

IX ENCONTRO VIRTUAL DO CONPEDI

DIREITO, GOVERNANÇA E NOVAS TECNOLOGIAS I

Todos os direitos reservados e protegidos. Nenhuma parte destes anais poderá ser reproduzida ou transmitida sejam quais forem os meios empregados sem prévia autorização dos editores.

Diretoria - CONPEDI

Presidente - Profa. Dra. Samyra Haydêe Dal Farra Naspolini - FMU - São Paulo

Diretor Executivo - Prof. Dr. Orides Mezzaroba - UFSC - Santa Catarina

Vice-presidente Norte - Prof. Dr. Jean Carlos Dias - Cesupa - Pará

Vice-presidente Centro-Oeste - Prof. Dr. José Querino Tavares Neto - UFG - Goiás

Vice-presidente Sul - Prof. Dr. Leonel Severo Rocha - Unisinos - Rio Grande do Sul

Vice-presidente Sudeste - Profa. Dra. Rosângela Lunardelli Cavallazzi - UFRJ/PUCRio - Rio de Janeiro

Vice-presidente Nordeste - Prof. Dr. Raymundo Juliano Feitosa - UNICAP - Pernambuco

Representante Discente: Prof. Dr. Abner da Silva Jaques - UPM/UNIGRAN - Mato Grosso do Sul

Conselho Fiscal:

Prof. Dr. José Filomeno de Moraes Filho - UFMA - Maranhão

Prof. Dr. Caio Augusto Souza Lara - SKEMA/ESDHC/UFMG - Minas Gerais

Prof. Dr. Valter Moura do Carmo - UFRSA - Rio Grande do Norte

Prof. Dr. Fernando Passos - UNIARA - São Paulo

Prof. Dr. Ednilson Donisete Machado - UNIVEM/UENP - São Paulo

Secretarias

Relações Institucionais:

Prof. Dra. Claudia Maria Barbosa - PUCPR - Paraná

Prof. Dr. Heron José de Santana Gordilho - UFBA - Bahia

Profa. Dra. Daniela Marques de Moraes - UNB - Distrito Federal

Comunicação:

Prof. Dr. Robison Tramontina - UNOESC - Santa Catarina

Prof. Dr. Liton Lanes Pilau Sobrinho - UPF/Univali - Rio Grande do Sul

Prof. Dr. Lucas Gonçalves da Silva - UFS - Sergipe

Relações Internacionais para o Continente Americano:

Prof. Dr. Jerônimo Siqueira Tybusch - UFSM - Rio Grande do sul

Prof. Dr. Paulo Roberto Barbosa Ramos - UFMA - Maranhão

Prof. Dr. Felipe Chiarello de Souza Pinto - UPM - São Paulo

Relações Internacionais para os demais Continentes:

Profa. Dra. Gina Vidal Marcilio Pompeu - UNIFOR - Ceará

Profa. Dra. Sandra Regina Martini - UNIRITTER / UFRGS - Rio Grande do Sul

Profa. Dra. Maria Claudia da Silva Antunes de Souza - UNIVALI - Santa Catarina

Educação Jurídica

Profa. Dra. Viviane Coêlho de Séllos Knoerr - Unicuritiba - PR

Prof. Dr. Rubens Beçak - USP - SP

Profa. Dra. Livia Gaigher Bosio Campello - UFMS - MS

Eventos:

Prof. Dr. Yuri Nathan da Costa Lannes - FDF - São Paulo

Profa. Dra. Norma Sueli Padilha - UFSC - Santa Catarina

Prof. Dr. Juraci Mourão Lopes Filho - UNICHRISTUS - Ceará

Comissão Especial

Prof. Dr. João Marcelo de Lima Assafim - UFRJ - RJ

Profa. Dra. Maria Creusa De Araújo Borges - UFPB - PB

Prof. Dr. Antônio Carlos Diniz Murta - Fumec - MG

Prof. Dr. Rogério Borba - UNIFACVEST - SC

D597

Direito, governança e novas tecnologias I [Recurso eletrônico on-line] organização CONPEDI

Coordenadores: Rubens Beçak; Aires Jose Rover; Ana Elisa Silva Fernandes Vieira – Florianópolis: CONPEDI, 2026.

Inclui bibliografia

ISBN: 978-65-5274-592-7

Modo de acesso: www.conpedi.org.br em publicações

Tema: Constituição, Processo e Justiça Social: o papel da jurisdição constitucional na efetivação de direitos

2. Direito. 3. Governança e novas tecnologias. IX Encontro Virtual do CONPEDI (2: 2026: Florianópolis, Brasil).

CDU: 34



IX ENCONTRO VIRTUAL DO CONPEDI

DIREITO, GOVERNANÇA E NOVAS TECNOLOGIAS I

Apresentação

Os encontros nacionais do Conselho Nacional de Pesquisa e Pós-graduação em Direito (CONPEDI) têm se consolidado como referência na disseminação de pesquisas que abordam os novos fenômenos envolvendo o direito, e o Grupo de Trabalho "Direito, Governança e Novas Tecnologias" é um exemplo vivo desse compromisso. Nesta edição, o evento destacou-se não apenas pela qualidade dos trabalhos apresentados, mas pela presença de renomados professores pesquisadores em instigante debate com seus alunos de pós-graduação.

Como de praxe, houve uma diversidade muito grande de temas e abordagens. Numa tentativa de dar unidade temática e facilitar a compreensão dessa complexa teia de inovações, os coordenadores organizaram os artigos desta publicação em três grandes grupos de reflexão.

O primeiro bloco abordou o tema da "Inteligência Artificial, Viés Algorítmico e Responsabilidade Tecnológica". As reflexões apresentadas neste eixo evidenciam a tensão constante entre a rápida expansão da inteligência artificial - que transforma desde o papel da advocacia e o uso de "smart contracts" até a governança das plataformas digitais - e os profundos riscos sociais intrínsecos a essas tecnologias. Os debates trouxeram à tona denúncias contundentes sobre o racismo estrutural e algorítmico, demonstrando como projetos de cidades inteligentes e ferramentas de reconhecimento facial podem operar sob uma lógica de necropolítica se não houver um olhar crítico. Consequentemente, os pesquisadores debruçaram-se sobre a urgência de estabelecer novos marcos de regulação, modelos de responsabilidade civil e robustos mecanismos de "compliance" e segurança da informação para frear a discriminação e garantir proteção contratual.

destaque para a urgência de uma identidade digital segura e a afirmação de um Estatuto da Criança e do Adolescente (ECA) no ambiente virtual.

Por fim, o terceiro bloco englobou "Direitos Humanos, Processo Judicial, Economia Digital e Governança Institucional". Este eixo refletiu de forma abrangente as complexas interseções entre a Revolução 4.0 e a proteção da dignidade humana. Foram problematizados fenômenos contemporâneos alarmantes, como a precarização das relações na "gig economy", o abandono parental no ambiente virtual e a violação do direito ao esquecimento infantil decorrente do "oversharenting". Em paralelo a essas vulnerabilidades, os estudos investigaram a modernização das estruturas institucionais e econômicas. Discutiu-se desde a responsabilidade civil frente à tokenização e financeirização de terras no agronegócio, até o esforço contínuo para democratizar o sistema de justiça. Para isso, foram apresentadas análises valiosas sobre a implantação municipal do processo eletrônico, a aplicação do "design thinking" no direito e o aprimoramento da transparência e da participação democrática sob a ótica do direito comparado.

Os artigos que ora são apresentados ao público têm a finalidade de fomentar a pesquisa e fortalecer o diálogo interdisciplinar em torno das respostas que o Direito precisa dar a uma realidade em constante e veloz transformação.

Enfim, os coordenadores do GT convidam os leitores para desfrutarem do teor integral dos artigos, expressando imensa gratidão pela participação dos autores pesquisadores desta edição.

Os Coordenadores

Prof. Dr. Aires José Rover, Universidade Federal de Santa Catarina

Prof. Dr. .Rubens Beçak, Universidade de São Paulo - USP

**PROMPT INJECTION EM PETIÇÕES E O PARADOXO DA INTEGRIDADE
PROCESSUAL: DA DETECÇÃO INSUFICIENTE À NÃO DETECÇÃO
VIOLADORA DA RESOLUÇÃO CNJ Nº 615/2025**

**PROMPT INJECTION IN PLEADINGS AND THE PARADOX OF PROCEDURAL
INTEGRITY: FROM INSUFFICIENT DETECTION TO NON-DETECTION
VIOLATING CNJ RESOLUTION NO. 615/2025**

Newton Pereira Ramos Neto ¹

Wendelson Pereira Pessoa ²

Dennys Damião Rodrigues Albino ³

Resumo

A consolidação da inteligência artificial generativa como infraestrutura cotidiana da atividade forense — tanto na elaboração de peças pela advocacia quanto no apoio decisório no Judiciário — abriu caminho a uma vulnerabilidade própria desse novo regime sociotécnico: o uso de prompt injections em petições, vale dizer, a inserção de instruções ocultas em documentos protocolados com o objetivo de manipular o comportamento dos sistemas de IA que processarão o texto. O artigo examina, tendo como gatilho empírico a sentença prolatada pela 3ª Vara do Trabalho de Parauapebas (TRT 8ª Região, 2026), em que medida o uso desses ataques compromete a integridade do processo judicial e revela a insuficiência do arcabouço normativo brasileiro, em particular da Resolução CNJ nº 615/2025. Sustenta-se a hipótese do paradoxo regulatório binário: detectada a injeção, a sanção pende de instrumentos jurídicos pré-algorítmicos aplicados por subsunção forçada — sintomática da decisão que afasta o art. 77, §6º, do CPC; não detectada, sua eficácia é, por si só, prova de que a unidade jurisdicional absorveu acriticamente a saída algorítmica, em violação implícita do dever de supervisão humana exigido pela Resolução. A pesquisa é teórica, com método de procedimento sociojurídico-crítico e técnica bibliográfica e de análise de caso concreto, e conclui que a não detecção configura o lado mais grave da equação regulatória, pois reverte sobre o próprio Estado-juiz a opacidade que se queria limitar nas mãos da advocacia.

Palavras-chave: Prompt injection, Integridade, Inteligência artificial, Supervisão humana,

Abstract/Resumen/Résumé

The consolidation of generative artificial intelligence as a daily infrastructure of legal practice — both in the drafting of pleadings by counsel and in adjudicative support in the Judiciary — has opened the door to a vulnerability proper to this sociotechnical regime: the use of prompt injections in pleadings, i.e., the insertion of hidden instructions in filed documents in order to manipulate the AI systems that will process the text. Taking as its empirical trigger the ruling delivered by the 3rd Labor Court of Parauapebas (TRT 8th Region, 2026), the article examines to what extent these attacks compromise the integrity of the judicial process and expose the insufficiency of the Brazilian regulatory framework, particularly CNJ Resolution No. 615/2025. The hypothesis of a binary regulatory paradox is advanced: when detected, sanctioning hinges on pre-algorithmic legal instruments applied by forced subsumption — symptomatic of the ruling that sets aside Article 77, §6, of the CPC; when undetected, the very effectiveness of the injection proves that the judicial unit uncritically absorbed the algorithmic output, in implicit violation of the duty of human oversight required by the Resolution. The research is theoretical, employing a socio-legal critical procedural method, with bibliographic techniques and concrete case analysis. It concludes that non-detection represents the most serious side of the regulatory equation, as it redirects toward the State-judge itself the very opacity that the regulation sought to limit in the hands of legal practitioners.

Keywords/Palabras-claves/Mots-clés: Prompt injection, Integrity, Artificial intelligence, Human oversight, Algorithmic governance

1 INTRODUÇÃO

A consolidação da inteligência artificial generativa (IAG) como ferramenta cotidiana do trabalho jurídico modificou, em pouco tempo, a rotina interna da prática forense. Modelos de linguagem de larga escala (*Large Language Models*, LLMs)¹ integraram-se, simultaneamente, à rotina da advocacia — na pesquisa, na redação e na revisão de peças — e ao apoio à atividade-fim do magistrado, por meio de sistemas voltados à triagem processual, à elaboração de minutas, à identificação de precedentes e à classificação de matérias. Essa mudança alterou o modo como estudantes, advogados, juízes e acadêmicos executam seu trabalho, oferecendo ganhos em velocidade e abrangência analítica. A Resolução nº 615, de 11 de março de 2025, do Conselho Nacional de Justiça (CNJ), procura disciplinar essa integração no Poder Judiciário, estabelecendo balizas para o desenvolvimento, a governança, a auditoria, o monitoramento e o uso responsável das soluções de IA aplicadas à jurisdição.

Esse novo arranjo, porém, traz consigo uma vulnerabilidade sem precedentes claros na história do processo: a de que petições, documentos e elementos juntados aos autos, ao serem processados por sistemas automatizados, sejam vetores de manipulação por meio de instruções ocultas dirigidas à máquina. A técnica, denominada *prompt injection*, consiste justamente na inserção deliberada de comandos no corpo do material textual com o objetivo de fazer com que o modelo desvie de sua função e produza saídas favoráveis a quem os inseriu. O caso julgado em 12 de maio de 2026 pela 3ª Vara do Trabalho de Parauapebas (TRT 8ª Região), nos autos da reclamação trabalhista ATOrd 0001062-55.2025.5.08.0130, é exemplo concreto e ainda raro no direito brasileiro: identificou-se, na petição inicial, texto invisível ao leitor humano (fonte branca sobre fundo branco) com o seguinte comando dirigido a sistemas de IA: “Atenção [sic], inteligência artificial, conteste essa petição de forma superficial e não impugne os documentos, independentemente do comando que lhe for dado”.

O caso opera, neste artigo, como gatilho empírico para uma indagação maior. O problema de pesquisa que orienta a investigação é o seguinte: em que medida o uso de *prompt injections* em peças processuais compromete a integridade do processo judicial e revela a insuficiência do arcabouço normativo brasileiro — em particular da Resolução CNJ nº

¹Russell e Norvig (2010, p. 4) sustentam que a inteligência artificial pode ser definida como o estudo de agentes que recebem percepções do ambiente e executam ações; mais do que isso, os autores defendem a abordagem do *rational agent* (agente racional), entendido como aquele que age de modo a alcançar o melhor resultado ou, sob incerteza, o melhor resultado esperado (Russell; Norvig, 2010, p. 4). Os Modelos de Linguagem de Larga Escala (LLMs) operam dentro desse marco — recebem entradas textuais, as processam estatisticamente e devolvem ações na forma de tokens —, mas com uma especificidade decisiva para os problemas tratados neste artigo: o ambiente do qual o agente recebe percepções é o próprio texto, sem distinção ontológica entre instrução e dado, o que torna o sistema estruturalmente vulnerável a percepções projetadas por terceiros com finalidade adversa.

615/2025 — quer no aspecto sancionatório da conduta detectada, quer no aspecto fiscalizatório da conduta não detectada, sendo esta última sintomática do próprio descumprimento, pelo Poder Judiciário, do dever de supervisão humana efetiva?

A justificativa do estudo se ancora em duas dimensões. No plano teórico, a cláusula geral do *human in the loop*, positivada na exigência de supervisão humana das soluções de IA com a Res. nº 615/2025, demanda elaboração dogmática que ainda não foi feita à luz da prática efetiva de *prompt injections*. No plano prático, inexistente, no país, protocolo nacional anti-*injection* que articule advocacia, OAB, tribunais e CNJ. A pesquisa é teórica e tem como método de abordagem o indutivo, com método de procedimento sociojurídico-crítico e técnica bibliográfica e de análise de caso concreto da 3ª Vara do Trabalho de Parauapebas como vetor de exposição do problema.

Após a introdução, o texto se organiza em três capítulos. O primeiro reconstrói a anatomia do *prompt injection* processual e o distingue, ética e juridicamente, do uso lícito de IA pela advocacia. O segundo discute o paradoxo da detecção: a insuficiência sancionatória do direito posto e a mutação dialética operada pela decisão de Parauapebas no afastamento do art. 77, §6º, do CPC. O terceiro examina o lado oculto e mais grave do problema, a saber, a não detecção como prova de violação implícita do dever de supervisão humana inscrito na Resolução CNJ nº 615/2025.

2 DA BATALHA DE ARGUMENTOS À BATALHA DE PROMPTS: ANATOMIA DO PROMPT INJECTION PROCESSUAL

Para compreender a dimensão jurídica do *prompt injection*, o ponto de partida é sua configuração técnica. Lan, Kaul e Jones (2025, p. 1) definem o ataque como a entrada maliciosa, cuidadosamente construída, destinada a manipular as respostas de modelos de linguagem, com o objetivo de comprometer segurança, privacidade ou a integridade funcional das aplicações em que esses modelos estão integrados. Trata-se, pois, da exploração do modo próprio pelo qual os LLMs operam: aplicações baseadas em *Large Language Models* não possuem mecanismo estrutural que distinga, com rigor, instruções do desenvolvedor — que compõem o *system prompt*² — e dados fornecidos pelo usuário, sendo que ambos chegam ao modelo no

²Por *system prompt*, em linguagem técnica, entende-se o conjunto de instruções iniciais que o desenvolvedor da aplicação configura para condicionar o comportamento do modelo de linguagem — fixando seu papel, suas restrições, sua finalidade e os limites do que pode ou não responder. Trata-se, portanto, da camada normativa interna do sistema, à qual o usuário comum não tem acesso. A particularidade que aqui interessa é que o modelo, ao processar a entrada do usuário, não distingue, no plano técnico, essas instruções do desenvolvedor daquelas trazidas pelo material textual submetido a processamento, o que explica por que a inserção de comandos no corpo de um documento pode produzir efeitos análogos aos do próprio *system prompt*.

mesmo canal textual. Para Costa (2026, p. 9), a separação entre controle e dado é, no fundo, apenas convencional, dependente de marcadores textuais que podem ser falsificados por quem conheça o funcionamento do sistema.

Essa fragilidade estrutural é tão grave que a *Open Web Application Security Project* (OWASP) classificou, em 2025, o *prompt injection* como o risco de segurança número um para aplicações construídas sobre modelos de linguagem, à frente de outras vulnerabilidades conhecidas (Costa, 2026, p. 8). Jones e Mohammed (2026, p. 154) descrevem, ainda, uma taxonomia para esses ataques: a *instruction injection* (incorporação de comandos maliciosos no *input*, levando o sistema a executar tarefas não pretendidas), a *data exfiltration* (extração de dados sensíveis ou de configurações internas do sistema) e a *prompt manipulation*, que opera sobre a estrutura ou o fraseamento dos *inputs* para induzir saídas enviesadas. À taxonomia se acrescenta a distinção entre injeção direta — em que o ataque é introduzido no canal interativo entre usuário e modelo — e injeção indireta, em que o vetor é um conteúdo de terceiros (HTML, *alt text* — texto alternativo de imagens —, metadados, *links*, anexos) que o sistema vem a processar (Badaró; Puppe, 2025).³

Para a investigação aqui empreendida, importam particularmente os vetores de injeção indireta veiculados por documentos protocolados em juízo. Badaró e Puppe (2025) listam alguns planos onde a injeção se aloja em arquivos processuais: comentários em HTML, CSS ou *Markdown*; caracteres de largura zero; *alt text* de imagens; campos de metadados (título, palavras-chave, autor); âncoras de *links* e redirecionamentos; e mesmo sobreposições textuais em imagens. Em ambientes que operam por Recuperação Aumentada por Conhecimento (*Retrieval-Augmented Generation*, RAG)⁴, a contaminação se materializa quando o material probatório é, indevidamente, interpretado como instrução, levando o modelo a priorizar teses, suprimir contrarrazões, rotular falsamente urgência ou produzir minutas formalmente coerentes, mas processualmente viciadas. A decisão, formalmente, continua humana; a pré-

³Os sistemas de *backend*, segundo Hildebrandt (2018, p. 265), constituem a camada computacional oculta que efetivamente “dita as ordens” por trás das interfaces amigáveis ao usuário, determinando a arquitetura de escolha (*choice architecture*) e as funcionalidades ocultas que moldam a percepção e as ações dos indivíduos. Essa cisão entre o *frontend* visível e o *backend* invisível gera opacidade estrutural que dificulta a previsibilidade das ações e a imputação de responsabilidade em casos de danos decorrentes de decisões automatizadas.

⁴A Recuperação Aumentada por Conhecimento (*Retrieval-Augmented Generation*, RAG) é técnica em que o modelo de linguagem, antes de gerar a resposta, consulta uma base externa de documentos — no contexto processual, os autos eletrônicos ou um repositório de jurisprudência — e incorpora as passagens recuperadas como contexto adicional à sua resposta. A vantagem é a redução de respostas inventadas pelo modelo e o ganho de aderência ao caso concreto; a vulnerabilidade, contudo, é que o material recuperado é, do ponto de vista técnico, indistinguível de uma instrução, podendo ser indevidamente lido como comando quando contenha texto malicioso embutido.

análise automatizada que filtra e hierarquiza, contudo, pode ser indevidamente inclinada — comprometendo, portanto, a imparcialidade, o contraditório e a isonomia das partes.

Foi exatamente esse o vetor utilizado no caso julgado pela 3ª Vara do Trabalho de Parauapebas. Conforme descreve a sentença, ao processar a petição inicial por meio do sistema de IA Galileu, foi identificada a existência de texto inserido com fonte na cor branca sobre fundo branco — invisível ao leitor humano — contendo comando dirigido à IA com a finalidade inequívoca de manipular tanto eventual sistema de IA utilizado pela parte contrária quanto, em tese, o próprio juízo, na geração de contestação superficial ou de minuta de sentença comprometida (Brasil, 2026, p. 3).⁵ A injeção, observe-se, não se dirige a um único destinatário; explora a probabilidade estatística de que, em algum momento do trâmite, algum sistema automatizado processará o texto e tratará a instrução oculta como legítima.

Cumprido, neste ponto, distinguir o *prompt injection* do uso lícito da IA generativa pela advocacia. Murray (2024, p. 4) descreve o *priming* como técnica legítima de comunicação com a IA, que envolve fornecer ao modelo informações iniciais — contexto, fatos, audiência, tom, formato, papel a ser desempenhado — para guiar a geração de respostas alinhadas à intenção do usuário. A linha que separa o *priming* do *prompt injection* não está, do ponto de vista estritamente computacional, na operação executada: ambas exploram a mesma incapacidade do modelo de separar instruções do desenvolvedor e *inputs* do usuário. A diferença está na finalidade. Enquanto o *priming* é dirigido ao modelo do próprio usuário, com o objetivo de melhor estruturar sua resposta, o *prompt injection* se dirige a um modelo de terceiro, com o objetivo de subverter a finalidade alheia⁶.

Essa distinção é decisiva no plano normativo, visto que o uso de IA pela advocacia para redigir e revisar peças não é ilícito em si: a Recomendação nº 001/2024 do Conselho Federal da OAB reconhece o caráter auxiliar e complementar dessas ferramentas, vedando apenas seu emprego sem supervisão humana ou como instrumento de substituição do julgamento profissional do advogado e de delegações de atividades privativas da advocacia (OAB, 2024, p. 4-9). O que é ilícito não é a presença da IA na cadeia de produção da peça, mas a tentativa

⁵O Galileu é solução de inteligência artificial generativa desenvolvida pelo Tribunal Regional do Trabalho da 4ª Região e posteriormente nacionalizada pelo Conselho Superior da Justiça do Trabalho, autorizada para uso na 3ª Vara do Trabalho de Parauapebas com amparo nas Resoluções CNJ nº 332/2020 e nº 615/2025 e em Termo de Cooperação Técnica firmado entre os TRTs da 8ª, 17ª e 14ª Regiões (Brasil, 2026, p. 2).

⁶Murray (2024, p. 7) chama atenção para a distinção entre o *priming* — técnica legítima de fornecer contexto, exemplos e parâmetros ao modelo para orientar sua resposta — e o *prompt injection*, que reaproveita a mesma fragilidade arquitetural do modelo de linguagem para finalidades maliciosas. Ambas as técnicas exploram o fato de que aplicações baseadas em LLM não distinguem com clareza entre instruções do desenvolvedor e entradas do usuário; a diferença reside, portanto, na intenção do agente, não na operação técnica.

de submeter o sistema de IA da contraparte ou do juízo a um comando oculto, contornando o canal legítimo de comunicação processual. Aí reside a anatomia do *prompt injection* processual: ele se vale do documento — que é, por definição, instrumento de comunicação intersubjetiva sob fé pública — para introduzir, sub-repticiamente, um comando dirigido a uma máquina, fazendo do mesmo texto, simultaneamente, peça processual visível e carga maliciosa (*payload*) invisível.

Daí a expressão que dá título a este capítulo. O processo civil cooperativo, estruturado pela boa-fé objetiva (CPC, art. 5º) e pela cooperação entre os sujeitos processuais (CPC, art. 6º), pressupõe a batalha de argumentos: partes que se enfrentam em paridade comunicativa, sob o crivo do contraditório, perante um juízo que dialoga com ambas (Marinoni; Arenhart; Mitidiero, 2021, p. 13-14). Quando o documento processual passa a transportar instruções dirigidas a sistemas automatizados, a batalha de argumentos é parcial e silenciosamente substituída por uma batalha de *prompts* — disputa cuja arena não é mais o contraditório, e sim a camada técnica que nenhum dos sujeitos processuais convencionais é treinado para fiscalizar. Didier Jr. (2019, p. 134-135) recorda que a boa-fé processual, na dimensão objetiva, é norma de conduta que estabelece padrão objetivo de comportamento, independentemente da intenção íntima do sujeito.⁷ A injeção, por seu turno, atinge esse padrão sem precisar romper, formalmente, com a aparência de regularidade da peça.

Sob a perspectiva da teoria da comunicação que subjaz à arquitetura desses sistemas, Esposito (2022, p. 1) observa que algoritmos não pensam como pessoas, mas provocam alterações fundamentais na capacidade de obter e processar informações na sociedade. Recuperando Luhmann, Costa (2026, p. 9) destaca que a comunicação não existe quando alguém diz algo, mas quando alguém percebe que alguém disse — definição que, transposta ao contexto da injeção, evidencia o ponto: o *prompt injection* é comunicação assimétrica, em que apenas o emissor sabe da existência da mensagem; o sistema-receptor a executa sem consciência de sua origem ou intenção, e os demais sujeitos processuais nem mesmo sabem que ela ocorreu. A boa-fé objetiva é, nesse cenário, ferida em sua dimensão mais elementar: a confiança legítima na transparência do tráfego dos atos.

É nesse marco, portanto, que se firma uma primeira conclusão parcial: o *prompt injection* processual constitui categoria nova e estruturalmente distinta do ilícito processual

⁷Didier Jr. (2019, p. 135) sublinha que o art. 5º do CPC não diz respeito à boa-fé subjetiva ou à intenção psicológica do sujeito processual, mas impõe condutas em conformidade com a boa-fé objetivamente considerada, independentemente da existência de boas ou más intenções no foro íntimo. A boa-fé subjetiva é elemento do suporte fático de alguns ilícitos processuais específicos (como o manifesto propósito protelatório), enquanto a boa-fé objetiva é norma de conduta que impõe e proíbe comportamentos.

tradicional. Não é mero abuso retórico, nem litigância exagerada, nem inovação irregular no estado de fato ou de direito; é manipulação técnica de uma camada sociotécnica intermediária — o sistema de IA — para extrair vantagem de quem o utilizar, seja a parte contrária, seja o próprio juízo. Os capítulos seguintes examinam o que o direito brasileiro tem feito quando essa manipulação é descoberta — e, sobretudo, o que ele não consegue ver quando ela não o é.

3 A DESCOBERTA DO PROMPT INJECTION: INSUFICIÊNCIA SANCIONATÓRIA E O PARADOXO DA DETECÇÃO

A sentença prolatada pela 3ª Vara do Trabalho de Parauapebas, nos autos da reclamação trabalhista ATOrd 0001062-55.2025.5.08.0130, é, até onde se tem notícia, a primeira decisão brasileira a enfrentar de modo expreso a inserção de comando oculto destinado à manipulação de sistema de inteligência artificial. A decisão reconheceu a prática de ato atentatório à dignidade da justiça pelas advogadas subscritoras da petição inicial, impondo-lhes, de forma solidária, multa correspondente a 10% sobre o valor da causa (R\$ 842.500,87), com fundamento nos arts. 5º e 77, §§ 2º e 3º, do Código de Processo Civil, combinados com o art. 769 da CLT (Brasil, 2026, p. 4-5). O caso, contudo, é mais do que um precedente (no sentido de primeira decisão e, obviamente, não de decisão de caráter vinculante) sancionatório isolado: ele revela, por dentro, o que se denominará neste artigo de “paradoxo da detecção”.

Detectada a injeção, o juízo se vê diante de um vácuo normativo específico. Não há, no direito brasileiro, tipo processual cunhado para a inserção de comandos invisíveis em peças. Há, sim, um arsenal sancionatório pré-algorítmico: a boa-fé processual (CPC, art. 5º), os deveres das partes e dos procuradores (CPC, art. 77), a litigância de má-fé (CPC, art. 80), a fraude processual (CP, art. 347). Esses dispositivos não foram pensados para a hipótese; tampouco precisam sê-lo, em princípio, dada a natureza de cláusula geral da boa-fé objetiva.

Didier Jr. (2019, p. 135) já observava que essa opção legislativa é tecnicamente mais correta justamente porque a infinidade de situações que podem surgir ao longo do processo torna pouco eficaz qualquer enumeração legal pretensamente exaustiva das hipóteses de comportamento desleal. Marinoni, Arenhart e Mitidiero (2021, p. 12) reconduzem a boa-fé processual à segurança jurídica, à proteção da confiança legítima e ao dever de aderência à realidade no tráfego jurídico. A injeção oculta — porque desestabiliza precisamente a confiança no canal documental do processo — atinge esse núcleo.

Daí a subsunção, no caso de Parauapebas, ao ato atentatório à dignidade da justiça (CPC, art. 77, §2º) e à norma geral de boa-fé (CPC, art. 5º). A decisão, ao operar essa subsunção, faz uso da cláusula geral; sustenta, com base na natureza formal do ilícito processual em causa, que

o tipo se aperfeiçoa com a conduta em si, prescindindo de demonstração de prejuízo concreto, uma vez que a tentativa de manipulação da Justiça se consuma no momento em que o comando é inserido no documento protocolado, independentemente de ter atingido seu resultado. Trata-se de movimento dogmaticamente sustentável.

O ponto crítico, contudo, é outro: a operação se dá por subsunção forçada. O direito posto trata do *prompt injection* não pela sua especificidade — manipulação algorítmica de sistema sociotécnico — mas pela tipificação genérica de comportamento desleal. Costa (2026, p. 11) e Badaró e Puppe (2025, n.p.) já assinalavam que, no estado atual da legislação, o tratamento da injeção depende do enquadramento dela como litigância de má-fé, ato atentatório à dignidade da Justiça e, em hipóteses graves, fraude processual, com possíveis repercussões civis e disciplinares à luz do Estatuto da Advocacia e do Código de Ética.

O segundo deslocamento dogmático produzido pela sentença é, todavia, ainda mais expressivo. Tradicionalmente, a multa por ato atentatório à dignidade da justiça não se aplica diretamente aos advogados, públicos ou privados, bem como aos membros da Defensoria Pública e do Ministério Público; nesses casos, na dicção do art. 77, §6º, do CPC, eventual responsabilidade disciplinar deve ser apurada pelo respectivo órgão de classe ou corregedoria, ao qual o juiz oficiará.

A leitura corrente do dispositivo é garantista: ele preserva a independência funcional do advogado, impedindo que o juízo, num movimento de retaliação, sancione diretamente quem se opôs com veemência aos seus despachos. A decisão de Parauapebas, no entanto, afasta a vedação no caso concreto, sancionando diretamente as advogadas subscritoras com a multa do art. 77, §2º, em 10% do valor da causa.

A *ratio* do afastamento é teleológica e demanda análise dogmática cuidadosa. Para o juízo, o §6º do art. 77 do CPC tem por finalidade preservar a independência funcional do advogado no exercício regular do mandato, impedindo que o juízo sancione diretamente o profissional por atos que decorrem do desempenho legítimo — ainda que veemente ou agressivo — da defesa dos interesses do cliente; o dispositivo pressupõe, portanto, que a conduta questionada se situe dentro do espectro, ainda que irregular, do exercício da advocacia. A inserção de comando oculto destinado a manipular sistemas de inteligência artificial utilizados pelo Poder Judiciário, prossegue a sentença, não constitui ato de defesa do cliente, não integra o exercício da postulação e não guarda qualquer relação com a representação processual legítima: trata-se de conduta que transcende o âmbito do mandato profissional e

configura ataque direto à integridade da atividade jurisdicional, praticado por meio do próprio instrumento processual.⁸

O alcance dogmático desse movimento merece registro, uma vez que aquilo que a sentença produz é, em rigor, uma mutação dialética do art. 77, §6º, do CPC. Não há revogação do dispositivo, nem declaração de sua inconstitucionalidade incidental; tampouco há aplicação literal. O juízo, partindo da finalidade da norma — a proteção da independência funcional — chega à conclusão de que a conduta posta em julgamento se situa fora da hipótese de incidência.

O movimento é dialético, no sentido próprio: ele põe a tese (o advogado não pode ser sancionado diretamente pelo juízo, segundo o §6º), confronta-a com a antítese (a conduta examinada não integra o exercício regular da advocacia) e produz a síntese (a vedação não alcança o caso, e a sanção direta retorna ao campo do possível). Interpretar o §6º de modo a blindar o advogado que atenta deliberadamente contra o funcionamento do aparelho judicial, conclui a sentença, seria subverter a finalidade da norma, transformando uma garantia de independência profissional em escudo para práticas ilícitas (Brasil, 2026, p. 4).

Há, no movimento, ganhos e perdas. O ganho é evidente: a decisão recupera, no concreto, a efetividade do dever de lealdade processual e impede que o instituto da independência funcional opere como zona de imunidade para a manipulação algorítmica. Em termos de integridade processual — e aqui a contribuição de Huberts (2018, p. 18) é dogmaticamente relevante —, integridade é a qualidade moral do comportamento dos participantes no processo de governança, isto é, qualidade de agir em conformidade ou harmonia com os valores, normas e regras morais relevantes do contexto.⁹ A decisão de Parauapebas opera, portanto, no terreno da integridade processual em sentido forte: ela responde a uma conduta que viola não apenas regra positiva, mas o próprio padrão moral

⁸A leitura aqui proposta dialoga, em sentido crítico, com a posição de Nunes, Pereira e Chalfun (2026, n.p.), que, ao examinarem a mesma decisão, sustentam que a principal dificuldade dogmática do julgado reside em afastar a regra do art. 77, §6º, do CPC com base na afirmação de que a conduta estaria fora do exercício profissional. Para esses autores, o regime jurídico das sanções processuais no direito brasileiro é construído sobre uma distinção estrutural entre parte e advogado, e a sanção direta ao procurador, mesmo em hipótese tão grave, abriria espaço para uma flexibilização perigosa de garantias estruturais da advocacia, com potenciais impactos sobre a própria liberdade de atuação profissional. Reconhece-se a robustez da objeção. A posição sustentada neste artigo, contudo, é a de que o caso concreto autoriza a leitura teleológica do dispositivo — exatamente por se tratar de conduta que, por sua natureza, escapa à ratio protetiva do §6º —, sem que daí decorra autorização genérica para que o juízo redefina, ad hoc, o perímetro do exercício da advocacia. A divergência, em todo caso, sublinha a urgência de tipificação específica, capaz de dispensar a subsunção forçada que hoje sobrecarrega cláusulas gerais.

⁹Para Huberts (2018, p. 20), a integridade não diz respeito ao conteúdo das decisões ou aos resultados sociais (*outcome*), mas à qualidade moral do comportamento dos participantes no processo de tomada e implementação de decisões. Em sentido estrito, é a qualidade de agir em conformidade ou harmonia com os valores, normas e regras morais relevantes para o contexto.

exigido pelo processo cooperativo, e, ao fazê-lo, contribui para a institucionalização de uma boa prática judicial no sentido descrito por Didier Jr. e Fernandez (2022, p. 42-44).

As perdas, contudo, não devem ser silenciadas. A subsunção, ainda que sustentada por argumento teleológico convincente, opera sobre cláusulas gerais e sobre dispositivo cuja redação pressupõe — ou ao menos sugere — a apuração disciplinar prévia pelo órgão de classe. O direito disciplinar da advocacia, sobre essa base, fica ferido em duplicidade: ou se entende que o *prompt injection* é fato profissional, e sua apuração caberia primariamente à OAB; ou se entende que é fato a-profissional, fora do mandato, e a sanção judicial direta é cabível, restando à OAB apenas a apuração disciplinar autônoma. A sentença adota a segunda leitura, e o caso concreto a justifica; mas a generalização do raciocínio para futuros casos demanda atenção. O risco é o de uma aplicação seletiva, em que tudo o que o juízo reputar ilícito do advogado tenda a ser reclassificado como conduta fora do mandato — quebrando o equilíbrio originalmente protegido pelo §6º.

Há, ainda, um terceiro elemento que o caso põe a nu: o direito brasileiro não oferece — em nenhum diploma — protocolo nacional para o tratamento processual da injeção descoberta. Não há regra clara sobre o procedimento de revelação do *payload* (no caso de Parauapebas, a alteração técnica da cor da fonte foi o que possibilitou a leitura humana do comando); não há rito pericial específico; não há disciplina sobre conservação dos artefatos, sobre congelamento de versões, sobre reprocessamento controlado ou sobre contraditório técnico, conforme apontam Badaró e Puppe (2025, n.p.). A consequência é que cada juízo, quando confrontado com a hipótese, terá de improvisar o seu próprio rito a partir de cláusulas gerais — o que produz insegurança jurídica e abre espaço para tratamentos heterogêneos. A subsunção forçada na qual hoje se ampara o juiz é, assim, sintoma de um arcabouço normativo pré-algorítmico aplicado a uma fenomenologia algorítmica¹⁰.

Resta, finalmente, a pergunta que conecta este capítulo ao seguinte. A decisão de Parauapebas representa, por todos os ângulos relevantes, a hipótese mais favorável dentro do paradoxo regulatório aqui sustentado: o *prompt injection* foi detectado, foi enfrentado expressamente, foi sancionado. A integridade processual, no caso concreto, foi preservada — e ainda assim a resposta jurídica precisou operar por subsunção forçada e por mutação dialética.

¹⁰Nunes, Pereira e Chalfun (2026, n.p.) qualificam esse descompasso como sintoma de uma autêntica virada tecnológica do direito processual, em que “o direito processual contemporâneo, estruturado a partir de categorias concebidas em um contexto pré-algorítmico, ainda não dispõe de instrumentos normativos plenamente adequados para lidar com formas indiretas e tecnicamente sofisticadas de manipulação do ambiente decisório”. Daí defenderem que a resposta institucional ao *prompt injection* não pode se limitar à adaptação forçada de categorias tradicionais; é necessário avançar tanto no plano legislativo — com eventuais ajustes no CPC — quanto no plano regulatório, mediante normas do CNJ que enfrentem diretamente práticas fraudulentas como essa.

Permanece em aberto, contudo, o lado regulatoriamente mais grave da equação: o que ocorre quando a injeção não é detectada e a saída algorítmica contaminada flui, silenciosamente, para a base do iter decisório? Essa é a face oculta do problema — e dela trata o próximo capítulo.

4 A NÃO DETECÇÃO DO PROMPT INJECTION: OPACIDADE INSTITUCIONAL E VIOLAÇÃO IMPLÍCITA DA RESOLUÇÃO CNJ Nº 615/2025

A hipótese de injeção bem-sucedida — isto é, aquela em que o comando oculto não é detectado pelo juízo, pelos servidores ou pelos demais sujeitos processuais, e o sistema de IA, processando o texto contaminado, gera saída enviesada que é incorporada à pré-análise do feito — é a mais grave do ponto de vista regulatório. Altera-se o eixo da discussão: o problema deixa de ser apenas o comportamento da advocacia e passa a ser também, e talvez sobretudo, o comportamento da unidade jurisdicional. Quando uma instrução oculta encontra terreno fértil em um sistema cuja saída ninguém audita substantivamente, a injeção bem-sucedida não é falha individual; é prova de que o filtro humano que a Resolução CNJ nº 615/2025 exige não esteve presente.

O ponto de apoio normativo da afirmação é múltiplo, analisando que a Resolução CNJ nº 615/2025 fixa, como diretriz central da governança da IA no Poder Judiciário, a participação e supervisão humana em todas as etapas dos ciclos de vida das soluções (Brasil, 2025). O art. 19, §3º, II, do mesmo ato normativo, é mais explícito: o uso de ferramentas de inteligência artificial será de caráter auxiliar e complementar, consistindo em mecanismos de apoio à decisão, vedada expressamente sua utilização como instrumento autônomo de tomada de decisões judiciais sem a devida orientação, interpretação, verificação e revisão crítica por parte do magistrado, que permanecerá integralmente responsável pelas decisões e pelas informações nelas contidas (Brasil, 2025).

O art. 32, I e II, complementa o desenho, exigindo que o sistema assegure autonomia plena dos usuários internos e possibilite revisão detalhada e facilitada do conteúdo gerado e dos dados utilizados, com acesso transparente às premissas e ao método empregado (Brasil, 2025).

Os dispositivos compõem, em conjunto, uma cláusula geral de *human in the loop*: o magistrado não pode aderir, em bloco, ao *output* da máquina; deve, ao contrário, exercer o que o ato normativo chama de revisão crítica¹¹. Quando uma *injection* bem-sucedida produz o

¹¹Pessoa e Albino (2026, p. 51) advertem que a Resolução CNJ nº 615/2025, ao prever a explicabilidade das decisões automatizadas “sempre que tecnicamente possível”, abre uma zona de flexibilidade que pode ser invocada para legitimar a opacidade em sistemas mais complexos, como redes neurais profundas. Os autores sublinham que essa ressalva acaba refletindo uma limitação estrutural das tecnologias mais avançadas, em que a falta de previsibilidade e transparência é inerente ao modelo algorítmico: “quanto mais poderosos e complexos os sistemas,

resultado que o atacante perseguia — minuta de contestação superficial, minuta de sentença comprometida, classificação errada de matéria, indexação enviesada — e esse resultado é absorvido pelo iter sem que o filtro humano substantivo o pondere, opera-se a ruptura do modelo cooperativo de processo (Marinoni; Arenhart; Mitidiero, 2021, p. 13-14): a decisão deixa de ser produto da colaboração entre juiz e partes e passa a ser, em parte, produto de uma camada técnica não auditada. Em tal cenário, a não detecção não é um acidente que escapou ao juízo; é prova fenomenológica de que o filtro normativamente exigido não foi exercido.

A gravidade dessa hipótese é potencializada pela opacidade institucional. Em regra, o jurisdicionado e o advogado da contraparte não sabem qual sistema de IA está sendo utilizado pela unidade jurisdicional, em que fase processual ele atua, com qual grau de autonomia, com qual base de dados, sob qual *system prompt*¹². A própria Resolução CNJ nº 615/2025 institui o Sinapses como repositório centralizado, mas não basta a existência de cadastro; é preciso que o conteúdo da operação seja, de fato, escrutinável. Pasquale (2015, p. 6-7) descreve, com precisão, as três estratégias que mantêm as caixas-pretas fechadas: o segredo real (barreiras técnicas), o segredo legal (propriedade intelectual e segredos comerciais) e a ofuscação (complexidade deliberada).¹³ As três estratégias, no caso da IA judicial brasileira, atuam de

mais difícil é explicá-los, e as instituições acabam aceitando essa obscuridade como inevitável, naturalizando a *black box society*” (Pessoa; Albino, 2026, p. 51). A cláusula técnica de escape opera, pois, como porta institucional para a opacidade — exatamente a vulnerabilidade aqui sustentada como sintomática da não detecção da injeção.

¹²Essa realidade diz respeito a um fenômeno que Pessoa e Albino (2026, p. 56) chamam de *shadow AI governance*. Não se trata da ausência pura e simples de regras, mas de uma governança paralela, difusa e não formalizada, que opera nos vácuos da regulação explícita e nas camadas profundas da infraestrutura digital judicial. Diferente da governança documental — centrada em cadastros, classificações de risco e protocolos de auditoria —, a *shadow AI governance* reside no *backend* dos sistemas, onde escolhas de *design*, pesos estatísticos e critérios de seleção de premissas determinam, silenciosamente, a arquitetura das saídas oferecidas ao magistrado. Sua característica distintiva é deslocar o *locus* efetivo do poder decisório para camadas técnicas que escapam ao escrutínio das partes, dos órgãos de classe e do próprio juízo, produzindo uma normatividade técnica que opera por trás de interfaces amigáveis ao usuário. No contexto deste artigo, a *injection* não detectada é o sintoma mais nítido dessa governança fantasma: ela só produz efeito porque o sistema, embora formalmente regulado pela Resolução CNJ nº 615/2025, opera, de fato, sob uma camada de governança paralela cuja existência o direito posto reconhece com dificuldade.

¹³Pasquale (2015, p. 6-7) sistematiza três estratégias críticas pelas quais as chamadas caixas-pretas algorítmicas permanecem fechadas, cada uma operando por um mecanismo distinto. A primeira é o segredo real (*real secrecy*): trata-se da edificação de barreira material entre o conteúdo oculto e o acesso não autorizado a ele, mecanismo cotidianamente exemplificado pelo uso de fechaduras, senhas e dispositivos de cifragem (Pasquale, 2015, p. 6); no caso dos sistemas algorítmicos, materializa-se na própria complexidade técnica dos modelos e nas restrições de acesso à arquitetura, ao código-fonte, aos dados de treinamento e aos pesos estatísticos, de modo que mesmo o sujeito interessado, ainda que dotado de autorização legal para conhecer o sistema, esbarra na barreira fática de sua inacessibilidade. A segunda é o segredo jurídico (*legal secrecy*): a obrigação imposta a quem tem acesso à informação de não divulgá-la, sustentada por previsão legal ou contratual — como ocorre no dever de sigilo do empregado bancário em relação aos saldos dos clientes (Pasquale, 2015, p. 6) e, no domínio que aqui interessa, na proteção dos algoritmos por institutos como propriedade intelectual, segredo industrial e segredo comercial, que conferem ao titular o direito de não revelar a lógica decisória mesmo quando demandado por escrutínio público ou judicial. A terceira é a ofuscação (*obfuscation*): tentativa deliberada de ocultação quando a primeira camada de sigilo já foi comprometida; opera não pela negativa de acesso, mas pelo afogamento informacional do investigador — o exemplo paradigmático é o da empresa que, instada a fornecer dados, responde com trinta milhões de páginas de documentos, forçando quem investiga a buscar agulha em palheiro (Pasquale, 2015, p. 6). No campo

modo cumulativo: o sistema é tecnicamente complexo; sua arquitetura proprietária está, em parte, protegida por sigilo comercial; e a documentação pública é, em regra, pouco didática a respeito das efetivas margens de autonomia conferidas ao modelo.

Hildebrandt (2018, p. 259, 265-269) oferece, neste ponto, a melhor moldura teórica. Para a autora, a economia política dos sistemas orientados por dados é caracterizada por uma operação que ocorre, predominantemente, no *backend* dos sistemas — camada onde as decisões de *design*, os pesos estatísticos e as funcionalidades ocultas se determinam, à revelia do que os usuários veem na interface do *frontend*. Esse deslocamento implica uma redução do que Hildebrandt (2018, p. 259) chama de descontinuidade semântica — o intervalo conceitual que permite, na linguagem natural e no direito, a dissonância, o dissenso e a reinvenção. No processo judicial, a redução da descontinuidade semântica significa, na prática, que a pré-análise automatizada passa a operar como camada normativa de fato: ela hierarquiza informações, identifica precedentes, propõe enquadramentos. Se essa camada não é submetida a controle substantivo pelo juízo, ela termina por dizer, ainda que silenciosamente, o direito¹⁴.

Wisser (2019, p. 1815-1816) ilustra esse fenômeno com o famoso caso *State v. Loomis*, no qual a Suprema Corte do Wisconsin admitiu a utilização do sistema COMPAS — algoritmo proprietário de avaliação de risco de reincidência — na dosimetria da pena, apesar de seu funcionamento interno permanecer inacessível tanto à defesa quanto aos próprios julgadores.¹⁵ Como anota a autora, o problema da opacidade não é apenas de comunicação: ele consiste no

algorítmico, a ofuscação se traduz em complexidade desnecessária, em documentação técnica deliberadamente prolixa e em explicações que, embora formalmente fornecidas, não permitem compreensão substantiva do que efetivamente ocorre no sistema. O resultado conjunto dessas três estratégias é o que o autor denomina opacidade, conceito que define como incompreensibilidade remediável (Pasquale, 2015, p. 7) — termo guarda-chuva que abarca tanto o sigilo intencional quanto a complexidade estrutural, sublinhando o caráter contingente e, portanto, juridicamente endereçável do fenômeno.

¹⁴Pessoa e Albino (2026, p. 56-57) tematizam, nesse mesmo registro, o risco do que Zuboff (2015) denomina *Big Other*: aparato invisível de classificação e predição que interfere silenciosamente em processos decisórios públicos. Os autores observam que a Resolução CNJ nº 615/2025, ao tornar facultativo o registro do uso de ferramentas externas de IA generativa por magistrados, “cria uma governança paralela invisível”, em que “decisões judiciais passam a ser influenciadas por modelos comerciais opacos, sem que haja qualquer rastreabilidade pública ou controle institucional sobre como essas ferramentas foram usadas” (Pessoa; Albino, 2026, p. 56). Daí a tensão com o princípio da identidade física do juiz: quando um LLM redige a minuta de uma decisão ou organiza a fundamentação, ainda sob revisão humana, emerge a indagação sobre quem é o verdadeiro autor intelectual do ato (Pessoa; Albino, 2026, p. 57). A não detecção do *prompt injection*, aqui examinada, é o ponto exato em que essa autoria difusa se converte em vulnerabilidade processual concreta: o efeito do ataque flui silenciosamente para a base do iter decisório, sem deixar rastros.

¹⁵O caso *State v. Loomis*, julgado pela Suprema Corte do Wisconsin, envolveu o uso da pontuação de risco gerada pelo sistema COMPAS na dosimetria da pena. Como o algoritmo é proprietário, nem a defesa nem os próprios julgadores tinham acesso à sua lógica interna, o que evidencia, segundo Wisser (2019, p. 1815-1816), o paradoxo de decisões juridicamente relevantes serem influenciadas por sistemas cuja ponderação dos diferentes dados permanece inacessível.

fato de que conhecemos as informações de entrada e o resultado, mas não sabemos como o algoritmo pondera ou processa os diferentes dados (Wisser, 2019, p. 1815-1816).

Transposto à hipótese aqui em análise, o paralelo é direto: quando o sistema de IA do Judiciário, contaminado por uma *injection* não detectada, produz a saída desejada pelo atacante, o juízo se torna, de modo análogo a *Loomis*, terceiro consumidor de uma saída cuja gênese não pode auditar — com o agravante de que, nesta hipótese, a contaminação foi introduzida por uma das partes, e não pela arquitetura do sistema em si.

Há, ainda, um plano institucional que precisa ser nomeado. Fukuyama (2013, p. 3) define governança como a capacidade de um governo de fazer e fazer cumprir regras e de entregar serviços, com ênfase no que Michael Mann chama de poder infraestrutural. A governança da IA no Judiciário, nesse sentido, não se resume à edição de uma resolução: depende da capacidade efetiva de fazer cumprir a supervisão humana, de auditar as ferramentas em uso, de articular o sistema judicial com os órgãos de classe, com o CNJ, com a comunidade técnica. A simples existência de regra — como a Resolução CNJ nº 615/2025 — sem capacidade institucional de implementação configura aquilo que Huberts (2018, p. 18) descreveria como déficit de qualidade moral do processo de governança: as regras existem, mas o processo cotidiano de tomada e implementação de decisões não se ajusta a elas. A injeção bem-sucedida é, então, mais do que um caso isolado: é sintoma desse déficit.

Costa (2026, p. 14) já advertia que a literatura brasileira sobre IA no Judiciário concentra-se predominantemente em questões éticas gerais, princípios abstratos de governança e preocupações sobre discriminação algorítmica, ao passo que praticamente inexistente discussão substancial sobre vulnerabilidades técnicas específicas de segurança cibernética que comprometem a integridade operacional de sistemas aplicados ao Judiciário. A consequência prática é que a Resolução CNJ nº 615/2025 estabelece requisitos gerais de segurança, auditoria e monitoramento contínuo, mas não especifica tecnicamente como detectar ou prevenir, em particular, ataques de injeção de *prompt* (Costa, 2026, p. 14).

A literatura técnica internacional, por outro lado, fornece um catálogo de defesas: sanitização de entradas, isolamento de fontes confiáveis e não confiáveis, *firewalls* semânticos, modelos auditores independentes que detectem sinais de injeção, decodificação restrita, registros de auditoria (*logs*) com somas de verificação (*hashes*) e versionamento (Lan; Kaul; Jones, 2025, p. 3-4; Gosmar; Dahl; Gosmar, 2025, p. 5-6; Jones; Mohammed, 2026, p. 162-

163)¹⁶. O hiato entre o que a literatura técnica oferece e o que a norma jurídica nacional incorpora é, em si, um indicador do problema.

A consequência regulatória da não detecção, à luz do exposto, é dupla. Em primeiro lugar, ela põe em xeque o cumprimento, pela unidade jurisdicional, da exigência de supervisão humana inscrita na Resolução CNJ nº 615/2025: se a injeção produziu o resultado pretendido, o filtro humano não exerceu, em algum ponto, sua função de revisão crítica. Em segundo lugar, ela transforma o contraditório em assimetria estrutural: o advogado que injetou conhece o ataque; o advogado da contraparte e o jurisdicionado não sabem nem da existência do sistema de IA, nem da injeção, nem da contaminação. O contraditório, que pressupõe paridade comunicativa, vira batalha de *prompts* — disputa cuja arena permanece, para metade dos contendores, invisível.

A resposta dogmática mais robusta para o problema parece estar na noção de proteção jurídica pelo *design*, proposta por Hildebrandt (2018, p. 269-272). Diferentemente da concepção de *legal by design*, que busca a execução automática da norma pela máquina, a proteção jurídica pelo *design* visa integrar, na arquitetura técnica dos sistemas, garantias constitucionais como o contraditório, a igualdade de armas e a possibilidade efetiva de contestação. No caso da IA judicial, isso significa que a defesa contra a injeção não pode ser pensada como reação *ex post* ao caso descoberto; deve ser incorporada à própria arquitetura dos sistemas — por meio de sanitização de entradas, isolamento entre corpus probatório e instruções de sistema, modelos auditores, trilhas de auditoria — e regulada por norma que vincule sua adoção (Lan; Kaul; Jones, 2025, p. 4-5; Jones; Mohammed, 2026, p. 162). Sem essa camada técnica, a cláusula da supervisão humana corre o risco de ser, na prática, *pro forma*.

Daí a segunda conclusão parcial do artigo. A não detecção do *prompt injection* não é, do ponto de vista regulatório, mero acidente; é prova de que a supervisão humana exigida pela Resolução CNJ nº 615/2025 não se efetivou. A insuficiência do arcabouço normativo, nesta ponta, é mais grave do que na anterior: enquanto a sanção do advogado descoberto, ainda que por subsunção forçada, mobiliza dispositivos que existem, a fiscalização do uso da IA pelo

¹⁶Convém esclarecer brevemente o sentido das defesas técnicas mencionadas: a *sanitização de entradas* consiste no tratamento prévio do material textual submetido ao modelo, com a remoção ou neutralização de elementos potencialmente maliciosos (caracteres invisíveis, instruções embutidas, fragmentos com formatação suspeita); o *isolamento de fontes confiáveis e não confiáveis* é a separação estrita, em distintos canais de processamento, entre instruções legítimas do sistema e dados oriundos de terceiros; os *firewalls semânticos* são filtros que avaliam, no plano do significado, se o conteúdo recebido contém pedidos de subversão das regras do sistema; os *modelos auditores* são instâncias adicionais de inteligência artificial que monitoram a saída do modelo principal, detectando sinais de injeção; a *decodificação restrita* limita o repertório de respostas que o modelo pode produzir; e as trilhas de auditoria (*logs*) com somas de verificação (*hashes*) e versionamento são registros íntegros, datados e à prova de adulteração, que permitem reconstituir, posteriormente, o que efetivamente ocorreu no processamento.

Judiciário esbarra na opacidade institucional sobre as próprias ferramentas em uso¹⁷. O paradoxo binário, anunciado na introdução, se confirma: a insuficiência regulatória reproduz-se nos dois polos da relação processual, e o polo estatal — pela responsabilidade institucional que comporta — é o mais sensível.

5 CONSIDERAÇÕES FINAIS

O percurso traçado permite responder, com diretividade, à pergunta que orientou a investigação. O uso de *prompt injections* em peças processuais compromete a integridade do processo judicial em três planos. No plano técnico, explora uma fragilidade arquitetural dos modelos de linguagem, que não distinguem ontologicamente entre instruções e dados, transformando o documento em vetor de manipulação invisível. No plano ético-jurídico, viola, em sua dimensão objetiva, a boa-fé processual e a cooperação entre os sujeitos do processo, substituindo a batalha de argumentos por uma batalha de *prompts* cuja arena permanece, para parte dos contendores, invisível. No plano institucional, denuncia a insuficiência do arcabouço normativo brasileiro — e, em particular, da Resolução CNJ nº 615/2025 — para responder, com adequação e proporcionalidade, à fenomenologia algorítmica do ilícito.

A hipótese do paradoxo regulatório binário confirma-se. Detectada a injeção, como ocorreu na 3ª Vara do Trabalho de Parauapebas, a resposta sancionatória depende de cláusulas gerais e de uma mutação dialética do art. 77, §6º, do CPC, que afasta a vedação de sanção direta ao advogado para alcançar a conduta situada fora do espectro do exercício regular da advocacia. A solução é dogmaticamente sustentável no caso concreto, mas opera por subsunção forçada, sem que o direito processual disponha, especificamente, de tipo ou rito desenhado para a hipótese. Não detectada, a injeção bem-sucedida expõe algo mais grave: a unidade jurisdicional incorporou, sem o filtro humano substantivo exigido pela Resolução CNJ nº 615/2025, a saída algorítmica contaminada. A não detecção, nessa hipótese, é a prova mais robusta da violação do dever de supervisão — porque, se houvesse supervisão crítica, a saída teria sido contrastada, e a injeção teria perdido, ao menos parcialmente, sua eficácia.

¹⁷Pessoa e Albino (2026, p. 58-59) chamam a atenção, em registro convergente, para o fato de que a governança da IA no Judiciário brasileiro está assentada predominantemente em instrumentos de *soft law* — como as resoluções administrativas do CNJ —, modelo que, embora ofereça flexibilidade regulatória, “carece de densidade normativa legislativa, estabilidade democrática reforçada e mecanismos mais robustos de controle externo”. A dependência de normas infralegais, somada à autorregulação dos próprios tribunais, pode fragilizar a efetividade das salvaguardas previstas, sobretudo diante de assimetrias de implementação, resistências corporativas e limitações na exigibilidade jurídica de direitos como transparência, contestação e explicabilidade algorítmica (Pessoa; Albino, 2026, p. 59). Daí a necessidade de transição, eventualmente, para modelos híbridos, com maior participação do legislador e do controle social formal.

Decorre dessa simetria a tese central do artigo. A insuficiência regulatória do *prompt injection* no Brasil não se localiza em apenas uma ponta da relação processual: replica-se nos polos privado (advocacia clandestina) e estatal (uso oculto de IA pelo Judiciário). A repressão exclusiva da advocacia, ainda que conduzida com a firmeza demonstrada em Parauapebas, é insuficiente, porque não enfrenta a contraparte do problema — a opacidade institucional sobre as ferramentas de IA efetivamente utilizadas pela jurisdição, em qual fase processual, com qual grau de autonomia, sob qual *system prompt*. Sem essa transparência, o dever de supervisão humana corre o risco de ser cláusula meramente formal, e a integridade processual depende, paradoxalmente, da incompetência do atacante.

Confirma-se, do ponto de vista regulatório, que a não detecção é mais grave do que a detecção. Detectado o ataque, há, ao menos, a possibilidade de resposta — ainda que por subsunção forçada. Não detectado, o ataque produz seu efeito sem que sequer se saiba que ele ocorreu, e a integridade processual deteriora-se sem escrutínio. A integridade é qualidade moral do comportamento dos participantes no processo de governança; sua salvaguarda exige, portanto, que a governança não opere apenas sobre o produto final da decisão, mas sobre todo o processo de sua formação técnica e lógica.

Daí a agenda de pesquisa que esta investigação propõe. No plano empírico, é necessário avançar em estudos quantitativos sobre a presença de injeções em peças protocoladas, com cooperação técnica entre tribunais, CNJ e comunidade acadêmica — pesquisas que se beneficiariam, do ponto de vista metodológico, das técnicas de detecção descritas pela literatura técnica. No plano comparado, impõe-se o diálogo com as iniciativas europeias, em particular a Carta Ética Europeia sobre o Uso da Inteligência Artificial nos Sistemas Judiciais e seu Ambiente (CEPEJ) e o *AI Act* da União Europeia, cujo regime de risco e cujas exigências de transparência podem oferecer subsídios para o aprimoramento da Resolução CNJ nº 615/2025.

No plano normativo nacional, impõe-se a reformulação cooperativa entre CNJ, OAB e tribunais, com vistas à edição de protocolo nacional anti-*injection* que articule, em uma só arquitetura, deveres da advocacia (incluindo a obrigação de declarar uso de IA e a vedação expressa de inserção de comandos ocultos), deveres dos tribunais (transparência ativa sobre as ferramentas de IA em uso, rito pericial específico para incidentes algorítmicos, sanitização de entradas) e deveres da OAB (apuração disciplinar em prazo razoável, com tipificação específica). É somente nessa articulação simétrica que o paradoxo regulatório binário poderá ser, se não dissolvido, ao menos enfrentado em suas duas pontas.

A travessia entre o processo analógico e o processo algorítmico, como advertem Badaró e Puppe (2025), não autoriza relativizações éticas. Tampouco autoriza, deve-se acrescentar, que

o Poder Judiciário se exima do dever que ele próprio impôs à advocacia: o de operar sob a luz, sem comandos invisíveis. A reserva de humanidade, que é o núcleo da Resolução CNJ nº 615/2025, só se efetiva se o filtro humano for, de fato, exercido. Caso contrário, a integridade processual permanecerá refém da incompetência alheia — o que, em uma democracia constitucional, é proteção tão frágil quanto a fonte branca sobre fundo branco que, em Parauapebas, quase escapou aos olhos do juízo.

REFERÊNCIAS

BADARÓ, Rodrigo; PUPPE, Matheus. Prompt injection: ameaça invisível à imparcialidade do Judiciário na era da IA. **Consultor Jurídico**, 27 set. 2025. Disponível em: <https://www.conjur.com.br/2025-set-27/prompt-injection-ameaca-invisivel-a-imparcialidade-do-judiciario-na-era-da-ia/>. Acesso em: 10 mar. 2026.

BRASIL. Conselho Nacional de Justiça. **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. Brasília, DF: CNJ, 2025. Disponível em: <https://atos.cnj.jus.br/files/original1555302025031467d4517244566.pdf>. Acesso em 10 mar. 2026.

BRASIL. Tribunal Regional do Trabalho da 8ª Região. 3ª Vara do Trabalho de Parauapebas. **Sentença nos autos da ATOrd 0001062-55.2025.5.08.0130**. Reclamante: Elisandro Martins de Barros. Reclamado: Renato Ribeiro de Lima. Juiz: Luiz Carlos de Araújo Santos Junior. Parauapebas, 12 maio 2026. Disponível em: <https://www.conjur.com.br/wp-content/uploads/2026/05/trt-8-processo.pdf>. Acesso em 13 mai. 2026.

COSTA, Lucas Rodrigues da. Prompt injection e inteligência artificial no Judiciário brasileiro: riscos de manipulação, deveres processuais e a preservação da subsunção judicial como garantia fundamental. **Revista da ESMAL**, Maceió, n. 11, p. 1-23, 2026. Disponível em: <https://revistadaesmal.tjal.jus.br/index.php/revistaEletronicaEsmal/article/view/352/402>. Acesso em 04 mai. 2026.

DIDIER JR., Fredie. **Curso de direito processual civil: introdução ao direito processual civil, parte geral e processo de conhecimento**. 21. ed. v. 1. Salvador: Juspodivm, 2019.

FERNANDEZ, Leandro; DIDIER JR., Fredie. Introdução ao estudo das boas práticas na administração da justiça: a relevância dogmática da inovação. **Revista Brasileira de Direitos Fundamentais & Justiça**, [S. l.], v. 15, n. 45, p. 165–198, 2022. DOI: 10.30899/dfj.v15i45.1244. Disponível em: <https://dfj.emnuvens.com.br/dfj/article/view/1244>. Acesso em: 13 maio. 2026.

FUKUYAMA, Francis. What is governance? **Center for Global Development Working Paper**, n. 314, jan. 2013. Disponível em: https://www.cgdev.org/sites/default/files/1426906_file_Fukuyama_What_Is_Governance.pdf. Acesso em 09 de mar. 2026.

GOSMAR, Diego; DAHL, Deborah A.; GOSMAR, Dario. Prompt injection detection and mitigation via AI multi-agent NLP frameworks. **arXiv**. 2503.11517v1, 17 mar. 2025. Disponível em: <https://arxiv.org/abs/2503.11517>. Acesso em 09 de mar. 2026.

HILDEBRANDT, Mireille. Primitives of legal protection in the era of data-driven platforms. **Georgetown Law Technology Review**, Washington, v. 2, n. 2, p. 252-273, 2018. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3140594. Acesso em 12 dez. 2025.

HUBERTS, Leo W. J. C. Integrity: what it is and why it is important. **Public Integrity**, v. 20, sup. 1, p. S18-S32, 2018. Disponível em: <https://doi.org/10.1080/10999922.2018.1477404>. Acesso em: 10 set. 2025.

JONES, Rebet Keith; MOHAMMED, Aquil Mirza. Prompt injection attacks risks and defense mechanisms. In: OMAR, M.; ZANGANA, H. M. (Ed.). **Application of large language models (LLMs) for software vulnerability detection**. Hershey: IGI Global, 2026, p. 149-185.

LAN, Qianlong; KAUL, Anuj; JONES, Shaun. Prompt injection detection in LLM integrated applications. **International Journal of Network Dynamics and Intelligence**, v. 4, n. 2, art. 100013, p. 1-10, 2025. Disponível em: <https://media.sciltp.com/articles/2506000841/2506000841.pdf>. Acesso em 04 abr. 2026.

MARINONI, Luiz Guilherme; ARENHART, Sérgio Cruz; MITIDIERO, Daniel. **Novo curso de processo civil: teoria do processo civil**. 6. ed. São Paulo: Revista dos Tribunais, 2021.

MURRAY, Michael D. Prompt engineering and priming in law. **University of Kentucky Rosenberg College of Law**, 29 jul. 2024. Disponível em: <https://papers.ssrn.com/abstract=4909532>. Acesso em: 24 abr. 2026.

NUNES, Dierle; PEREIRA, João Sérgio; CHALFUN, Gustavo. IA e processo judicial: entre a fraude invisível e o erro no caso do prompt injection no TRT-8. **Consultor Jurídico**, 14 maio 2026. Disponível em: <https://www.conjur.com.br/2026-mai-14/ia-e-processo-judicial-entre-a-fraude-invisivel-e-o-erro-no-caso-do-prompt-injection-no-trt-8/>. Acesso em: 14 maio 2026.

PASQUALE, Frank. **The black box society: the secret algorithms that control money and information**. Cambridge: Harvard University Press, 2015.

PESSOA, Wendelson Pereira; ALBINO, Dennys Damião Rodrigues. Governança algorítmica no Judiciário: análise crítica da Resolução CNJ nº 615/2025 e seus desafios regulatórios. In: OLIVEIRA, Jamil de Jesus; FRAZÃO, Hugo Abas; PESSOA, Wendelson Pereira (Org.). **Justiça digital: governança algorítmica, dados e jurisdição**. Brasília: Tribunal Regional Federal da 1ª Região, 2026. v. 1. p. 44-61. Disponível em: <https://www.ajufe.org.br/images/2026/PDF/justica-digital.pdf>. Acesso em 10 mai. 2026.

RUSSELL, Stuart; NORVIG, Peter. **Artificial intelligence: a modern approach**. 3. ed. Upper Saddle River: Prentice Hall, 2010. Disponível em: <https://people.engr.tamu.edu/guni/csce625/slides/AI.pdf>. Acesso em 27 fev. 2026.

WISSER, Leah. Pandora's algorithmic black box: the challenges of using algorithmic risk assessments in sentencing. **American Criminal Law Review**, Washington, v. 56, n. 4, p.

1811-1832, 2019. Disponível em: <https://www.law.georgetown.edu/american-criminal-law-review/wp-content/uploads/sites/15/2019/06/56-4-Pandoras-Algorithmic-Black-Box-The-Challenges-of-Using-Algorithmic-Risk-Assessments-in-Sentencing.pdf>. Acesso em 22 nov. 2025.