaz em qualquer contexto.

O objetivo geral é construir parâmetros jurídicos e organizacionais para uma supervisão efetiva. Os objetivos específicos são distinguir riscos técnicos e cognitivos, relacioná-los às exigências processuais, examinar o alcance do marco regulatório do CNJ e propor instrumentos de implementação suscetíveis de avaliação. Preserva-se a responsabilidade decisória do magistrado e reconhece-se que ferramentas adequadamente delimitadas podem contribuir para o trabalho judicial.

A contribuição pretendida não reside na descoberta de que modelos de linguagem podem errar ou de que usuários podem confiar excessivamente na automação. Consiste em articular esses

problemas à estrutura do contraditório e em formular um critério de suficiência institucional da supervisão. Sob esse critério, o controle precisa permitir reconhecer o erro relevante, conferir suas fontes e impedir sua incorporação ao ato. A seção final desenvolve essa proposta em termos de responsabilidades, recursos e indicadores.

2 DELIMITAÇÃO METODOLÓGICA

O estudo é jurídico-dogmático, qualitativo e de orientação sociotécnica. Emprega revisão narrativa de literatura e análise documental para desenvolver um argumento normativo sobre a integração de ferramentas generativas ao apoio à decisão judicial. O recorte principal é o processo civil brasileiro; referências a outros domínios servem à comparação ou à identificação de riscos, sem pressupor identidade entre regimes jurídicos. Não se realiza levantamento de gabinetes, experimento com magistrados ou avaliação direta de sistemas em produção.

O núcleo documental compreende a Constituição de 1988, o Código de Processo Civil, a Lei Geral de Proteção de Dados Pessoais e a Resolução CNJ nº 615/2025, consultada em sua versão compilada com a alteração da Resolução nº 674/2026. O Regulamento (UE) 2024/1689, a Recomendação da UNESCO sobre Ética da Inteligência Artificial e o AI Risk Management Framework do NIST são utilizados como referências comparativas, respeitadas suas diferentes naturezas e âmbitos de aplicação. A atualização documental desta versão encerrou-se em 22 de setembro de 2026.

A seleção bibliográfica foi intencional e orientada pelos problemas do manuscrito: limites da geração textual, confiança em automação, supervisão, rastreabilidade e fundamentação judicial. Foram priorizadas fontes com autoria e identificação bibliográfica verificáveis, textos acadêmicos pertinentes ao mecanismo discutido e documentos normativos oficiais. As buscas de conferência foram dirigidas por títulos, autores, identificadores e atos normativos, em portais institucionais, repositórios acadêmicos e páginas de editoras. Trata-se de revisão narrativa, sem protocolo de busca exaustiva, contagem de registros ou pretensão de representatividade do conjunto da literatura.

As fontes cumprem funções distintas. A teoria processual oferece o parâmetro de legitimidade da decisão; os estudos de interação humano-automação indicam mecanismos e riscos possíveis; os trabalhos sobre modelos de linguagem delimitam problemas de desempenho; os atos normativos definem obrigações e espaços de escolha institucional. O argumento resulta da articulação crítica entre essas funções, e não da equivalência entre uma evidência experimental e uma conclusão jurídica.

São examinados, de forma ilustrativa, Mata v. Avianca, decidido em 2023 pelo Tribunal Federal do Distrito Sul de Nova York, e a decisão conjunta de 2025 em Ayinde e Al-Haroun, da High Court inglesa. A escolha considera a disponibilidade de decisões oficiais e sua pertinência para o dever de

verificar fontes. Os eixos de leitura são a natureza da informação defeituosa, a responsabilidade de conferência e a resposta institucional. Não se trata de amostra jurisprudencial nem de estudo comparado exaustivo. Os episódios envolvem atuação profissional perante os tribunais e não comprovam padrões de comportamento dos magistrados brasileiros.

A análise distingue três planos: descrição de normas e documentos; inferências sobre mecanismos sociotécnicos; proposições de política pública. Essa separação impede que hipóteses plausíveis sejam apresentadas como resultados medidos. A literatura mobilizada também contém estudos em inglês e em ambientes distintos da jurisdição nacional, além de preprints sujeitos a revisão. Por isso, a conclusão sustenta a necessidade de condições institucionais de controle, sem estimar a magnitude dos riscos nem declarar comprovada a superioridade do modelo proposto.

3 CONTRADITÓRIO E ARQUITETURA DA DECISÃO JUDICIAL

A concepção de processo como procedimento em contraditório, desenvolvida por Fazzalari (2006), fornece o eixo jurídico da investigação. Nessa perspectiva, a participação dos destinatários dos efeitos do provimento integra a estrutura de sua formação. A qualidade processual não pode ser reduzida à correção aparente do texto final. Interessa também se as partes puderam participar da construção das razões que orientaram a decisão. A aplicação dessa concepção aos ambientes algorítmicos é uma elaboração do presente artigo, e não uma formulação tecnológica atribuída ao autor.

O suporte digital pode modificar as condições materiais dessa participação. Interfaces definem quais documentos são exibidos primeiro; mecanismos de busca ordenam resultados; resumos escolhem fatos e argumentos. Essas operações não substituem necessariamente o julgamento, mas interferem no campo de informação disponível ao decisor. A questão é saber quando a mediação facilita o exame do processo e quando passa a limitar, sem controle adequado, aquilo que será considerado relevante.

A fundamentação deve responder a razões processualmente relevantes, e não apenas reproduzir uma estrutura textual reconhecível. Se uma síntese automatizada elimina uma objeção capaz de alterar a solução, a leitura exclusiva dessa síntese pode impedir que a decisão a enfrente. Se a pesquisa apresenta somente julgados favoráveis a uma conclusão inicial, o conjunto de alternativas fica artificialmente estreito. Em ambos os exemplos, a regularidade gramatical da minuta não permite identificar o defeito. A conferência depende do retorno ao material processual e às fontes jurídicas.

Convém distinguir a rastreabilidade documental da fundamentação jurídica. A primeira permite identificar documentos, versões e intervenções relevantes; a segunda explicita as razões pelas quais determinada solução é juridicamente justificável. Um registro completo de operações não transforma uma decisão deficiente em decisão motivada. De modo inverso, o dever de fundamentar não exige

expor toda operação mental do julgador. A função jurídica da rastreabilidade é auxiliar a verificação das premissas efetivamente utilizadas, sem confundir documentação técnica com justificação do provimento.

Essa distinção evita exigir transparência ilimitada de todo fluxo de trabalho. Informações protegidas por sigilo, dados pessoais de terceiros e registros sem pertinência para a controvérsia demandam tratamento próprio. O parâmetro defendido é funcional: quando a intervenção automatizada tiver relevância para uma premissa decisória contestada, deve existir um caminho institucional para esclarecer sua base documental e corrigir o erro. O alcance do acesso depende de fundamentação, proporcionalidade e proteção dos direitos envolvidos.

O risco de perda de habilidades, frequentemente denominado deskilling, deve ser situado nesse contexto. A delegação continuada de leitura, pesquisa e comparação pode reduzir oportunidades de exercício dessas competências. Contudo, demonstrar perda efetiva entre profissionais do Judiciário exigiria investigação longitudinal, com medidas de desempenho e controle de fatores concorrentes. Neste artigo, o fenômeno é tratado como risco a monitorar, não como resultado estabelecido para a magistratura brasileira. Preservar a capacidade de revisão requer formação e prática de verificação, mas também instrumentos que tornem essa prática viável.

4 LIMITES DOS MODELOS E DA REVISÃO HUMANA ISOLADA

4.1 PLAUSIBILIDADE TEXTUAL E VERIFICAÇÃO JURÍDICA

Modelos de linguagem geram sequências textuais a partir de padrões aprendidos e das informações disponíveis em sua operação. A coerência da resposta não assegura que uma afirmação seja verdadeira, que uma norma esteja vigente ou que um precedente seja aplicável. Ji *et al.* (2023) sistematizam diferentes formas de produção de conteúdo não sustentado na geração de linguagem natural. O conceito de alucinação é útil para caracterizar esse problema, desde que não substitua a identificação do erro concreto e de seu efeito jurídico.

No apoio à jurisdição, é necessário separar ao menos três falhas. A primeira é a inexistência da fonte, como uma decisão inventada. A segunda é a infidelidade ao documento, quando se atribui a um julgado real uma proposição que ele não contém. A terceira é a inadequação jurídica da utilização, quando a fonte existe e foi corretamente transcrita, mas não sustenta a conclusão para o caso. A última não se resolve pela confirmação de um número processual ou de um endereço eletrônico: exige análise dos fundamentos, dos fatos relevantes e da relação entre a decisão citada e a controvérsia.

Kalai *et al.* (2025) relacionam a persistência de alucinações a incentivos de treinamento e avaliação que favorecem respostas por tentativa em detrimento do reconhecimento da incerteza. Esse

argumento ajuda a explicar por que respostas assertivas não equivalem a conhecimento verificado. Não autoriza, porém, afirmar que todos os sistemas têm o mesmo desempenho ou que técnicas de recuperação, ferramentas externas e procedimentos de avaliação são inúteis. A confiabilidade deve ser aferida por tarefa, versão, ambiente e critérios de validação.

Purushothama *et al.* (2025) relatam sensibilidade de interpretações jurídicas em inglês a variações de formulação das perguntas e limitações na correspondência com julgamentos humanos. A publicação nos anais consultados consiste em resumo de uma página. Ela sustenta o alerta geral sobre estabilidade, mas não permite reconstruir integralmente o experimento a partir desse documento. A inferência pertinente ao presente estudo é limitada: interfaces e comandos podem constituir variáveis relevantes da avaliação de ferramentas interpretativas, cuja robustez não deve ser presumida.

A divergência entre um modelo e participantes humanos, por si só, também não demonstra erro jurídico. Em questões interpretativas, pode haver desacordo razoável, e a opinião majoritária de uma amostra não constitui critério suficiente de validade. O problema se torna juridicamente relevante quando a variação decorre de aspectos sem pertinência para a controvérsia, não recebe justificação ou impede o controle das premissas. Avaliações técnicas precisam, assim, ser conectadas a padrões jurídicos explícitos, em vez de reduzir a correção da decisão à concordância estatística.

4.2 CONFIANÇA EM AUTOMAÇÃO E CUSTO DE CONFERÊNCIA

Parasuraman e Riley (1997) distinguem formas de uso, uso inadequado, não utilização e imposição indevida de automação. Lee e See (2004) mostram que a confiança deve ser ajustada às capacidades efetivas do sistema e às características do contexto. Esses referenciais afastam a ideia de que mais confiança é sempre desejável: tanto a dependência excessiva quanto a rejeição injustificada podem prejudicar o trabalho. No ambiente jurídico, a forma persuasiva de uma resposta é um sinal insuficiente de confiabilidade.

A revisão de uma minuta pode exigir localizar documentos, comparar passagens, verificar vigência normativa e reconstruir a pertinência de precedentes. Há, portanto, diferença entre o tempo de ler um texto e o tempo de conferir suas premissas. Quando a organização contabiliza somente a rapidez da produção, o trabalho de verificação pode ficar invisível. O risco institucional consiste em atribuir ao revisor a responsabilidade por falhas sem lhe assegurar condições para detectá-las.

Endsley (2017) examina dificuldades da supervisão de sistemas automatizados, inclusive as relacionadas à consciência situacional. A transferência desse debate para o Judiciário requer cautela, pois tarefas, formação e consequências diferem entre domínios. Ainda assim, o referencial torna plausível uma pergunta de pesquisa: a revisão de sínteses prontas permite ao profissional manter

compreensão suficiente dos autos para perceber omissões? A resposta depende de investigação contextual; não pode ser deduzida apenas da existência de uma assinatura humana.

O mecanismo proposto é condicional. Uma saída fluente pode induzir confiança; se a conferência for onerosa e o fluxo premiar a rapidez, aumenta a possibilidade de aceitação sem exame suficiente. Essa possibilidade não elimina a agência do profissional nem transforma toda falha em efeito inevitável da organização. Responsabilidade individual e responsabilidade institucional podem coexistir. A governança é necessária justamente para reduzir condições previsíveis de erro sem converter a tecnologia em justificativa automática para condutas inadequadas.

Também é preciso considerar a hipótese contrária: instrumentos bem desenhados podem tornar a revisão melhor e menos custosa do que seria sem assistência. Jain *et al.* (2025), em preprint sobre verificação factual, identificam benefícios de formas de complementaridade entre avaliações humanas e de IA e diferenças conforme o tipo de ajuda oferecida. A apresentação de evidências e resultados de busca aparece como alternativa mais promissora do que a exposição de avaliações prontas para estimular confiança adequada. Esses achados não validam diretamente um modelo judicial, mas impedem uma rejeição indiferenciada de toda assistência automatizada.

Consequentemente, a tese de insuficiência dirige-se à supervisão isolada, cuja efetividade é presumida sem demonstração. Uma política adequada deve comparar o fluxo assistido com alternativas reais, incluindo o trabalho sem IA, e observar erros remanescentes após a revisão. O objetivo não é alcançar infalibilidade, mas justificar a escolha institucional com base em riscos, benefícios e capacidade de correção.

4.3 REFERÊNCIAS FALSAS E OS LIMITES DA ILUSTRAÇÃO JURISPRUDENCIAL

Em Mata v. Avianca, processo nº 22-cv-1461 (PKC), a decisão de 22 de junho de 2023 examinou a apresentação de decisões inexistentes geradas pelo ChatGPT e a persistência em defendê-las após questionamentos. O tribunal aplicou sanção de US$ 5.000, em responsabilidade conjunta e solidária dos profissionais e do escritório, sob a Rule 11. O fundamento não foi uma proibição geral de IA: a decisão destacou o dever dos advogados de assegurar a correção do material apresentado (ESTADOS UNIDOS, 2023).

Na decisão conjunta Ayinde e Al-Haroun, de 6 de junho de 2025, a High Court inglesa enfrentou referências falsas em episódios associados ao uso efetivo ou suspeito de IA. O julgamento reafirmou a conferência em fontes autorizadas e a responsabilidade de quem utiliza pesquisa produzida por terceiros. Também tratou da adequação da formação, supervisão e regulação profissionais, ampliando o foco para medidas organizacionais (INGLATERRA E PAÍS DE GALES, 2025, §§ 3 e 7–9).

Os dois documentos mostram a relevância prática do problema da verificação. Ao mesmo tempo, a detecção judicial das referências falsas constitui evidência de que controles podem funcionar. O que os episódios não permitem demonstrar é a frequência dos erros não detectados, o efeito causal de metas de produtividade ou a insuficiência de toda revisão humana. Seu uso adequado é ilustrar como informação não verificada chega ao processo e como a resposta pode combinar responsabilização e prevenção.

No Brasil, a avaliação deve ser feita à luz do regime próprio. Os deveres de boa-fé e de veracidade e as hipóteses dos arts. 79 a 81 do CPC exigem análise da conduta e de seus pressupostos, com contraditório e decisão fundamentada. A existência de uma citação errada não autoriza transpor automaticamente sanções estrangeiras ou qualificar toda falha como litigância de má-fé. A responsabilidade processual da parte também não se confunde com a responsabilidade disciplinar ou civil do advogado.

Se o próprio órgão judicial incorpora a referência falsa à fundamentação, a questão se desloca para a validade e a correção do ato, conforme o vício e a via processual cabível. Esse deslocamento reforça a necessidade de prevenção institucional, sem confundir a conduta de quem apresenta uma peça com a de quem exerce jurisdição. A fonte comum do risco tecnológico não uniformiza os deveres jurídicos dos diferentes sujeitos.

4.4 USO NÃO GOVERNADO E EXPLICAÇÕES APARENTES

Adota-se a expressão Shadow AI para designar o uso funcional de ferramentas de IA à margem dos procedimentos de informação, autorização ou controle aplicáveis. A definição não abrange automaticamente toda ferramenta externa ou contratada individualmente. A distinção importa porque a irregularidade depende do regime de uso e das condições de proteção, não somente da origem comercial do serviço. A ausência de registros institucionais pode dificultar a identificação de dados enviados, versões utilizadas e resultados incorporados ao trabalho.

A segurança também envolve o conteúdo processado. A OWASP (2025) descreve ataques de prompt injection, nos quais entradas manipulam o comportamento de aplicações com modelos de linguagem. Um documento submetido para análise pode conter comandos destinados a influenciar a síntese ou a seleção de informações. No processo judicial, isso permite formular um cenário de ameaça: tratar instruções inseridas no material das partes como se fossem comandos legítimos do sistema. Trata-se de risco técnico a prevenir, sem presumir sua ocorrência generalizada nos autos brasileiros.

O controle exige separar conteúdo analisado de instruções autorizadas, limitar permissões e validar a saída. Nenhuma dessas medidas, isoladamente, garante proteção completa. A arquitetura deve

reduzir o alcance do dano, permitir interromper a operação e preservar evidências de incidentes. Como proposta de política pública, a segurança precisa alcançar o fluxo completo, incluindo documentos recuperados, integrações e tratamento do resultado.

Outro equívoco consiste em equiparar a explicação produzida pelo modelo à reconstituição fiel de seu funcionamento. Mehta (2025), em preprint de autoria individual, relata subnotificação de influências nas explicações de modelos em tarefas experimentais. O resultado recomenda cautela com justificativas textuais autogeradas. Uma resposta que descreve etapas convincentes não comprova, por esse motivo, que o sistema tenha efetivamente considerado todas as fontes relevantes.

Para o Direito, essa limitação recomenda privilegiar razões verificáveis: documento de origem, passagem pertinente, enquadramento normativo e resposta às objeções relevantes. Não é necessário conhecer cada operação interna do modelo para identificar uma citação falsa; tampouco basta conhecer a sequência textual apresentada pelo sistema para validar uma interpretação. A transparência útil é aquela que permite conferir uma premissa ou exercer uma contestação juridicamente relevante.

5 MARCO NORMATIVO E CONDIÇÕES DE EFETIVIDADE

O parâmetro constitucional decorre dos arts. 5º, LIV e LV, e 93, IX, da Constituição. No processo civil, os arts. 9º e 10 vedam que o exercício do poder decisório dispense a participação exigível, e o art. 489, § 1º, disciplina deficiências da fundamentação. A aplicação desses dispositivos à assistência algorítmica não transforma todo uso de IA em nulidade. Exige identificar se o modo de utilização comprometeu alguma garantia ou produziu um vício juridicamente relevante (BRASIL, 1988; 2015).

A exigência de enfrentar argumentos capazes de infirmar a conclusão e de demonstrar a pertinência dos precedentes é especialmente importante. Um sistema pode apresentar um conjunto de decisões verdadeiras e, ainda assim, não responder ao argumento decisivo da parte. A validade do apoio tecnológico deve, portanto, ser avaliada pela função que desempenha na formação do ato. Quantidade de citações, fluência e aparência de tecnicidade são indicadores inadequados quando dissociados do problema jurídico.

A LGPD acrescenta obrigações relativas ao tratamento de dados. São pertinentes os princípios do art. 6º, o regime do Poder Público no art. 23, o registro de operações previsto no art. 37 e as medidas de segurança do art. 46. A aplicação depende da finalidade, da base legal e do contexto de tratamento; a disponibilidade de um documento em sistema judicial não autoriza indistintamente seu reaproveitamento em serviços externos. O art. 20, por sua vez, refere-se a decisões tomadas unicamente

com base em tratamento automatizado que afetem interesses do titular: não deve ser descrito como garantia irrestrita de revisão humana para qualquer decisão assistida (BRASIL, 2018).

A Resolução CNJ nº 615/2025 não se limita à supervisão: prevê gestão de riscos, medidas de governança e avaliação de impacto para soluções de alto risco. A alteração da Resolução nº 674/2026 incide na composição do comitê do art. 15. O texto admite contratação individual nas condições do art. 19; seu § 6º distingue registro interno devido e menção facultativa na decisão. No anexo, AR4 trata de juízos conclusivos de aplicação do Direito a fatos, enquanto BR4 contempla textos de apoio sob instruções e revisão do magistrado (CNJ, 2025; 2026).

Essa arquitetura impede sustentar que o regulador simplesmente ignorou a governança. A questão crítica é outra: em que condições a supervisão exigida consegue operar como diferença substantiva entre assistência e substituição decisória? Se o magistrado recebe uma solução pronta e apenas a confirma, a denominação formal de apoio não basta para esclarecer a função exercida. A classificação deve considerar o uso concreto, e não apenas o nome comercial da aplicação ou a existência de uma etapa final de aprovação.

O contraste entre juízo conclusivo e texto de apoio oferece, assim, um ponto específico de discussão. A hipótese deste artigo é que a documentação do fluxo e da intervenção humana pode ajudar a avaliar essa fronteira. Não se propõe substituir o anexo normativo por uma classificação genérica de risco intermediário. Propõe-se investigar se as condições que justificam o enquadramento permanecem presentes durante o uso e se mudanças de finalidade exigem nova avaliação.

Também não se deve confundir registro interno com acesso irrestrito das partes a todos os comandos ou com obrigação geral de declarar IA no texto do pronunciamento. A proposta aqui defendida é de aperfeiçoamento: usos materialmente relevantes para premissas controvertidas devem contar com procedimento de esclarecimento e contestação, respeitado o sigilo. Seu fundamento argumentativo está na efetividade das garantias processuais; a extensão concreta desse procedimento demanda disciplina institucional e não pode ser apresentada como reprodução literal da resolução.

Como referência comparativa, o Regulamento (UE) 2024/1689 inclui determinados usos na administração da justiça no anexo III, item 8, e trata da supervisão no art. 14. A comparação deve preservar a distinção entre sistemas abrangidos e atividades auxiliares, bem como entre o regime europeu e o brasileiro. Não se atribui ao regulamento aplicação automática aos tribunais nacionais nem se realiza aqui análise de seu cronograma de implementação (UNIÃO EUROPEIA, 2024).

A Recomendação da UNESCO de 2021 oferece um parâmetro ético internacional, enquanto o AI RMF do NIST organiza a gestão de riscos nas funções Govern, Map, Measure e Manage. Nenhum desses instrumentos, por si, cria o mesmo tipo de obrigação de uma lei brasileira. Sua utilidade está

em orientar perguntas sobre responsabilidade, avaliação e monitoramento (UNESCO, 2021; TABASSI, 2023). A recepção crítica deve evitar tanto a importação automática quanto a utilização apenas retórica de princípios.

6 SUPERVISÃO AMPLIADA COMO PROPOSTA DE POLÍTICA PÚBLICA

6.1 UM CRITÉRIO DE SUFICIÊNCIA INSTITUCIONAL

Denomina-se supervisão ampliada o arranjo em que a revisão individual é apoiada por competências, informações, recursos e procedimentos de responsabilização. O conceito preserva a decisão humana e distribui deveres de suporte entre gestão do tribunal, equipes técnicas, unidades judiciais e estruturas de controle. A ampliação não autoriza delegar o mérito da causa a um segundo modelo: acrescenta condições para que o responsável possa avaliar o apoio recebido.

Propõem-se quatro requisitos cumulativos como critério analítico de suficiência institucional. O primeiro é informacional: acesso tempestivo às fontes e identificação de limitações relevantes. O segundo é cognitivo: formação adequada à tarefa e possibilidade de perceber erros e omissões. O terceiro é organizacional: tempo e distribuição de trabalho compatíveis com a conferência. O quarto é decisório: autoridade efetiva para rejeitar o resultado, interromper o uso e solicitar correção. A ausência de qualquer requisito compromete a justificativa de confiar na revisão como salvaguarda.

Esse critério não é um teste empírico validado nem pretende esgotar as condições de qualidade jurisdicional. Sua utilidade é tornar a política verificável. Em vez de registrar apenas que uma pessoa aprovou o texto, a avaliação pode perguntar se ela tinha acesso aos autos, se conferiu os pontos relevantes e se o desenho permitia discordar. A responsabilidade do signatário permanece, mas a organização passa a responder também pela adequação das condições de trabalho e do sistema oferecido.

6.2 FONTES VERIFICÁVEIS E INFRAESTRUTURA DE DADOS

A geração aumentada por recuperação, conhecida pela sigla RAG, combina recuperação de documentos com geração de respostas. Lewis *et al.* (2020) apresentam uma arquitetura dessa natureza para tarefas intensivas em conhecimento. No domínio judicial, a técnica pode apoiar a conexão entre afirmações e fontes identificáveis. A proposta não é que a recuperação elimine alucinações, mas que facilite verificações que seriam mais difíceis diante de uma resposta sem lastro documental explícito.

Para cumprir essa função, a interface deve permitir abrir a fonte original e localizar a passagem utilizada. O repositório precisa distinguir ementa e inteiro teor, manter metadados de origem e atualização e preservar relações entre decisões quando relevantes. Uma referência existente não deve

receber aparência de validação jurídica automática: autenticidade do documento, fidelidade da síntese e pertinência ao caso são verificações diferentes.

Há limites que não desaparecem com RAG. A busca pode deixar de recuperar um documento decisivo, privilegiar precedentes de determinada orientação ou trazer material superado. A geração pode extrapolar os trechos recuperados. Por isso, propõe-se registrar quais fontes sustentam as proposições centrais e sinalizar ausência, conflito ou insuficiência de suporte. Quando a base não permite resposta segura, o fluxo deve oferecer encaminhamento para pesquisa adicional, sem incentivar o preenchimento fictício de lacunas.

Um data lake institucional pode integrar essa infraestrutura, mas o volume de dados não assegura governança ou soberania. Armazenamento interno também não elimina dependências de fornecedores nem resolve problemas de qualidade. A escolha entre repositórios centralizados, federados ou outras arquiteturas deve considerar custos, interoperabilidade, atualização, segurança e restrições de acesso. O requisito jurídico relevante é a capacidade de preservar e conferir a origem da informação, e não a adoção obrigatória de uma tecnologia de armazenamento.

Devem ser definidos prazos de retenção, perfis de acesso e critérios de minimização dos registros. Guardar indiscriminadamente todos os comandos e documentos pode ampliar exposição de dados e custos sem melhorar o controle. Como proposta, o registro deve preservar informação suficiente para apurar o uso relevante e seus efeitos, com acesso proporcional à finalidade. A rastreabilidade não deve criar um novo repositório de dados sensíveis desnecessários.

6.3 ATENÇÃO INSTITUCIONAL E REFORMULAÇÃO DO BÔNUS DINÂMICO

A atenção necessária à supervisão é um recurso organizacional. A proposta de bônus dinâmico pode ser preservada se receber definição precisa: reserva variável de tempo e capacidade de revisão para tarefas com maior risco jurídico ou maior dificuldade de conferência. Não se presume aqui bonificação pecuniária, alteração remuneratória ou recompensa vinculada ao resultado do julgamento. Eventuais incentivos financeiros exigiriam fundamento jurídico e avaliação própria, alheios ao recorte desenvolvido.

O mecanismo sugerido funciona como redistribuição de carga. Um caso com fontes conflitantes, prova complexa ou referência não localizada recebe mais tempo de análise e, quando adequado, apoio adicional. Critérios de acionamento devem ser públicos no plano institucional, documentados e sujeitos a revisão. A declaração de confiança do próprio modelo não pode ser o único gatilho: uma resposta altamente confiante pode estar errada, e a dificuldade jurídica nem sempre coincide com a incerteza estatística.

A arquitetura de escolha discutida por Thaler e Sunstein (2008) ajuda a examinar como a apresentação das opções influencia comportamentos. É necessário, entretanto, distinguir um nudge de uma obrigação ou incentivo material. Exibir a fonte antes do botão de aprovação pode alterar o percurso de revisão; exigir dupla conferência é uma regra procedimental; ampliar tempo disponível é uma medida de gestão. Chamar todas essas intervenções de nudge enfraqueceria a precisão conceitual da proposta.

O desenho deve evitar efeitos perversos. Contar divergências em relação à IA pode estimular discordância performática; premiar concordância pode reforçar submissão; exigir justificativa extensa para toda recusa pode desincentivar o controle. Recomenda-se avaliar a qualidade da conferência por amostragem e por critérios de pertinência, sem transformar cada passo em burocracia cujo custo ultrapasse seu benefício. O bônus dinâmico permanece uma hipótese de intervenção institucional a ser testada.

6.4 IMPLEMENTAÇÃO E AVALIAÇÃO

Uma política implementável precisa definir quem decide sobre o uso, quem acompanha o desempenho e quem responde a incidentes. À administração do tribunal cabe assegurar recursos, delimitar usos e exigir condições contratuais de controle. Às equipes técnicas cabe documentar versões, avaliar segurança e oferecer mecanismos de recuperação de fontes. Às unidades judiciais cabe conferir a adequação do apoio ao caso e relatar problemas. Às estruturas de governança cabe examinar resultados agregados e propor correções, preservada a independência no julgamento.

A implementação pode começar por um piloto de escopo delimitado, com tarefas, documentos e critérios de interrupção definidos previamente. Propõe-se avaliar, em ambiente controlado, três condições: trabalho sem assistência, assistência acompanhada de revisão convencional e assistência com revisão apoiada por fontes e tempo reservado. Quando a aleatorização não for viável, a comparação deve registrar diferenças de complexidade, experiência e carga de trabalho. Esses procedimentos constituem uma agenda futura; não foram executados nesta pesquisa.

Os indicadores precisam relacionar produtividade e qualidade. Entre os candidatos estão a proporção de referências inexistentes, a proporção de afirmações sem suporte documental, a frequência de omissões de argumentos relevantes e o tempo total de trabalho, incluindo conferência e retrabalho. A avaliação deve ainda observar correções posteriores e capacidade de localizar a origem dos erros. Para cada medida, é preciso explicitar unidade de análise, denominador, gravidade e método de aferição.

Uma taxa de rejeição das respostas, isoladamente, não é medida de sucesso: pode refletir erros do modelo, cautela do usuário ou inadequação do caso ao sistema. Da mesma forma, elevada aceitação não demonstra qualidade. Propõe-se examinar amostras com avaliadores qualificados, registrar desacordos e distinguir erro verificável de divergência interpretativa razoável. A concordância entre revisores também precisa ser contextualizada, sem convertê-la em substituto automático da argumentação jurídica.

O piloto deve prever critérios para suspensão ou restrição do uso quando houver incidentes graves, perda de rastreabilidade ou desempenho inferior ao padrão previamente definido. A atualização de modelos e bases exige nova avaliação proporcional à mudança. Os resultados podem fundamentar expansão, revisão ou abandono da ferramenta; uma política de inovação responsável deve admitir as três possibilidades.

Há uma objeção importante: controles adicionais podem consumir parte do ganho de eficiência. A resposta não é negar esse custo, mas incluí-lo na avaliação. Uma ferramenta que acelera a redação e aumenta desproporcionalmente o tempo de conferência pode ser inadequada para determinada tarefa. Em outros contextos, a recuperação bem organizada de fontes pode reduzir o custo global. A política deve comparar esses resultados, evitando confundir texto produzido com trabalho concluído de forma juridicamente adequada.

Uma segunda objeção refere-se ao risco de padronização do próprio controle. Listas de conferência excessivamente rígidas podem criar nova ritualização. Por isso, o modelo proposto combina requisitos mínimos com juízo profissional e revisão dos protocolos. Não pretende transformar toda decisão em uma sequência mecânica de validações, mas assegurar que o apoio tecnológico não torne invisíveis erros, fontes e alternativas relevantes.

7 CONCLUSÃO

A supervisão humana isolada é insuficiente como garantia abstrata de controle da inteligência artificial generativa no apoio à decisão judicial. Essa conclusão não decorre de uma prova de incapacidade universal dos revisores, mas da distinção entre presença formal e possibilidade efetiva de intervenção. A assinatura de um responsável não demonstra, por si, acesso às fontes, competência para detectar falhas, tempo de conferência ou liberdade organizacional para rejeitar o resultado.

A articulação entre contraditório, fundamentação e literatura sociotécnica permite situar o risco na produção das razões da decisão. Uma ferramenta pode influenciar o julgamento antes da redação final, ao selecionar e resumir informações. A governança precisa alcançar essas etapas, sem equiparar

registros técnicos à motivação jurídica ou pressupor que todo uso de automação produz um vício processual.

O marco do CNJ já reúne mecanismos de governança que ultrapassam a revisão individual. A crítica mais consistente dirige-se às condições de efetividade e às fronteiras entre apoio e substituição decisória. O artigo propõe que essas condições sejam examinadas mediante requisitos informacionais, cognitivos, organizacionais e decisórios, acompanhados de avaliação contextual. O modelo de supervisão ampliada combina fontes verificáveis, infraestrutura de dados, reserva de tempo para conferência e capacidade de correção institucional.

As técnicas de recuperação e o bônus dinâmico de atenção são meios possíveis dessa política, não garantias de sucesso. Sua adequação depende de testes, custos, proteção de dados e efeitos sobre o trabalho. As decisões estrangeiras e os estudos examinados sustentam a relevância dos mecanismos discutidos, mas não permitem estimar sua incidência no Judiciário brasileiro nem provar a superioridade da intervenção proposta.

A agenda subsequente exige investigar fluxos reais, comparar erros antes e depois da revisão e medir o custo total da assistência. Até que esses resultados estejam disponíveis, a contribuição é normativa e metodológica: oferecer critérios pelos quais a promessa de supervisão possa ser submetida a escrutínio. A integração de IA à jurisdição deve ser avaliada pela preservação da possibilidade de reconstruir, contestar e revisar razões, e não apenas pela velocidade com que produz textos.

REFERÊNCIAS

BRASIL. Constituição da República Federativa do Brasil de 1988. Brasília, DF: Presidência da República, 1988. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 22 set. 2026.

BRASIL. Lei nº 13.105, de 16 de março de 2015. Código de Processo Civil. Brasília, DF: Presidência da República, 2015. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13105.htm. Acesso em: 22 set. 2026.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais. Brasília, DF: Presidência da República, 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 22 set. 2026.

CONSELHO NACIONAL DE JUSTIÇA (CNJ). Resolução nº 615, de 11 de março de 2025. Estabelece diretrizes para o desenvolvimento, utilização e governança de soluções desenvolvidas com recursos de inteligência artificial no Poder Judiciário. Texto compilado com a Resolução nº 674/2026. Brasília, DF: CNJ, 2025. Disponível em: https://atos.cnj.jus.br/files/compilado1243232026032669c529ebd7bbb.pdf. Acesso em: 22 set. 2026.

CONSELHO NACIONAL DE JUSTIÇA (CNJ). Resolução nº 674, de 25 de março de 2026. Altera o art. 15 da Resolução CNJ nº 615/2025. Brasília, DF: CNJ, 2026. Disponível em: https://atos.cnj.jus.br/files/original1224352026032669c525831ce70.pdf. Acesso em: 22 set. 2026.

ENDSLEY, Mica R. From Here to Autonomy: Lessons Learned From Human–Automation Research. Human Factors, v. 59, n. 1, p. 5–27, 2017. DOI: https://doi.org/10.1177/0018720816681350.

ESTADOS UNIDOS. United States District Court for the Southern District of New York. Mata v. Avianca, Inc. Processo nº 22-cv-1461 (PKC). Opinion and Order on Sanctions, documento 54, 22 jun. 2023. Juiz P. Kevin Castel. Disponível em: https://www.nhd.uscourts.gov/sites/default/files/pdf/Mata-v-Avianca-sanctions-order.PDF. Acesso em: 22 set. 2026.

FAZZALARI, Elio. Instituições de direito processual. Tradução de Elaine Nassif. Campinas: Bookseller, 2006.

INGLATERRA E PAÍS DE GALES. High Court of Justice. Ayinde v. London Borough of Haringey; Al-Haroun v. Qatar National Bank QPSC. [2025] EWHC 1383 (Admin). Processos AC-2024-LON-003062 e CL-2024-000435, 6 jun. 2025. Disponível em: https://www.judiciary.uk/wp-content/uploads/2025/06/Ayinde-v-London-Borough-of-Haringey-and-Al-Haroun-v-Qatar-National-Bank.pdf. Acesso em: 22 set. 2026.

JAIN, Rishub; BRIDGERS, Sophie; JANZER, Lili; GREIG, Rory; TEH, Tian Huey; MIKULIK, Vladimir. Human-AI Complementarity: A Goal for Amplified Oversight. arXiv, 2025. Preprint, versão 1, 30 out. 2025. Disponível em: https://arxiv.org/abs/2510.26518v1. Acesso em: 22 set. 2026.

JI, Ziwei et al. Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, v. 55, n. 12, art. 248, 2023. DOI: https://doi.org/10.1145/3571730.

KALAI, Adam Tauman; NACHUM, Ofir; VEMPALA, Santosh S.; ZHANG, Edwin. Why Language Models Hallucinate. arXiv, 2025. Preprint. Disponível em: https://arxiv.org/abs/2509.04664. Acesso em: 22 set. 2026.

LEE, John D.; SEE, Katrina A. Trust in Automation: Designing for Appropriate Reliance. Human Factors, v. 46, n. 1, p. 50–80, 2004. DOI: https://doi.org/10.1518/hfes.46.1.50_30392.

LEWIS, Patrick et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, v. 33, 2020. Disponível em: https://papers.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html. Acesso em: 22 set. 2026.

MEHTA, Deep Pankajbhai. Can We Trust AI Explanations? Evidence of Systematic Underreporting in Chain-of-Thought Reasoning. arXiv, 2025. Preprint, submetido em 25 dez. 2025, identificador 2601.00830. Disponível em: https://arxiv.org/abs/2601.00830. Acesso em: 22 set. 2026.

OPEN WORLDWIDE APPLICATION SECURITY PROJECT (OWASP). LLM01:2025 Prompt Injection. OWASP Gen AI Security Project, 2025. Disponível em: https://genai.owasp.org/llmrisk/llm01-prompt-injection/. Acesso em: 22 set. 2026.

PARASURAMAN, Raja; RILEY, Victor. Humans and Automation: Use, Misuse, Disuse, Abuse. Human Factors, v. 39, n. 2, p. 230–253, 1997. DOI: https://doi.org/10.1518/001872097778543886.

PURUSHOTHAMA, Abhishek; MIN, Junghyun; WALDON, Brandon; SCHNEIDER, Nathan. Not ready for the bench: LLM legal interpretation is unstable and uncalibrated to human judgments. In: PROCEEDINGS OF THE NATURAL LEGAL LANGUAGE PROCESSING WORKSHOP 2025. Suzhou: Association for Computational Linguistics, 2025. p. 317. Resumo publicado nos anais. DOI: https://doi.org/10.18653/v1/2025.nllp-1.22. Disponível em: https://aclanthology.org/2025.nllp-1.22/. Acesso em: 22 set. 2026.

TABASSI, Elham. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg: National Institute of Standards and Technology, 2023. NIST AI 100-1. Disponível em: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf. Acesso em: 22 set. 2026.

THALER, Richard H.; SUNSTEIN, Cass R. Nudge: Improving Decisions About Health, Wealth, and Happiness. New Haven: Yale University Press, 2008.

UNIÃO EUROPEIA. Regulamento (UE) 2024/1689 do Parlamento Europeu e do Conselho, de 13 de junho de 2024. Cria regras harmonizadas em matéria de inteligência artificial. Jornal Oficial da União Europeia, 2024. Disponível em: https://eur-lex.europa.eu/eli/reg/2024/1689/oj. Acesso em: 22 set. 2026.

UNESCO. Recommendation on the Ethics of Artificial Intelligence. Paris: UNESCO, 2021. Disponível em: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics. Acesso em: 22 set. 2026.