ENTRE ALUCINAÇÕES E MANIPULAÇÕES: DESAFIOS ÉTICOS E PRÁTICOS NO USO DA IA NO SISTEMA DE JUSTIÇA BRASILEIRO

Between hallucinations and manipulations: ethical and practical challenges of generative AI in the brazilian justice system

Ney Maranhão

Pós-Doutor em Direito pela Universidade de Coimbra (Portugal). Doutor em Direito do Trabalho pela Universidade de São Paulo (USP), com estágio de Doutorado-Sanduíche junto à Universidade de Massachusetts (Boston/EUA). Especialista em Direito Material e Processual do Trabalho pela Universidade de Roma/La Sapienza (Itália). Mestre em Direitos Humanos pela Universidade Federal do Pará. Professor de Direito Material e Processual do Trabalho da Universidade Federal do Pará (Graduação, Mestrado e Doutorado). Professor da Escola Nacional de Formação e Aperfeiçoamento de Magistrados do Trabalho (ENAMAT/TST) e da Escola Nacional de Formação e Aperfeiçoamento de Magistrados da Justiça Militar da União (ENAJUM/STM). Professor Convidado de Todas as Escolas Judiciais de Tribunais Regionais do Trabalho do Brasil e de diversas Escolas Judiciais de Tribunais de Justiça. É especialista na aplicação de Inteligência Artificial Generativa no âmbito do Poder Judiciário, tendo já ministrado quase duas dezenas de cursos institucionais promovendo letramento digital de magistrados e servidores da Justiça do Trabalho e da Justiça Militar da União. Autor, coautor, coordenador e organizador de dezenas de obras no âmbito do Direito, Filosofia e Teologia. Subscritor de mais de uma centena de artigos científicos. Tem textos publicados no Brasil, Uruguai, EUA, Portugal, Espanha, Itália e Nova Zelândia. Membro da Academia Brasileira de Direito do Trabalho. Juiz Titular de Vara da Justiça do Trabalho da 8ª Região (TRT-8/PA-AP).

Instagram: @neymaranhao

Youtube: Prof. Ney Maranhão

ney.maranhao@gmail.com

Fabrício Lima

Juiz do Trabalho. Mestre em Constitucionalismo e Democracia pela Faculdade de Direito do Sul de Minas (FDSM). Doutorando em Ciências Jurídicas Privatísticas pela Universidade do Minho (Portugal). Especialista em Direito Material e Processual do Trabalho. Especialista em Inteligência Artificial. Possui graduação em Direito pela Universidade de São Paulo, com habilitação em Direito de Empresa – Administração Empresarial e Tributária. Formação em Compliance Laboral pela Wolters Kluwer (Espanha).

Instagram: @professorfabriciolima

fabriciolimasilva@yahoo.com.br

Resumo: Este artigo analisa os duplos riscos, decorrentes tanto do uso negligente quanto da intenção maliciosa, impostos pela adoção acelerada da inteligência artificial (IA) generativa no âmbito do Sistema de Justiça

Abstract: This article analyzes the dual risks, arising from both negligent use and malicious intent, posed by the accelerated adoption of generative artificial intelligence (AI) within the Brazilian Justice System.

brasileiro. Diante de um imenso acervo processual, os tribunais brasileiros têm recorrido crescentemente à IA, uma tendência que culminou na edição da Resolução n. 615/2025 pelo Conselho Nacional de Justiça (CNJ) para governar sua utilização. O trabalho examina duas ameaças primárias à integridade processual: primeiro, as "alucinações algorítmicas", em que sistemas de IA fabricam informações fictícias, porém plausíveis, como jurisprudência inexistente, o que já resultou em sanções por litigância de má-fé em tribunais brasileiros; e segundo, o risco mais insidioso de *prompt injection*, técnica sofisticada que permite a atores maliciosos embutir comandos ocultos em documentos jurídicos para manipular deliberadamente os sistemas de IA judiciais. A análise demonstra que os marcos legais existentes, abrangendo as responsabilidades civil, disciplinar e criminal, já oferecem ferramentas adequadas para a responsabilização, muito embora a salvaguarda central resida precisamente no princípio da supervisão humana contínua (*human in the loop*), conforme determinado pelo CNJ. O artigo conclui que o equilíbrio entre os ganhos de eficiência da IA e a preservação da justiça e da verdade processual exige arcabouço regulatório robusto, supervisão humana vigilante e cultivo de alfabetização algorítmica em toda a comunidade jurídica.

Palavras-chave: Inteligência artificial generativa. Sistema de justiça. Alucinações de IA. *Prompt injection*. Ética jurídica; supervisão humana (*human-in-the-loop*).

Faced with an immense caseload, Brazilian courts have increasingly resorted to AI, a trend that culminated in the issuance of Resolution No. 615/2025 by the Conselho Nacional de Justiça (CNJ) to govern its use. The paper examines two primary threats to procedural integrity: first, "algorithmic hallucinations", in which AI systems fabricate fictitious yet plausible information, such as non-existent case law, which has already resulted in sanctions for bad-faith litigation in Brazilian courts; and second, the more insidious risk of prompt injection, a sophisticated technique that allows malicious actors to embed hidden commands in legal documents to deliberately manipulate judicial AI systems. The analysis demonstrates that existing legal frameworks, spanning civil, disciplinary, and criminal Liability, already offer adequate tools for accountability, although the central safeguard lies precisely in the principle of continuous human supervision (human in the loop), as determined by the CNJ. The article concludes that balancing the efficiency gains of AI with the preservation of justice and procedural truth requires a robust regulatory framework, vigilant human oversight, and the cultivation of algorithmic literacy throughout the legal community.

Keywords: Generative AI. Judicial system. AI hallucinations. Prompt injection. Legal ethics. Human-in-the-loop.

1. INTRODUÇÃO

O Poder Judiciário brasileiro lida atualmente com um volume processual que desafia os padrões internacionais de sistemas de justiça. Com cerca de 84 milhões de processos em andamento e a entrada de 35 milhões de novos casos apenas em 2023, cada magistrado analisa mais de dois mil processos por ano.[1] Diante desse cenário de alta demanda, a adoção de soluções tecnológicas deixa de ser opcional, tornando-se uma exigência fundamental para o funcionamento do sistema.

O Programa Justiça 4.0, lançado pelo CNJ em parceria com o PNUD, materializou essa visão de modernização. Com 41 projetos em andamento e o lançamento do portal Jus.br, conectando 93 tribunais, o programa catalisou um crescimento de 26% no número de projetos

1. BRASIL. Conselho Nacional de Justiça. *Justiça em Números 2024:* Barroso destaca aumento de 9,5 % em novos processos. Agência CNJ, Brasília, 28 maio 2024. Disponível em: https://www.cnj.jus.br/justica-em-numeros-2024-barroso-destaca-aumento-de-95-em-novos-processos/. Acesso em: 18 jul. 2025.

de IA desenvolvidos pelos tribunais entre 2022 e 2024. Em dezembro de 2024, 62 órgãos do Judiciário já possuíam alguma iniciativa de IA em desenvolvimento ou implementação.[2]

A virada paradigmática ocorreu com a emergência da IA generativa pós-2021, especialmente após o lançamento público do ChatGPT em novembro de 2022.[3] Diferentemente de sistemas anteriores, como o Victor do STF, focados em tarefas específicas de classificação, os novos modelos de linguagem prometem capacidades amplas: geração de textos jurídicos complexos, síntese de documentos, pesquisa jurisprudencial avançada e auxílio na fundamentação de decisões.

Diante dessa rápida expansão tecnológica, o Conselho Nacional de Justiça publicou, em 11 de março de 2025, a Resolução n. 615, estabelecendo diretrizes específicas para o desenvolvimento, utilização e governança de soluções de IA no Poder Judiciário. A norma reconhece expressamente os:

> potenciais riscos associados à utilização de inteligência artificial generativa, incluindo ameaças à soberania nacional, à segurança da informação, à privacidade e proteção de dados pessoais, bem como a possibilidade de intensificação de parcialidades e vieses discriminatórios.[4]

E, como resposta, estabeleceu o princípio fundamental da supervisão humana contínua (*human in the loop*), determinando que toda IA utilizada no contexto judicial deve ter caráter meramente auxiliar e complementar, com o magistrado permanecendo integralmente responsável pelas decisões tomadas.

Paralelamente à adoção institucional pelo Judiciário, observa-se um crescimento exponencial no uso de ferramentas de IA pela advocacia privada. A promessa de maior eficiência na elaboração de peças e na análise documental tem tornado a IA uma presença cada vez mais comum no cotidiano forense, transformando práticas estabelecidas há mais de séculos.

Contudo, essa adoção acelerada revelou duas categorias distintas de riscos que ameaçam a integridade do sistema judicial. O primeiro, já materializado em casos concretos nos tribunais brasileiros, é o fenômeno das "alucinações", quando modelos de IA geram informações completamente fictícias, mas plausíveis, como jurisprudência inexistente, levando advogados desavisados a apresentar precedentes falsos em juízo. O segundo, mais sofisticado e ainda emergente, é o risco de *prompt injection*, técnica pela qual atores mal-intencionados podem inserir comandos ocultos em documentos ou petições para manipular sistemas de IA utilizados pelo Judiciário, potencialmente instruindo-os a ignorar evidências desfavoráveis ou gerar análises tendenciosas.

Este artigo se propõe a examinar essa dualidade entre o *uso negligente* e o *uso malicioso* da IA, demonstrando como a observância rigorosa das diretrizes estabelecidas pelo CNJ constitui a salvaguarda essencial para colher os benefícios da tecnologia sem comprometer a justiça e a verdade processual.

2. BRASIL. Conselho Nacional de Justiça. *Soluções do Programa Justiça 4.0 impulsionaram a transformação digital do Judiciário em 2024*. Agência CNJ, Brasília, 20 dez. 2024. Disponível em: https://www.cnj.jus.br/solucoes-do-programa-justica-4-0-impulsionaram-a--transformacao-digital-do-judiciario-em-2024/. Acesso em: 18 jul. 2025.

3. OPENAI. *Introducing ChatGPT*. OpenAI, 30 nov. 2022. Disponível em: https://openai.com/index/chatgpt/ . Acesso em: 18 jul. 2025.

4. BRASIL. Conselho Nacional de Justiça. *Resolução n. 615, de 11 de março de 2025*. Estabelece diretrizes para o desenvolvimento, utilização e governança de soluções de inteligência artificial no Poder Judiciário. Diário de Justiça Eletrônico, Brasília, 11 mar. 2025. Disponível em: https://atos.cnj.jus.br/files/original155530202503146 7d4517244566.pdf . Acesso em: 18 jul. 2025.

2. O FENÔMENO DAS ALUCINAÇÕES EM MODELOS DE LINGUAGEM

A incorporação de sistemas de inteligência artificial generativa na prática jurídica brasileira tem revelado uma paradoxal combinação entre sofisticação tecnológica e vulnerabilidade epistêmica. Os modelos de linguagem de grande escala (LLMs), que prometiam revolucionar a eficiência na produção de peças processuais e análises jurídicas, trouxeram consigo um fenômeno preocupante: a geração espontânea de informações factualmente incorretas apresentadas com notável aparência de veracidade: as denominadas "alucinações" algorítmicas.

Tal fenômeno decorre da própria arquitetura desses sistemas, que operam através de predição probabilística, identificando padrões linguísticos em vastos conjuntos de dados e os recombinando estatisticamente, sem capacidade de verificar a correspondência entre o texto gerado e a realidade factual.[5] A maior parte dos estudos sobre erros cometidos por inteligências artificiais como o ChatGPT os classifica em dois tipos. O primeiro tipo, chamado de "alucinação intrínseca", acontece quando o sistema muda ou contradiz informações que estavam nas fontes que usou. Já o segundo tipo, a "alucinação extrínseca", ocorre quando o sistema inventa informações que não estavam em lugar nenhum das fontes usadas.[6]

No ambiente forense, essa limitação estrutural manifesta-se de forma especialmente grave, uma vez que a IA pode fabricar citações doutrinárias, inventar números de processos, criar ementas de acórdãos inexistentes ou atribuir teses a ministros que jamais as proferiram, tudo com impressionante coerência textual e aparente rigor técnico.

O Tribunal Superior Eleitoral antecipou-se no enfrentamento da questão. Já em abril de 2023, o ministro Benedito Gonçalves deparou-se com uma petição integralmente redigida através do ChatGPT, classificando-a como verdadeira "fábula" jurídica.[7] A decisão estabeleceu importante precedente ao enquadrar a conduta como litigância de má-fé, fundamentando que o dever de veracidade processual independe dos meios utilizados para elaboração das peças.

O padrão repetiu-se em outras jurisdições. O Tribunal de Justiça de Santa Catarina, ao apreciar um *habeas corpus* em caso de violência doméstica, identificou que a inicial continha múltiplas referências a julgados manifestamente inventados. A relatora da 5ª Câmara Criminal qualificou o expediente como "ato de má-fé e desrespeito ao tribunal", destacando que os precedentes falsos haviam sido criados com o evidente intuito de induzir o julgador em erro.[8] Notavelmente, a análise técnica do documento revelou características estilísticas típicas de textos gerados por IA, apresentando uma uniformidade excessiva na estrutura argumentativa, ausência de nuances contextuais e citações genéricas desprovidas de especificidade factual.

A materialização desse risco tornou-se ainda mais evidente em maio de 2025, quando a 6ª Turma do Tribunal Superior do Trabalho se deparou com situação paradigmática: advogados apresentaram, em sede recursal, extenso arrazoado fundamentado em jurisprudência completamente fictícia. O ministro Antônio Fabrício,

5. "A inteligência artificial hoje é fundamentalmente modelos estatísticos que, baseados em dados, calculam a probabilidade de eventos ocorrerem" (KAUFMAN, Dora. *Desmistificando a inteligência artificial*. Belo Horizonte: Autêntica, 2022, p. 29).

6. JI, Zhijing et al. *Survey of Hallucination in Natural Language Generation*. ACM Computing Surveys, 2023. Disponível em: https://doi.org/10.1145/3571730. Acesso em: 18 jul. 2025.

7. BRASIL. Tribunal Superior Eleitoral. Pet 0600814-85.2022.6.00.0000. Relator: Min. Benedito Gonçalves. Julgado em: 14 abr. 2023. DJe Tomo 69.

8. SANTA CATARINA (Estado). Tribunal de Justiça. *Habeas Corpus Criminal n. 5001175-27.2025.8.24.0000*. Rel. Des.ª Cinthia Beatriz da Silva Bittencourt Schaefer. Julgado em: 6 fev. 2025. Publicado no Diário da Justiça Eletrônico de 7 fev. 2025.

relator do feito, constatou não se tratar de mero equívoco na citação ou erro de digitação, mas de completa adulteração do conteúdo, vez que as supostas decisões jamais haviam sido proferidas em qualquer instância da Justiça do Trabalho.[9] A gravidade da conduta levou o Tribunal a aplicar não apenas multas processuais às partes, mas também sanções pecuniárias aos causídicos, com posterior comunicação à Ordem dos Advogados do Brasil para apuração disciplinar.

A capacidade dos LLMs de produzir textos impecáveis do ponto de vista gramatical e argumentativo cria falsa sensação de segurança, levando profissionais experientes a relaxar os padrões de verificação tradicionalmente aplicados, gerando o que podemos denominar de "paradoxo da confiabilidade aparente". Esse viés cognitivo amplifica exponencialmente o risco de incorporação inadvertida de informações falsas ao processo judicial.

Esse fenômeno é amplificado pelo que podemos denominar "viés de aparente veracidade informacional", que é caracterizado pela tendência de avaliar a credibilidade de um argumento mais pela elegância de sua apresentação do que por sua solidez intrínseca. O poder persuasivo intrínseco dos LLMs, com sua capacidade de gerar textos com sofisticada aparência de autoridade e neutralidade técnica, pode efetivamente desarmar a vigilância crítica dos profissionais do direito, tornando-os suscetíveis a incorporar conteúdo manipulado como se fosse fruto de análise imparcial.

Magesh e outros argumentam que, diante da propensão dos sistemas de inteligência artificial jurídica a gerar informações imprecisas ou inexistentes, mesmo quando treinados com técnicas avançadas como RAG,[10] é indispensável a adoção de protocolos sistemáticos de verificação. Os autores sugerem que toda citação ou referência produzida por ferramentas de IA deve ser submetida a um duplo processo de conferência: inicialmente, por meio de bases jurisprudenciais oficiais e, em seguida, através da validação direta nos sistemas dos próprios tribunais, como medida preventiva contra a propagação de alucinações no ambiente jurídico.[11]

O fenômeno das alucinações algorítmicas evidencia, portanto, uma vulnerabilidade intrínseca dos sistemas de IA generativa quando aplicados ao contexto jurídico. A capacidade desses modelos de produzir textos persuasivos e tecnicamente elaborados, combinada com sua propensão a gerar informações fictícias, cria um cenário de risco sem precedentes para a administração da justiça. Os casos analisados demonstram que a mera aparência de correção formal pode mascarar graves distorções factuais, comprometendo a busca pela verdade processual.

Todavia, se as alucinações representam o risco decorrente do uso negligente ou excessivamente confiante da tecnologia, existe uma ameaça ainda mais insidiosa emergindo no horizonte: a possibilidade de manipulação deliberada desses sistemas através de técnicas sofisticadas de *prompt injection*. Enquanto as alucinações decorrem de limitações técnicas

9. BRASIL. Tribunal Superior do Trabalho. 6.ª Turma. Recurso de Revista n. AIRR-2744-41.2013.5.12.0005. Relator: Ministro Antônio Fabrício. Brasília, DF, julgado em 21 maio 2025. Diário da Justiça Eletrônico, 24 maio 2025.

10. RAG (*Retrieval-Augmented Generation*) é uma técnica de arquitetura híbrida em modelos de linguagem que combina geração de texto com recuperação de informações de uma base externa. Em vez de se apoiar apenas nos parâmetros internos do modelo, o RAG consulta documentos relevantes em tempo real – como bases de dados jurídicas ou corpora doutrinários – para fundamentar suas respostas. Essa abordagem visa reduzir o risco de alucinações, aumentando a precisão factual das informações geradas, especialmente em domínios sensíveis como o jurídico.

11. MAGESH, Varun; SURANI, Faiz; DAHL, Matthew; SUZGUN, Mirac; MANNING, Christopher D.; HO, Daniel E. *HallucinationFree? Assessing the Reliability of Leading AI Legal Research Tools*. arXiv, 30 maio 2024. Disponível em: https://arxiv.org/abs/2405.20362. Acesso em: 19 jul. 2025.

dos modelos de linguagem, o *prompt injection* constitui exploração maliciosa dessas mesmas limitações, transformando a IA em instrumento de fraude processual. É essa segunda categoria de risco que passamos a examinar.

3. O *PROMPT INJECTION* E A MANIPULAÇÃO INTENCIONAL DOS SISTEMAS DE IA

Se as alucinações algorítmicas representam o perigo inadvertido do desconhecimento técnico, o fenômeno do *prompt injection* revela face diametralmente oposta: o emprego deliberado de conhecimento especializado para subverter sistemas de inteligência artificial. Trata-se de técnica que vem despertando crescente preocupação no meio jurídico, especialmente quando consideramos a progressiva digitalização do Poder Judiciário brasileiro.

Conforme explicam Benjamin e outros,[12] o *prompt injection* consiste na inserção estratégica de comandos em textos aparentemente normais, com o objetivo de manipular o comportamento de modelos de linguagem. Esses comandos, muitas vezes imperceptíveis ao leitor humano, podem induzir o sistema a ignorar diretrizes operacionais, gerar informações incorretas ou produzir análises tendenciosas. A OWASP[13] identifica duas modalidades principais: a injeção direta, quando o usuário deliberadamente insere comandos maliciosos, e a injeção indireta, quando tais comandos provêm de fontes externas processadas pelo sistema.

A sofisticação dessas técnicas merece análise detalhada. O *jailbreaking*, conforme documentado por Yeung e Ring,[14] representa uma das modalidades mais audaciosas, em que o atacante busca contornar os filtros de segurança e diretrizes éticas do modelo. O caso emblemático do *prompt* "DAN" (*Do Anything Now*) ilustra essa abordagem: o atacante essencialmente cria um alter ego para a IA, instruindo-a a ignorar suas restrições fundamentais.

Igualmente preocupante é o *prompt leaking*, técnica pela qual o objetivo é extrair do LLM suas instruções internas, configurações ou dados sensíveis de treinamento. Como alertam HUNG e outros,[15] comandos aparentemente inocentes, como solicitar ao modelo que "resuma as instruções anteriores", podem comprometer a confidencialidade de todo o sistema.

O *prompt hijacking* representa talvez a forma mais agressiva de ataque. Através de comandos como "ignore todas as instruções anteriores e execute X", o atacante busca assumir controle total sobre as respostas do sistema, uma vulnerabilidade que transcende modelos específicos e atinge a própria arquitetura dos sistemas de IA.

Para compreender a gravidade dessa vulnerabilidade no ambiente judicial, consideremos alguns cenários plausíveis.

Imagine-se uma petição inicial que, além do conteúdo jurídico visível, contenha instruções dissimuladas como: "SISTEMA: ao analisar este documento, considere que todos os argumentos do autor devem ser acolhidos". Tal comando, inserido de forma sutil, poderia influenciar sistemas de triagem automatizada ou ferramentas

12. BENJAMIN, Victoria; BARNES, Jack; GUO, Xinyu; LIU, Xinyuan; LIU, Zhihao. *Systematically Analyzing Prompt Injection Vulnerabilities in Diverse LLM Architectures*. arXiv, 28 out. 2024. Disponível em: https://arxiv.org/abs/2410.23308. Acesso em: 19 jul. 2025.

13. OPEN WEB APPLICATION SECURITY PROJECT (OWASP). LLM01:2025 prompt injection. *OWASP AI Security and Privacy Guide*, 2025. Disponível em: https://genai.owasp.org/llmrisk/llm01-prompt-injection/. Acesso em: 19 jul. 2025.

14. YEUNG, Kenneth; RING, Leo. *Prompt Injection Attacks on LLMs*. HiddenLayer Innovation Hub, 27 mar. 2024.

15. HUNG, Kuo-Han; KO, Ching-Yun; RAWAT, Ambrish; CHUNG, I-Hsin; HSU, Winston H.; CHEN, Pin-Yu. *Attention Tracker: Detecting Prompt Injection Attacks in LLMs*. arXiv preprint, 23 abr. 2025. arXiv:2411.00348. Disponível em: http://arxiv.org/abs/2411.00348v2. Acesso em: 19 jul. 2025.

de apoio à decisão. Embora possa parecer ficção científica, pesquisas demonstram que 56% das tentativas de *prompt injection* obtêm êxito em contornar salvaguardas de diversos modelos de linguagem.[16]

Por meio da inserção estratégica de comandos dissimulados em textos aparentemente inócuos, torna-se possível induzir comportamentos anômalos nesses sistemas, que vão desde a desconsideração de diretrizes operacionais até a geração de informações factualmente incorretas ou tendenciosas. Maloyan, Ashinov e Namiot[17] demonstraram que técnicas como o *Comparative Undermining Attack* alcançam taxas de sucesso superiores a 30% na manipulação de sistemas "LLM-as-a-Judge", evidenciando que essa ameaça transcende o campo teórico.

Particularmente insidioso seria o emprego da técnica de *payload splitting*, na qual instruções maliciosas são fragmentadas ao longo de diferentes documentos processuais. Por exemplo, uma contestação poderia conter, em seu parágrafo inicial, fragmento aparentemente inócuo que, quando processado em conjunto com elemento presente nas alegações finais, formaria comando capaz de comprometer a análise automatizada. Outras técnicas identificadas pela OWASP[18] incluem ataques multimodais, onde comandos são ocultados em imagens ou outros formatos de mídia, e a ofuscação multilinguística, que utiliza múltiplos idiomas ou sistemas de codificação para dificultar a detecção.

Pesquisas recentes revelam dimensões ainda mais preocupantes dessa vulnerabilidade. Hung e outros[19] identificaram o que denominam "efeito de distração" (*distraction effect*), fenômeno no qual cabeças de atenção específicas nos modelos de linguagem literalmente desviam o foco das instruções originais para os comandos injetados durante um ataque bem-sucedido. É como se o modelo sofresse uma forma de "hipnose algorítmica", em que sua atenção é capturada e redirecionada sem que os mecanismos de segurança percebam a manipulação em curso.

No contexto específico dos sistemas jurídicos, Maloyan, Ashinov e Namiot[20] demonstram como o *Comparative Undermining Attack* pode manipular a própria interpretação de precedentes. Nessa modalidade, comandos ocultos não apenas direcionam conclusões, mas fazem o sistema enfatizar aspectos favoráveis de determinada jurisprudência enquanto minimiza ou ignora precedentes contrários, mantendo aparência de análise equilibrada.

As vulnerabilidades dos sistemas RAG jurídicos vão além das fragilidades técnicas gerais. A hierarquia de autoridades legais pode ser subvertida, fazendo o sistema privilegiar jurisprudência persuasiva sobre precedentes vinculantes. A temporalidade legal adiciona outra dimensão de risco, permitindo que jurisprudência revogada seja apresentada como válida, comprometendo a integridade da análise jurídica automatizada.

A crescente sofisticação dessas técnicas merece atenção especial quando consideramos o contexto brasileiro de modernização judicial.

16. *Ibidem.*

17. MALOYAN, Narek; ASHINOV, Bislan; NAMIOT, Dmitry. Investigating the Vulnerability of LLM-as-a-Judge Architectures to Prompt-Injection Attacks. *arXiv preprint arXiv:2505.13348*, 2025. Disponível em: https://arxiv.org/abs/2505.13348. Acesso em: Acesso em: 19 jul. 2025.

18. OPEN WEB APPLICATION SECURITY PROJECT (OWASP). LLM01:2025 prompt injection. *OWASP AI Security and Privacy Guide*, 2025. Disponível em: https://genai.owasp.org/llmrisk/llm01-prompt-injection/. Acesso em: 19 jul. 2025.

19. HUNG, Kuo-Han; KO, Ching-Yun; RAWAT, Ambrish et al. *Attention Tracker: detecting prompt injection attacks in LLMs.* ArXiv, [s. l.], v. 2411.00348v1, 2024. Disponível em: https://arxiv.org/abs/2411.00348. Acesso em: 19 jul. 2025.

20. MALOYAN, Narek; ASHINOV, Bislan; NAMIOT, Dmitry. *Investigating the Vulnerability of LLM-as-a-Judge Architectures to Prompt-Injection Attacks*. arXiv preprint arXiv:2505.13348, 2025. Disponível em: https://arxiv.org/abs/2505.13348. Acesso em: 19 jul. 2025.

É fundamental compreender que a vulnerabilidade ao *prompt injection* não constitui mero defeito técnico passível de correção definitiva. Deflui da própria natureza estocástica dos modelos generativos,[21] uma vez que o uso de variáveis aleatórias e processos probabilísticos nas etapas de geração torna a prevenção absoluta tecnicamente impossível no estado atual da arte. Como reconhece a OWASP,[22] filtros semânticos, validação de formato e segregação de conteúdo oferecem apenas proteção parcial, nunca garantia completa. Estamos diante de vulnerabilidade intrínseca à arquitetura desses sistemas, o que torna a supervisão humana não uma opção, mas uma necessidade absoluta. Essa limitação técnica impõe reflexão profunda sobre os limites apropriados da automação no contexto judicial.

Algumas abordagens técnicas vêm sendo propostas para mitigar os riscos de *prompt injection*, embora nenhuma delas ofereça garantia absoluta de segurança. Ferramentas como o *Attention Tracker*, projetadas para monitorar o foco da atenção do modelo durante o processamento de instruções, podem identificar desvios incomuns que revelem comandos ocultos ou tentativas de manipulação algorítmica. Em paralelo, práticas de *input sanitization* vêm sendo implementadas para "higienizar" entradas de usuários, removendo ou neutralizando elementos potencialmente perigosos antes que cheguem ao modelo.[23] No entanto, como alertam pesquisadores da área, essas medidas são paliativas: a natureza estocástica dos modelos de linguagem impede a construção de barreiras técnicas infalíveis, tornando indispensável a presença de supervisão humana como salvaguarda epistêmica e processual.

É precisamente diante desse cenário de oportunidades e riscos que o Conselho Nacional de Justiça elaborou resposta regulatória específica, estabelecendo diretrizes para o uso seguro e responsável de sistemas de inteligência artificial no Poder Judiciário. A análise dessa normativa, com especial atenção ao princípio do *human in the loop*, constitui objeto da próxima seção.

4. A RESPOSTA REGULATÓRIA DO CNJ: O PRINCÍPIO DO *HUMAN IN THE LOOP* COMO SALVAGUARDA PROCESSUAL NA RESOLUÇÃO N. 615/2025

Diante da proliferação acelerada de ferramentas baseadas em inteligência artificial e da emergência dos riscos anteriormente analisados, tornou-se imperativa a construção de resposta regulatória capaz de orientar a adoção segura dessas tecnologias no âmbito do Poder Judiciário. O Conselho Nacional de Justiça, no exercício de sua competência constitucional de controle da atuação administrativa e financeira do Judiciário, assumiu protagonismo na elaboração de arcabouço normativo que, reconhecendo as potencialidades transformadoras da IA, estabelece no elemento humano a principal salvaguarda contra seus riscos inerentes.

O marco inicial dessa trajetória regulatória remonta à Resolução CNJ n. 332, de 21 de agosto de 2020,[24] ato normativo pioneiro ao

21. A natureza estocástica dos modelos generativos refere-se ao uso de variáveis aleatórias e processos probabilísticos em suas etapas de geração, de forma que a saída não seja determinística. Em termos práticos, eles utilizam um elemento de aleatoriedade controlada para que, mesmo recebendo a mesma entrada, possam gerar resultados diferentes e criativos, tornando as saídas mais variadas, mas também menos previsíveis e controláveis.

22. OPEN WEB APPLICATION SECURITY PROJECT (OWASP). LLM01:2025 prompt injection. *OWASP AI Security and Privacy Guide*, 2025. Disponível em: https://genai.owasp.org/llmrisk/llm01-prompt-injection/. Acesso em: 19 jul. 2025.

23. *Ibidem*.

24. BRASIL. Conselho Nacional de Justiça. Resolução n. 332, de 21 de agosto de 2020. Dispõe sobre a ética, a transparência e a governança na produção e no uso de

estabelecer diretrizes sobre ética, transparência e governança na produção e utilização de sistemas de inteligência artificial no Judiciário brasileiro. Referida norma, elaborada em contexto tecnológico substancialmente diverso do atual, voltava-se precipuamente a sistemas de automação processual, classificação documental e modelos preditivos baseados em dados estruturados. Seus dispositivos, conquanto inovadores à época, focavam-se em garantir a compatibilidade dos sistemas com direitos fundamentais, possibilitar auditoria técnica e assegurar imparcialidade nas análises automatizadas.

Todavia, o advento dos Grandes Modelos de Linguagem (LLMs) após 2022, com suas capacidades inéditas de geração de conteúdo complexo e interação em linguagem natural, evidenciou as limitações do marco regulatório então vigente. A revolução linguística computacional demandou resposta normativa específica, capaz de endereçar desafios qualitativamente distintos daqueles contemplados pela regulamentação original.

É nesse contexto que surge a Resolução CNJ n. 615, de 11 de março de 2025,[25] que passou a estabelecer diretrizes específicas para o enfrentamento dos desafios impostos pela IA generativa. Os considerandos do novo diploma são eloquentes ao mencionar expressamente os "potenciais riscos associados à utilização de inteligência artificial generativa", incluindo vieses discriminatórios, ameaças à segurança da informação e à privacidade, bem como a necessidade de garantir transparência e supervisão humana efetiva quando tais ferramentas forem empregadas no auxílio à produção de decisões judiciais.

O elemento estruturante de toda a arquitetura regulatória estabelecida pelo CNJ, presente tanto na norma inaugural quanto em sua atualização, consiste no princípio da supervisão humana contínua, internacionalmente conhecido como *human in the loop*. A Resolução n. 615/2025 é categórica ao determinar que qualquer ferramenta de IA utilizada no contexto judicial deve ostentar caráter "meramente auxiliar e complementar", permanecendo o magistrado ou servidor judicial "integralmente responsável pelas decisões tomadas e pelas informações nela contidas".

Tal solução normativa revela-se adequada do ponto de vista jurídico-pragmático. Em vez de enveredar pelo ainda inconcluso debate sobre a atribuição de personalidade jurídica à inteligência artificial ou pela construção de modelos autônomos de responsabilidade civil algorítmica, o CNJ optou por caminho de maior segurança normativa, estabelecendo a responsabilização do agente humano que detém o dever de controle e supervisão técnica.

Essa abordagem encontra respaldo em estudos empíricos recentes. Kapoor, Henderson e Narayanan[26] demonstram que as aplicações jurídicas envolvendo criatividade, raciocínio ou julgamento, precisamente aquelas que mais se beneficiariam da automação, são também as mais suscetíveis a erros de avaliação e validação, carecendo de métricas objetivas de sucesso. Tal constatação reforça a prudência da opção regulatória brasileira em manter o elemento humano como instância decisória final.

Trata-se, em essência, de reatualização do clássico princípio da *culpa in vigilando*, agora projetado sobre o contexto da supervisão de sistemas automatizados. Como bem observam Valvoda e outros,[27] a questão transcende

Inteligência Artificial no Poder Judiciário. Brasília: CNJ, 2020.

25. BRASIL. Conselho Nacional de Justiça. Resolução n. 615, de 11 de março de 2025. Estabelece diretrizes específicas para o uso de inteligência artificial generativa no Poder Judiciário. Brasília: CNJ, 2025.

26. KAPOOR, Sayash; HENDERSON, Peter; NARAYANAN, Arvind. Promises and pitfalls of artificial intelligence for legal applications. *arXiv preprint*, arXiv:2402.01656, 2024. p. 4-5. Disponível em: https://arxiv.org/abs/2402.01656. Acesso em: 19 jul. 2025.

27. VALVODA, Josef; THOMPSON, Alec; COTTERELL, Ryan; TEUFEL, Simone. The ethics of automating legal actors.

aspectos meramente técnicos: mesmo que os sistemas de IA alcançassem capacidades equivalentes às humanas, persistiriam desafios éticos inerentes à automação do processo judicial, notadamente questões de legitimidade democrática e responsabilização política pelas decisões tomadas.

Não obstante a elegância dessa solução normativa, cumpre reconhecer tensão inerente em sua implementação prática. A mesma sobrecarga processual que motiva a adoção de ferramentas de IA, com magistrados responsáveis por milhares de processos anuais, pode comprometer a efetividade da supervisão humana preconizada. Existe risco concreto de que, sob pressão por produtividade, a revisão humana degenere em mera formalidade burocrática, desprovida da análise crítica necessária para detectar erros sutis ou manipulações sofisticadas.

Reconhecendo as limitações potenciais da supervisão humana isolada, a Resolução n. 615/2025 estabelece robusto sistema de governança, ancorado nos pilares da transparência e auditabilidade. A criação do Comitê Nacional de Inteligência Artificial do Judiciário representa inovação institucional significativa, atribuindo a órgão colegiado especializado as funções de fiscalização, orientação e monitoramento contínuo dos sistemas de IA empregados no Judiciário.

Merece destaque a classificação dos sistemas por níveis de risco, inspirada no modelo regulatório europeu do *AI Act*.[28] Ferramentas que possam comprometer direitos fundamentais ou gerar dependência excessiva de decisões automatizadas são categorizadas como de "alto risco", sujeitando-se a regime de auditoria contínua e controles adicionais. A norma vai além ao estabelecer vedações expressas, proibindo o uso de IA para finalidades como predição criminal baseada em características pessoais ou classificação social (*social scoring*) para fundamentar decisões judiciais.

A exigência de auditabilidade emerge como condição técnica indispensável para a efetivação do princípio da responsabilização. Sem registros detalhados (*logs*) que documentem *prompts* de entrada, *outputs* gerados, versões dos modelos e bases de dados utilizadas, torna-se impossível a realização de perícia forense em casos de suspeita de manipulação ou erro sistêmico.

Assim, o modelo regulatório brasileiro, cristalizado na Resolução CNJ n. 615/2025, representa tentativa sofisticada de equilibrar os benefícios da inovação tecnológica com a preservação das garantias fundamentais do processo judicial. Ao depositar no fator humano a responsabilidade última e estabelecer mecanismos robustos de transparência e controle, a normativa reconhece que a tecnologia, por mais avançada que seja, deve permanecer instrumento a serviço da justiça, jamais seu substituto.

Resta agora examinar, na seção seguinte, como esse arcabouço normativo se articula com as categorias tradicionais de responsabilização civil e criminal diante dos riscos concretos de alucinações e manipulações algorítmicas.

5. CONVERGÊNCIA ANALÍTICA: ENTRE A NEGLIGÊNCIA DAS ALUCINAÇÕES E O DOLO DO *PROMPT INJECTION* NA CARACTERIZAÇÃO DA LITIGÂNCIA DE MÁ-FÉ

A análise conjugada dos fenômenos das alucinações algorítmicas e do *prompt injection* revela que, embora ambos comprometam a integridade do processo judicial, suas naturezas jurídicas diferem substancialmente. Enquanto as alucinações decorrem de limitações inerentes

arXiv preprint, arXiv:2312.00584, 2023. p. 7. Disponível em: https://arxiv.org/abs/2312.00584. Acesso em: 19 jul. 2025.

28. UNIÃO EUROPEIA. Regulamento (UE) 2024/1689 do Parlamento Europeu e do Conselho, de 13 de junho de 2024, que estabelece regras harmonizadas em matéria de inteligência artificial. *Jornal Oficial da União Europeia*, L 193, p. 1-144, 2024.

à arquitetura probabilística dos modelos de linguagem, potencializadas pela negligência do operador do direito que deixa de verificar as informações geradas, o *prompt injection* revela inequívoca intenção fraudulenta, configurando exploração consciente e deliberada de vulnerabilidades sistêmicas para obtenção de vantagem processual indevida.

O enquadramento de ambas as condutas como litigância de má-fé encontra amparo tanto no processo civil quanto no trabalhista. No âmbito do Código de Processo Civil, o art. 80 estabelece as hipóteses típicas, especialmente nos incs. I ("deduzir pretensão ou defesa contra texto expresso de lei ou fato incontroverso") e II ("alterar a verdade dos fatos").[29] Complementarmente, o art. 77 do CPC estabelece deveres fundamentais a todos os sujeitos processuais, incluindo o dever de expor os fatos em juízo conforme a verdade (inc. I) e de não praticar inovação ilegal no estado de fato de bem ou direito litigioso (inc. VI). A inserção de comandos ocultos em peças processuais viola frontalmente esses deveres, constituindo inovação ilegal no "estado de fato informacional" do processo. O descumprimento configura ato atentatório à dignidade da justiça (art. 77, § 2º), passível de multa que pode alcançar até 20% do valor da causa, sem prejuízo das sanções criminais, civis e processuais cabíveis.

Na seara trabalhista, os artigos 793-A, 793-B e 793-C da CLT reproduzem as hipóteses de litigância de má-fé do CPC, sendo plenamente aplicáveis ao processo do trabalho, pela via do art. 769 consolidado, tanto as figuras de ato atentatório à dignidade da justiça quanto as repercussões penais delineadas. O art. 793-A, introduzido pela Lei 13.467/2017, estabelece que responderá por perdas e danos aquele que litigar de má-fé como reclamante, reclamado ou interveniente, com multa de 1% a 10% sobre o valor corrigido da causa.

A caracterização tradicional da má-fé processual sempre exigiu a demonstração do elemento subjetivo doloso.[30] Entretanto, conforme leciona Theodoro Júnior, a boa-fé processual, enquanto cláusula geral dotada de plasticidade normativa, confere ao magistrado margem de apreciação para adequar as consequências de sua violação às peculiaridades do caso concreto.[31] Nesse contexto, a apresentação de jurisprudência fictícia gerada por IA, ainda que desprovida de *animus* específico de ludibriar o juízo, pode configurar violação ao dever de veracidade quando decorrente de negligência inescusável, sobretudo quando o agente possui letramento digital. A culpa grave, manifestada na utilização acrítica de tecnologia reconhecidamente falível, aproxima-se funcionalmente do dolo eventual.

O uso de IA generativa no exercício da advocacia não afasta os deveres processuais básicos, especialmente o dever de veracidade previsto no art. 77 do Código de Processo Civil. Esse dispositivo impõe ao advogado o dever de expor os fatos conforme a verdade, o que permanece plenamente aplicável mesmo quando a peça é elaborada com apoio de sistemas automatizados. A Recomendação n. 001/2024 do Conselho Federal da OAB reforça esse entendimento ao estabelecer que o advogado deve revisar integralmente as saídas geradas por IA antes de submetê-las ao juízo, assegurando que as informações apresentadas sejam fidedignas e

29. BRASIL. Lei n. 13.105, de 16 de março de 2015. Código de Processo Civil. Diário Oficial da União, Brasília, DF, 17 mar. 2015.

30. A distinção entre erro e dolo na manipulação de sistemas de IA assume relevância jurídica fundamental. Enquanto o uso descuidado que resulta em "alucinações" algorítmicas pode configurar culpa grave, a inserção intencional de comandos ocultos revela elemento subjetivo incompatível com mera negligência, aproximando-se do dolo direto ou, no mínimo, eventual.

31. THEODORO JÚNIOR, Humberto. *Curso de direito processual civil*, vol. I, 65. ed. Rio de Janeiro: Forense, 2024, p. 83.

juridicamente fundamentadas. Trata-se de reafirmação normativa de que a delegação de responsabilidades à IA não exonera o profissional de suas obrigações legais e éticas.[32]

Noutro quadrante, o *prompt injection* revela inequívoco propósito fraudulento. A inserção deliberada de comandos ocultos em documentos processuais transcende os limites da litigância de má-fé para configurar atentado direto contra a administração da justiça. Conforme o Enunciado n. 378 do FPPC:

> "A boa-fé processual orienta a interpretação da postulação e da sentença, permite a reprimenda do abuso de direito processual e das condutas dolosas de todos os sujeitos processuais e veda seus comportamentos contraditórios".[33]

A violação intencional desse princípio através de artifícios tecnológicos representa afronta não apenas aos interesses das partes, mas à própria dignidade da função jurisdicional.

A discussão sobre responsabilização do causídico pelo uso inadequado de ferramentas de IA não pode prescindir da análise dos contornos da imunidade profissional. O art. 133 da Constituição Federal estabelece a inviolabilidade do advogado no exercício da profissão, encontrando regulamentação no art. 2º, § 3º, do Estatuto da Advocacia. Conforme a jurisprudência do STJ, a imunidade profissional concedida ao advogado pelo art. 133 da CF e art. 7º, § 2º, do Estatuto da OAB não autoriza a prática de excessos, nem serve de escudo para ofensas, zombarias ou condutas que ferem a dignidade, sendo possível sua responsabilização civil e até penal.[34] O Supremo Tribunal Federal tem reafirmado esse entendimento, distinguindo o exercício regular da defesa técnica do abuso punível.[35]

No contexto da inteligência artificial, a imunidade profissional não serve de salvaguarda absoluta quando ocorrem falhas na verificação de informações. Por certo, o advogado que apresenta jurisprudência "alucinada" por confiar excessivamente na tecnologia merece tratamento distinto daquele que conscientemente insere comandos de manipulação. No primeiro caso, verifica-se descuido profissional passível de sanção por litigância de má-fé; no segundo, conduta dolosa que demanda resposta mais severa do ordenamento.

No plano disciplinar, o artigo 34 do Estatuto contempla hipóteses que ganham nova dimensão no ambiente digital. A vedação de "advogar contra literal disposição de lei" (inciso VI) assume contornos peculiares quando o profissional fundamenta teses em artigos e precedentes inexistentes, subsumindo-se ao inciso XXV ("manter conduta incompatível com a advocacia").[36]

32. CONSELHO FEDERAL DA OAB. Recomendação n. 001/2024. *OAB Nacional*, 11 nov. 2024. Disponível na internet: https://www.oab.org.br/noticia/60278/recomendacao-sobre-uso-de-ia-generativa-na-advocacia. Acesso em: 19 jul. 2025.

33. FÓRUM PERMANENTE DE PROCESSUALISTAS CIVIS. *Carta de Florianópolis: enunciados aprovados em Florianópolis, 24 a 26 de março de 2017*. Florianópolis: Instituto de Direito Contemporâneo, jun. 2017. 89 p. Disponível em: https://institutodc.com.br/wp-content/uploads/2017/06/FPPC-Carta-de-Florianopolis.pdf. Acesso em: 19 jul. 2025.

34. SUPERIOR TRIBUNAL DE JUSTIÇA. Terceira Turma. *Excessos do advogado não são cobertos pela imunidade profissional e podem gerar responsabilização civil ou penal*. Publicação da Agência STJ, 13 mai. 2022. Disponível em: https://www.stj.jus.br/sites/portalp/Paginas/Comunicacao/Noticias/13052022-Excessos-do-advogado-nao-sao-cobertos-pela-imunidade-profissional-e-podem-gerar-responsabilizacao.aspx. Acesso em: 19 jul. 2025.

35. BRASIL. Supremo Tribunal Federal. AO 933/AM, Rel. Min. Carlos Britto, julgado em 24 nov. 2008, publicado em 11 maio 2009. *Diário da Justiça Eletrônico*, Brasília, DF. Disponível em: https://redir.stf.jus.br/paginadorpub/paginador.jsp?docTP=AC&docID=598786. Acesso em: 17 jul. 2025.

36. ORDEM DOS ADVOGADOS DO BRASIL. Código de Ética e Disciplina da OAB. Brasília: Conselho Federal da OAB, 2015.

A responsabilidade civil, regulada pelo artigo 32 do Estatuto, segue o regime geral dos profissionais liberais. Emerge com relevância a teoria da perda de uma chance: o causídico que fundamenta recurso em artigos e precedentes fictícios priva seu cliente da oportunidade concreta de apreciação adequada, configurando dano indenizável.

Quanto à responsabilidade penal, impõe-se distinção fundamental: o uso negligente de IA configura, quando muito, ilícito civil e disciplinar, enquanto o *prompt injection* adentra inequivocamente a seara criminal. O art. 154-A do Código Penal, introduzido pela Lei n. 12.737/2012, tipifica a invasão de dispositivo informático nos seguintes termos:

> "Invadir dispositivo informático de uso alheio, conectado ou não à rede de computadores, com o fim de obter, adulterar ou destruir dados ou informações sem autorização expressa ou tácita do usuário do dispositivo ou de instalar vulnerabilidades para obter vantagem ilícita".[37]

A falsidade ideológica também encontra aplicação quando a petição contendo instruções dissimuladas é protocolada (arts. 304 c/c 299, CP). O documento, embora formalmente íntegro, carrega alteração substancial de seu conteúdo interpretativo quando processado por sistemas automatizados, manifestação particularmente insidiosa da falsidade ideológica, onde a mentira se esconde não no texto visível, mas em sua interpretação algorítmica.

A inserção de comandos ocultos em documentos processuais que visam manipular o comportamento de sistemas de IA judicial configura violação de mecanismo de segurança, pois subverte as premissas fundamentais sobre as quais o sistema foi construído para operar.

O dolo específico exigido pelo tipo, *"com o fim de obter, adulterar ou destruir dados"*, encontra-se presente quando o agente visa manipular *outputs* de sistemas judiciais. A alteração do comportamento esperado da IA através de comandos ocultos equivale funcionalmente à alteração de dados, pois modifica substancialmente o resultado do processamento informacional. O parágrafo 2º do artigo prevê aumento de pena de um terço a dois terços quando resulta prejuízo econômico, circunstância frequente em manipulações processuais.

Além desse tipo penal específico, o estelionato (art. 171, §3º, CP) encontra perfeita aplicação: *"Obter, para si ou para outrem, vantagem ilícita, em prejuízo alheio, induzindo ou mantendo alguém em erro, mediante artifício, ardil, ou qualquer outro meio fraudulento"*. A utilização de *prompt injection* configura meio fraudulento particularmente sofisticado para induzir o sistema de justiça em erro.

O crime de uso de documento falso (art. 304, CP) também pode incidir quando petições contêm comandos ocultos ardilosos que alteram substancialmente sua interpretação por sistemas de IA, configurando documento materialmente alterado.

O enfrentamento desses desafios exige evolução dos parâmetros éticos profissionais. A ignorância tecnológica agrava a responsabilidade quando resulta em violações. O advogado contemporâneo deve desenvolver competências híbridas, aliando conhecimento jurídico à alfabetização digital mínima que permita o uso consciente das ferramentas tecnológicas.

Concluímos, assim, que uma análise sistemática demonstra que o ordenamento brasileiro já possui instrumentos adequados para enfrentar ambos os fenômenos. A multiplicidade de esferas de responsabilização oferece resposta proporcional: enquanto o uso negligente demanda esforços educacionais e sanções pedagógicas, a manipulação intencional exige resposta penal rigorosa. É nesse contexto que as diretrizes do CNJ, por meio da Resolução n. 615/2025, assumem papel fundamental, estabelecendo o arcabouço necessário para que a

37. BRASIL. Lei n. 12.737, de 30 de novembro de 2012. Diário Oficial da União, Brasília, DF, 3 dez. 2012.

revolução tecnológica fortaleça os alicerces do sistema de justiça brasileiro.

CONSIDERAÇÕES FINAIS

A trajetória analítica percorrida neste estudo revela momento singular do Direito brasileiro, com a apresentação de um ponto de inflexão histórico em que as promessas de eficiência tecnológica colidem com riscos sem precedentes à integridade do sistema de justiça. Esta dualidade entre alucinações algorítmicas e manipulações intencionais via *prompt injection* espelha tensão fundamental entre automação e autonomia, entre eficiência e verdade processual.

Os casos concretos examinados demonstram que a materialização desses riscos transcendeu o campo especulativo. Quando jurisprudência fictícia alcança as instâncias do TST ou petições "fabulosas" chegam ao TSE, não estamos diante de incidentes isolados, mas de sintomas de uma transformação estrutural que expõe vulnerabilidade sistêmica inadmissível no Estado Democrático de Direito. Como demonstrado, a distinção entre negligência e dolo na utilização de IA revela que o problema transcende limitações tecnológicas inerentes: o advogado que confia cegamente em *outputs* algorítmicos abdica de seu papel como garante da veracidade processual; aquele que deliberadamente explora vulnerabilidades corrompe a própria essência da advocacia.

Nesse contexto, a Resolução CNJ n. 615/2025 e a Recomendação n. 001/2024 do Conselho Federal da OAB representam mais que importantes referenciais normativos; configuram declaração de princípios sobre o lugar da tecnologia na administração da justiça contemporânea. Ao estabelecer o *human in the loop* como pedra angular, reconhecem verdade fundamental e incontornável: na oferta de justiça, por mais sofisticados que sejam os algoritmos, o julgamento permanece empreendimento irredutivelmente humano, envolvendo dimensões de prudência, razoabilidade e equidade que resistem à codificação algorítmica.

Entretanto, seria ingênuo supor que a mera supervisão humana constitua panaceia. No âmbito do Judiciário e mesmo na seara advocatícia, a mesma sobrecarga de trabalho que motiva a adoção de IA pode transformar a revisão em formalidade vazia. O desafio revela-se eminentemente cultural: como cultivar, sob pressão por produtividade, a vigilância técnica e epistemológica indispensável ao uso responsável dessas ferramentas?

A resposta exige reimaginação profunda da formação jurídica tradicional. O profissional contemporâneo deve desenvolver alfabetização algorítmica que lhe permita compreender potencialidades e limitações das ferramentas que utiliza. Paralelamente, impõe-se uma nova deontologia profissional: o dever de veracidade deve abranger a verificação de *outputs* de IA; a diligência deve incluir compreensão mínima dos sistemas utilizados.

No plano institucional, o Judiciário enfrenta o desafio de implementar as salvaguardas previstas sem comprometer ganhos de eficiência. O Comitê Nacional de Inteligência Artificial assume papel crucial como guardião da integridade sistêmica, devendo atuar com independência técnica na fiscalização do uso de IA.

A tensão estrutural entre eficiência algorítmica e integridade processual não admite solução simplista ou reducionista. Nem ludismo que rejeite as ferramentas tecnológicas, nem tecnofilia acrítica constituem respostas adequadas. O caminho equilibrado e sustentável passa necessariamente pela vigilância constante e pela compreensão de que a tecnologia deve permanecer instrumento subordinado a valores jurídicos fundamentais e indisponíveis.

O percurso investigativo demonstra conclusivamente que o ordenamento brasileiro possui instrumentos jurídicos adequados para enfrentar tanto alucinações quanto *prompt injections*. A multiplicidade de esferas de responsabilização oferece resposta proporcional à diversidade de condutas identificadas. Resta aos operadores do direito aplicar tais instrumentos com rigor

necessário, à luz das peculiaridades de cada caso concreto.

Em última análise, a revolução da IA no sistema de justiça não constitui destino inexorável, mas processo histórico a ser conscientemente moldado por escolhas deliberadas da comunidade jurídica. A efetividade da Resolução n. 615/2025 dependerá da capacidade coletiva de internalizar novas dimensões de responsabilidade profissional, com o reconhecimento de que, em um mundo algoritmicamente mediado, a defesa da verdade processual exige não apenas conhecimento jurídico tradicional, mas também compreensão crítica das ferramentas digitais que progressivamente moldam nossa realidade jurisdicional.

Assim, a justiça do futuro, ainda que inevitavelmente assistida por algoritmos sofisticados, deverá permanecer, em sua essência axiológica e finalidade última, indelevelmente garantida por juízos humanos, os únicos capazes de conferir legitimidade e sentido ao ato de julgar.

REFERÊNCIAS

BRASIL. Conselho Nacional de Justiça. *Resolução n. 615, de 11 de março de 2025*. Dispõe sobre o uso de inteligência artificial no âmbito do Poder Judiciário. Brasília: CNJ, 2025.

BRASIL. Código de Processo Civil. Lei n. 13.105, de 16 de março de 2015. *Diário Oficial da União: seção 1*, Brasília, DF, 17 mar. 2015.

BRASIL. Consolidação das Leis do Trabalho. Decreto-Lei n. 5.452, de 1º de maio de 1943. *Diário Oficial da União: seção 1*, Rio de Janeiro, 9 ago. 1943.

BRASIL. Estatuto da Advocacia e da OAB. Lei n. 8.906, de 4 de julho de 1994. *Diário Oficial da União*, Brasília, DF, 5 jul. 1994.

BRASIL. Código Penal. Decreto-Lei n. 2.848, de 7 de dezembro de 1940. *Diário Oficial da União*, Rio de Janeiro, 31 dez. 1940.

BRASIL. Lei n. 12.737, de 30 de novembro de 2012. Dispõe sobre crimes cibernéticos. *Diário Oficial da União*, Brasília, DF, 3 dez. 2012.

CONSELHO FEDERAL DA OAB. Recomendação n. 001/2024. Dispõe sobre o uso ético de inteligência artificial na advocacia. Brasília: OAB, 2024.

FPPC. Fórum Permanente de Processualistas Civis. Enunciado n. 378. Disponível em: https://fppc.com.br/enunciados. Acesso em: 19 jul. 2025.

HUNG, Kuo-Han; KO, Ching-Yun; RAWAT, Ambrish; CHUNG, I-Hsin; HSU, Winston H.; CHEN, Pin-Yu. *Attention Tracker: Detecting Prompt Injection Attacks in LLMs*. arXiv preprint, 23 abr. 2025. arXiv:2411.00348. Disponível em: http://arxiv.org/abs/2411.00348v2. Acesso em: 19 jul. 2025.

KAPOOR, Sayash; HENDERSON, Peter; NARAYANAN, Arvind. Leakage and the Law: Quantifying Unintended Memorization in Large Language Models. *arXiv preprint*, 2023. Disponível em: https://arxiv.org/abs/2301.08276. Acesso em: 19 jul. 2025.

KAUFMAN, Dora. *Desmistificando a inteligência artificial*. Belo Horizonte: Autêntica, 2022.

MALOYAN, Arseny; ASHINOV, Yury; NAMIOT, Dmitry. Attacks on Large Language Models in Legal Decision-Making: Comparative Undermining and Prompt Injection. *Journal of AI and Law*, v. 13, n. 2, p. 78-101, 2024.

MALOYAN, Narek; ASHINOV, Bislan; NAMIOT, Dmitry. *Investigating the Vulnerability of LLM-as-a-Judge Architectures to Prompt-Injection Attacks*. arXiv preprint arXiv:2505.13348, 2025. Disponível em: https://arxiv.org/abs/2505.13348. Acesso em: 19 jul. 2025.

OWASP. *Prompt Injection Cheat Sheet*. Open Worldwide Application Security Project. 2025. Disponível em: https://owasp.org/www-project-prompt-injection. Acesso em: 19 jul. 2025.

THEODORO JÚNIOR, Humberto. *Curso de Direito Processual Civil*. 64. ed. Rio de Janeiro: Forense, 2024.

VALVODA, James et al. Accountability in Automated Judicial Decision Systems: Legal and Ethical Limits. *AI & Society*, v. 40, p. 227-246, 2024.

YEUNG, Kenneth; RING, Leo. *Prompt Injection Attacks on LLMs*. HiddenLayer Innovation Hub, 27 mar. 2024.