ReGeo ISSN: 2177-3246

O TERCEIRO LEITOR DA PETIÇÃO: A INTELIGÊNCIA ARTIFICIAL COMO DESTINATÁRIA OCULTA E A BARREIRA DE INTEGRIDADE ALGORÍTMICA NO ACESSO À JUSTIÇA

PROMPT INJECTION, BOA-FÉ PROCESSUAL E OS LIMITES ENTRE TRADUÇÃO SEMIÓTICA E MANIPULAÇÃO CLANDESTINA

PROMPT INJECTION, BUENA FE PROCESAL Y LOS LÍMITES ENTRE TRADUCCIÓN SEMIÓTICA Y MANIPULACIÓN CLANDESTINA

https://doi.org/10.56238/revgeov17n7-198

Data de submissão: 30/06/2026 **Data de publicação:** 31/07/2026

Cristiano Gomes Mazzini
Instituição: Universidade Federal de Rondônia
E-mail: cristianomazzini@gmail.com
Lattes: http://lattes.cnpq.br/9413151734286079
Orcid: https://orcid.org/0000-0003-1430-1069

RESUMO

Este artigo investiga o prompt injection em petições judiciais como barreira emergente ao acesso à justiça. Parte da constatação de que a inteligência artificial deslocou-se de ferramenta externa para componente da infraestrutura cognitiva do Poder Judiciário, integrando o que a doutrina denomina sexta onda renovatória. A pesquisa sustenta que essa incorporação amplia a capacidade institucional de processamento e compreensão das demandas, mas cria vulnerabilidade nova: a exposição da mediação algorítmica a comunicações ocultas dirigidas exclusivamente à máquina. O trabalho, de natureza teórica e qualitativa, adota o método hipotético-dedutivo e vale-se de análise documental de decisões brasileiras proferidas em 2026, de atos normativos do Conselho Nacional de Justiça e de literatura técnica de segurança da informação. A pesquisa caracteriza a inteligência artificial como destinatária operacional intermediária da comunicação processual, um terceiro leitor que interpreta o documento sem ocupar posição de sujeito, e identifica, a partir dessa categoria, uma dupla camada documental: a humana, visível e submetida ao contraditório, e a maquínica, invisível e orientada a modificar o processamento automatizado. Formula-se, com base nessas duas categorias, a barreira de integridade algorítmica como obstáculo interno emergente à sexta onda, e distingue-se a tradução semiótica legítima, que explicita e democratiza sentidos, da manipulação algorítmica clandestina, que oculta instruções e se apropria privadamente do canal tecnológico. A resposta adequada ao fenômeno, conclui-se, passa pela construção de governança fundada em ingestão segura de documentos, supervisão humana, transparência, auditabilidade e responsabilização proporcional, e não pelo abandono da tecnologia.

Palavras-chave: Acesso à Justiça. Inteligência Artificial. Prompt Injection. Boa-Fé Processual. Tradução Semiótica.

ABSTRACT

This article investigates prompt injection in court filings as an emerging barrier to access to justice. It departs from the observation that artificial intelligence has shifted from an external tool to a component of the cognitive infrastructure of the Judiciary, integrating what scholarship terms the sixth renovatory wave. The research argues that this incorporation expands institutional capacity to process and understand claims, but creates a new vulnerability: the exposure of algorithmic mediation to hidden communications addressed exclusively to the machine. The work, of a theoretical and qualitative

nature, adopts the hypothetical-deductive method and draws on documentary analysis of Brazilian decisions rendered in 2026, normative acts of the National Council of Justice, and technical information security literature. The research characterises artificial intelligence as an intermediary operational addressee of procedural communication, a third reader that interprets the document without occupying the position of a subject, and identifies, on that basis, a double documentary layer: the human one, visible and subject to the adversarial principle, and the machine one, invisible and oriented to modify automated processing. It formulates the algorithmic integrity barrier as an emerging internal obstacle to the sixth wave, and distinguishes legitimate semiotic translation, which makes meanings explicit and democratic, from clandestine algorithmic manipulation, which hides instructions and privately appropriates the technological channel. The adequate response, it concludes, lies in governance grounded in secure document ingestion, human oversight, transparency, auditability and proportionate accountability, rather than in the abandonment of technology.

Keywords: Access to Justice. Artificial Intelligence. Prompt Injection. Procedural Good Faith. Semiotic Translation.

RESUMEN

Este artículo investiga la inyección de instrucciones (prompt injection) en peticiones judiciales como una barrera emergente para el acceso a la justicia. Parte de la constatación de que la inteligencia artificial pasó de ser una herramienta externa a convertirse en un componente de la infraestructura cognitiva del Poder Judicial, integrando lo que la doctrina denomina sexta ola renovatoria. La investigación sostiene que esta incorporación amplía la capacidad institucional de procesamiento y comprensión de las demandas, pero genera una nueva vulnerabilidad: la exposición de la mediación algorítmica a comunicaciones ocultas dirigidas exclusivamente a la máquina. El estudio, de naturaleza teórica y cualitativa, adopta el método hipotético-deductivo y utiliza el análisis documental de decisiones judiciales brasileñas dictadas en 2026, actos normativos del Consejo Nacional de Justicia y literatura técnica de seguridad de la información. La investigación caracteriza la inteligencia artificial como destinataria operativa intermediaria de la comunicación procesal, una tercera lectora que interpreta el documento sin ocupar la posición de sujeto jurídico, e identifica, a partir de esta categoría, una doble capa documental: la humana, visible y sometida al contradictorio, y la maquínica, invisible y orientada a modificar el procesamiento automatizado. Con base en estas categorías, se formula la barrera de integridad algorítmica como un obstáculo interno emergente para la sexta ola y se distingue la traducción semiótica legítima, que explicita y democratiza los sentidos, de la manipulación algorítmica clandestina, que oculta instrucciones y se apropia privadamente del canal tecnológico. La respuesta adecuada al fenómeno, concluye el estudio, requiere la construcción de una gobernanza basada en la ingestión segura de documentos, supervisión humana, transparencia, auditabilidad y responsabilidad proporcional, y no el abandono de la tecnología.

Palabras clave: Acceso a la Justicia. Inteligencia Artificial. Prompt Injection. Buena Fe Procesal. Traducción Semiótica.

1 INTRODUÇÃO

Durante décadas, a preocupação dos operadores do direito com a integridade do processo concentrou-se na autenticidade dos documentos, na licitude das provas e na veracidade das afirmações levadas a juízo. Essas preocupações partilhavam um pressuposto comum: o destinatário do conteúdo processual era sempre um ser humano, capaz de ler, interpretar e reagir àquilo que lhe era apresentado. A incorporação da inteligência artificial à rotina do Poder Judiciário rompeu esse pressuposto. Sistemas de triagem, classificação temática, sumarização de peças e apoio à elaboração de minutas passaram a operar em escala nacional, de modo que a compreensão institucional das demandas passou a apoiar-se, ainda que de forma auxiliar, em uma camada tecnológica de processamento textual. A inteligência artificial deixou de ser um instrumento externo, acionado quando conveniente, para tornar-se componente da infraestrutura cognitiva por meio da qual o tribunal organiza, filtra e apreende a informação.

Esse deslocamento tem uma consequência que a doutrina processual ainda não enfrentou com o devido cuidado. Se a infraestrutura tecnológica intermedeia a percepção do conteúdo processual, ela também passa a ser alvo de manipulação. A técnica conhecida como prompt injection, já estudada de forma extensa pela literatura de segurança da informação, consiste em inserir, num conteúdo aparentemente comum, instruções dissimuladas destinadas a subverter o comportamento de um sistema de inteligência artificial. Transposta para o processo judicial, a técnica assume contorno específico: a parte ou o advogado inserem, na peça, comandos invisíveis ao olhar humano, redigidos para serem lidos apenas pela máquina, com o propósito de condicionar o modo como o sistema classifica, resume ou avalia o documento. O leitor humano acredita estar diante de um documento; a inteligência artificial pode estar diante de um comando.

A imagem que dá título a este trabalho procura nomear essa novidade. Toda petição sempre teve leitores conhecidos: o juízo, a quem se pede a tutela, e a parte contrária, a quem se instaura o contraditório. Com a mediação algorítmica, surge um terceiro leitor, o sistema de inteligência artificial que interpreta o documento antes do humano, ou ao lado dele. Esse terceiro leitor não é sujeito processual, não titulariza direitos nem deveres e não decide juridicamente, mas lê, e essa condição basta para que possa ser enganado. O prompt injection é a técnica de dirigir-se a esse leitor às escondidas, por meio de instruções que os demais participantes do processo desconhecem.

O problema enfrentado por este artigo pode ser assim enunciado: a adoção da inteligência artificial pelo Judiciário amplia o acesso à justiça, ao reduzir custos cognitivos e acelerar a organização das demandas, mas essa mesma adoção cria uma barreira nova, de natureza distinta das barreiras econômicas, organizacionais e procedimentais historicamente descritas pela teoria do acesso à justiça.

Trata-se de uma barreira de integridade, situada não no acesso ao sistema, mas na confiabilidade da mediação algorítmica que o sistema realiza. A hipótese central sustenta que, quando sistemas artificiais passam a atuar como destinatários operacionais de documentos processuais, surgem vulnerabilidades capazes de comprometer o contraditório, a igualdade e a boa-fé, e que a inserção de comandos ocultos em petições é a manifestação concreta dessa barreira. A resposta adequada não é o bloqueio da tecnologia, mas a construção de uma governança fundada em ingestão segura de documentos, supervisão humana, transparência, auditabilidade e responsabilização proporcional.

O fenômeno justifica a investigação por sua atualidade. Ao longo do primeiro semestre de 2026, tribunais brasileiros passaram a identificar, em peças processuais, comandos ocultos dirigidos a sistemas de inteligência artificial. Episódios antes tratados como hipótese acadêmica converteram-se em realidade documentada, com decisões que reconheceram atos atentatórios à dignidade da justiça e aplicaram sanções por litigância de má-fé. O Conselho Nacional de Justiça, por seu Comitê Nacional de Inteligência Artificial do Judiciário, editou manifestação técnica específica sobre a mitigação desses riscos, e entidades da advocacia lançaram ferramentas de detecção. O tema deixou de ser especulação e passou a exigir enquadramento teórico rigoroso.

O trabalho está organizado em cinco seções. Após esta introdução, a segunda seção reconstrói o referencial teórico, articulando as ondas renovatórias do acesso à justiça, a teoria da comunicação processual, o instrumental da semiótica e a literatura técnica sobre prompt injection. A terceira seção explicita a metodologia. A quarta, dedicada aos resultados e à discussão, desenvolve as quatro contribuições originais e confronta a resposta institucional brasileira, analisando criticamente decisões e atos normativos. A quinta apresenta as considerações finais. Registre-se, por fim, que a pesquisa não pretende validar a fundamentação das decisões examinadas, mas utilizá-las como evidência documental de um fenômeno estrutural; por essa razão, elas são descritas de forma despersonalizada e submetidas ao mesmo rigor crítico aplicado à tecnologia.

2 REFERENCIAL TEÓRICO

O enquadramento teórico deste trabalho articula quatro campos que, isoladamente, não dão conta do fenômeno investigado, mas que, combinados, permitem compreendê-lo em suas dimensões institucional, comunicacional, ética e técnica. A teoria do acesso à justiça fornece o vocabulário para situar a automação judicial como etapa de um processo histórico de superação de obstáculos, e para discutir se essa etapa comporta, ou não, o estatuto de uma nova onda renovatória. A semiótica, e em particular a noção de tradução intersemiótica, oferece o instrumental para distinguir operações que ampliam a compreensão do ato processual daquelas que a distorcem. A dogmática processual da boa-

fé objetiva e do contraditório substancial permite qualificar juridicamente a conduta de quem manipula a mediação algorítmica. E a literatura técnica de segurança da informação descreve, com precisão que a doutrina jurídica ainda não incorporou, o funcionamento concreto do prompt injection. A seguir, cada um desses campos é apresentado na ordem em que sustenta o argumento central do artigo.

2.1 AS ONDAS RENOVATÓRIAS E A SEXTA ONDA TECNOLÓGICA

A teoria do acesso à justiça encontra seu marco fundador no Projeto de Florença, coordenado por Mauro Cappelletti e Bryant Garth na década de 1970, cujos resultados foram sistematizados na obra Acesso à Justiça, traduzida no Brasil por Ellen Gracie Northfleet e publicada em 1988. Os autores identificaram três movimentos de reforma, denominados ondas renovatórias, cada qual voltado a superar um tipo de obstáculo ao acesso efetivo. A primeira dirigiu-se ao obstáculo econômico e concentrou-se na assistência jurídica aos que não podem suportar os custos do processo. A segunda enfrentou o obstáculo organizacional, ligado à tutela dos interesses difusos e coletivos, e forçou a revisão de noções básicas como legitimação e coisa julgada. A terceira, à qual os autores deram o nome de enfoque do acesso à justiça, é a mais abrangente, e aqui cabe desfazer um equívoco corrente: ela não se reduz aos meios consensuais de solução de conflitos. A mediação, a conciliação e a arbitragem integram essa onda, mas não a esgotam. Cappelletti e Garth propuseram um enfoque amplo, que envolve reformas procedimentais, institucionais, preventivas e extrajudiciais, voltado ao conjunto de instituições, pessoas e procedimentos utilizados para processar e prevenir disputas.

As três primeiras ondas constituem formulação clássica e consolidada. As chamadas quarta, quinta e sexta ondas são desenvolvimentos doutrinários posteriores, cuja terminologia carece de uniformidade. Kim Economides, na segunda metade dos anos 1990, propôs uma quarta onda voltada à dimensão ética das profissões jurídicas e ao acesso dos próprios operadores do direito à justiça e à sua formação. Parte da doutrina situa a tecnologia como quarta onda, associando-a ao processo eletrônico; outra parte a situa como sexta, reservando-lhe especificamente a inteligência artificial. Diante dessa falta de uniformidade, adota-se aqui uma sistematização funcional, baseada não na numeração, mas nos obstáculos enfrentados. Sob essa perspectiva, a transformação tecnológica do acesso à justiça pode ser lida como sequência de três camadas. A digitalização reduz barreiras físicas, mas cria a exclusão digital e exige o que se pode chamar de acesso cognitivo. A transnacionalização projeta a tutela para além das fronteiras estatais. A automação inteligente introduz a inteligência artificial, a jurimetria e os modelos generativos na atividade jurisdicional.

É a essa última camada, aqui designada sexta onda, que o presente estudo se refere, com a ressalva expressa de que o termo não corresponde a consenso terminológico consolidado. O que a

caracteriza não é o uso de ferramentas digitais em si, mas a automação de tarefas que antes exigiam cognição humana: classificar, resumir, recuperar e correlacionar informações. Sua correspondência institucional é visível no plano internacional. A Organização para a Cooperação e Desenvolvimento Econômico, no documento sobre governança com inteligência artificial publicado em 2025, sustenta que as tecnologias digitais e os dados têm potencial significativo para apoiar o acesso à justiça e para sustentar a resiliência, a eficiência, a efetividade e a equidade dos sistemas, desde que submetidos a governança adequada. A ressalva é decisiva: o ganho não é incondicional, depende de uma arquitetura de controle. Essa condicionalidade é o que o presente trabalho procura desenvolver, ao sustentar que a integridade da mediação algorítmica é pressuposto para que a sexta onda cumpra sua promessa de ampliação do acesso.

2.2 COMUNICAÇÃO PROCESSUAL, SEMIÓTICA E TRADUÇÃO INTERSEMIÓTICA

O processo é, em sua essência, um fenômeno comunicativo, e compreendê-lo exige instrumental que ultrapassa a dogmática e recorre à semiótica, ciência que estuda os signos e os processos de significação. Charles Sanders Peirce (2005), um dos fundadores da disciplina, concebeu o signo como algo que representa uma coisa para alguém sob determinado aspecto, numa relação triádica entre o representamen, o objeto e o interpretante. O interpretante, nessa arquitetura, não é o intérprete, mas o efeito de sentido que o signo produz. A distinção importa para este estudo porque permite pensar a interpretação como processo que pode ocorrer em diferentes suportes, inclusive não humanos, sem que se atribua a esses suportes a condição de sujeito.

Roman Jakobson (2010), em ensaio clássico sobre os aspectos linguísticos da tradução, distinguiu três modalidades. A intralingual reformula signos verbais por meio de outros signos da mesma língua. A interlingual verte signos verbais de uma língua para outra. A intersemiótica, ou transmutação, interpreta signos verbais por meio de sistemas de signos não verbais, e é essa terceira modalidade que interessa de perto ao direito contemporâneo. A tradução semiótica dos atos judiciais, entendida como esforço de ampliar a equivalência comunicacional entre o conteúdo jurídico e a compreensão do destinatário, opera na fronteira entre o verbal e o visual, entre o técnico e o acessível. Traduzir uma decisão para linguagem compreensível, valer-se de recursos visuais que expliquem o dispositivo, reformular o pedido em termos inteligíveis: tudo isso é atividade de tradução no sentido jakobsoniano, e tudo isso reduz a assimetria entre o Estado-juiz e o cidadão.

Esse movimento tem tradução normativa no Brasil. A Recomendação CNJ 144, de 25 de agosto de 2023, recomendou aos tribunais o uso de linguagem simples, clara e acessível, com o emprego, sempre que possível, de elementos visuais que facilitem a compreensão, e o Pacto Nacional do

Judiciário pela Linguagem Simples consolidou eixos de simplificação da comunicação com o cidadão. A linguagem simples, o Visual Law, a acessibilidade digital e o desenho centrado no usuário manifestam uma mesma aspiração: que o acesso não se esgote na possibilidade formal de peticionar, mas se realize na inteligibilidade efetiva da experiência processual.

Aqui está um ponto teórico central deste artigo. A tradução semiótica legítima é aquela que torna o conteúdo visível, explicita sentidos e se submete ao controle intersubjetivo de todos os participantes. Ela democratiza a comunicação porque amplia o número de sujeitos capazes de compreender e reagir ao ato. É por contraste com esse pano de fundo que o prompt injection revela sua natureza: opera em sentido inverso, oculta em vez de explicitar, apropria-se privadamente do canal em vez de democratizá-lo.

2.3 BOA-FÉ OBJETIVA E A DIMENSÃO SUBSTANCIAL DO CONTRADITÓRIO

A resposta jurídica ao prompt injection encontra fundamento primeiro na boa-fé objetiva processual. O artigo 5º do Código de Processo Civil dispõe que aquele que de qualquer forma participa do processo deve comportar-se de acordo com a boa-fé. É preciso afastar, de plano, a leitura que reduz a boa-fé à intenção psicológica do agente. A doutrina processual brasileira, na esteira de Fredie Didier Jr (2026)., sustenta que o artigo 5º consagra a boa-fé objetiva, um padrão normativo de comportamento exigível, aferível pela conduta externa e independente das intenções. Essa raiz processual comunica-se com a tradição civilista de Judith Martins-Costa (2024) e de António Menezes Cordeiro (2011), que desenvolveram a compreensão da boa-fé como cláusula geral capaz de impor deveres de conduta e de permitir o controle do exercício de posições jurídicas. Transposta ao processo, a boa-fé objetiva dispensa a enumeração exaustiva das condutas desleais, porque a infinidade de situações possíveis torna ineficaz qualquer catálogo fechado. Do artigo 6º deriva o dever de cooperação, e dele os deveres de lealdade, transparência e proteção da confiança, além da vedação ao comportamento ardiloso e ao abuso de posições processuais.

A compreensão contemporânea do contraditório reforça esse fundamento e permite localizar com precisão a nocividade do prompt injection. A garantia do contraditório, prevista no artigo 5º, inciso LV, da Constituição, deixou de ser lida em chave meramente formal. A doutrina passou a reconhecer no contraditório uma dimensão formal, consistente na participação das partes, e uma dimensão substancial, correspondente ao direito de influenciar o convencimento do julgador. Didier Jr. observa que não basta permitir que a parte participe do processo; é indispensável assegurar-lhe possibilidade efetiva de influir no conteúdo da decisão. Cândido Rangel Dinamarco (2023), na mesma linha, ensina

que o contraditório tem dupla destinação: dirige-se às partes, para assegurar-lhes participação efetiva, e ao magistrado, que deve construir a decisão a partir do diálogo processual.

Dessa evolução extrai-se um aspecto que a literatura sobre inteligência artificial no processo costuma ignorar. O contraditório não protege qualquer forma de influência, apenas aquela exercida por argumentos conhecidos, submetidos ao debate e passíveis de impugnação. Trata-se de influência transparente, dialógica, construída em ambiente de igualdade. É exatamente esse pressuposto de publicidade que o prompt injection viola: opera pela invisibilidade e escapa por completo ao debate.

2.4 A TÉCNICA DO PROMPT INJECTION NA LITERATURA DE SEGURANÇA DA INFORMAÇÃO

A expressão prompt injection designa uma classe de ataques contra modelos de linguagem, na qual um agente insere instruções capazes de fazer o modelo desviar-se das orientações legítimas que recebeu e passar a executar comandos alinhados aos interesses do atacante. A entidade OWASP, em seu Top 10 for Large Language Model Applications, classifica o prompt injection como o risco de número um, identificado pela sigla LLM01. Um aspecto técnico dessa definição tem consequência jurídica direta: segundo a OWASP, a vulnerabilidade ocorre quando entradas alteram o comportamento ou a saída do modelo de maneiras não pretendidas, e essas entradas podem afetar o modelo ainda que sejam imperceptíveis a humanos, bastando que o conteúdo seja processado pela máquina. A injeção de comandos não precisa, portanto, ser humanamente visível ou legível. Essa observação separa aquilo que o humano vê daquilo que a máquina lê, separação que o direito ainda precisa apreender com rigor.

A literatura distingue duas modalidades. A injeção direta ocorre quando o próprio usuário insere as instruções na interface do sistema. A injeção indireta ocorre quando o modelo processa conteúdo proveniente de fontes externas, como documentos, páginas e arquivos, e nesse conteúdo estão embutidas as instruções. Kai Greshake e coautores, em estudo publicado nos anais do 16th ACM Workshop on Artificial Intelligence and Security em 2023, demonstraram que as aplicações integradas a modelos de linguagem borram a fronteira entre dado e instrução, de modo que adversários podem explorá-las remotamente ao injetar comandos em dados que provavelmente serão recuperados. O National Institute of Standards and Technology, no documento sobre aprendizado de máquina adversarial de referência AI 100-2e2025, incorporou a injeção direta e indireta de comandos à sua taxonomia e reconheceu que as mitigações disponíveis são, em grande parte, empíricas, sem garantia teórica de prevenção completa. O prompt injection não se confunde com o jailbreak: neste, é o próprio usuário quem tenta persuadir a inteligência artificial a desrespeitar suas regras para obter respostas

normalmente bloqueadas; naquele, a manipulação está embutida em conteúdo externo, muitas vezes sem conhecimento do usuário que submete o documento à análise.

O ataque por documento processual é caso paradigmático de injeção indireta. A petição não é digitada na interface pelo atacante, mas juntada aos autos e processada posteriormente pela ferramenta institucional, e as instruções maliciosas viajam embutidas no próprio arquivo. As técnicas de ocultação são variadas e de baixa barreira de entrada. A manifestação técnica do Comitê Nacional de Inteligência Artificial do Judiciário cataloga, entre os vetores conhecidos, o texto invisível ou de baixíssimo contraste, a fonte diminuta, os caracteres de largura zero, as sobreposições, os metadados de arquivos, os comentários, as instruções fora da área visível da página, o conteúdo inserido em legendas de imagens e o texto extraído por reconhecimento óptico de caracteres induzido a erro. O traço comum a esses vetores é a divergência deliberada entre o que o humano vê e o que a máquina lê. Nisso o prompt injection difere estruturalmente da falsidade documental clássica. Na falsidade tradicional, o conteúdo visível é alterado e pode, em tese, ser detectado por inspeção humana; no prompt injection, o conteúdo visível permanece íntegro, e a manipulação ocorre na camada de processamento automatizado, invisível ao leitor no fluxo normal de trabalho.

3 METODOLOGIA

A pesquisa é de natureza teórica e qualitativa e adota o método hipotético-dedutivo. Parte da hipótese enunciada na introdução, segundo a qual a incorporação da inteligência artificial à infraestrutura cognitiva do Judiciário cria uma barreira de integridade cuja manifestação é o prompt injection, e submete essa hipótese a teste mediante confronto com o referencial teórico e com o material empírico disponível. Segundo Marconi e Lakatos (2003), o método hipotético-dedutivo parte da formulação de um problema, avança pela proposição de conjecturas que o expliquem provisoriamente e submete essas conjecturas a tentativas de falseamento, na esteira do critério proposto por Popper. É esse percurso que a presente investigação reproduz: a hipótese central funciona como conjectura a ser confrontada com a doutrina, os atos normativos e as decisões judiciais que compõem o corpus empírico, e não como conclusão antecipada que a pesquisa se limita a ilustrar. A técnica predominante é a análise documental, combinada com revisão bibliográfica de fontes jurídicas e de literatura técnica de segurança da informação.

O corpus empírico reúne três conjuntos de documentos. O primeiro compreende decisões judiciais brasileiras proferidas em 2026 que identificaram e sancionaram o uso de prompt injection em peças processuais. Essas decisões são examinadas de forma despersonalizada, sem identificação do tribunal, da vara, do magistrado, das partes ou dos advogados envolvidos, uma vez que ao estudo

interessa o fenômeno, não a individualização dos casos. O segundo conjunto abrange atos e documentos normativos e técnicos do Conselho Nacional de Justiça, com destaque para a Resolução CNJ 615 de 2025 e para a manifestação técnica de seu Comitê Nacional de Inteligência Artificial do Judiciário sobre a mitigação de riscos de prompt injection. O terceiro conjunto reúne a literatura técnica internacional sobre a matéria, incluindo os documentos da OWASP e do National Institute of Standards and Technology e a produção acadêmica correlata.

O tratamento do material segue três operações. A primeira é a descrição despersonalizada das ocorrências, destinada a preservar o foco no fenômeno e a evitar exposição indevida de sujeitos processuais. A segunda é a análise crítica dos enquadramentos jurídicos adotados nas decisões examinadas, submetendo-os ao mesmo rigor aplicado à tecnologia; a pesquisa não depende, portanto, da correção integral desses enquadramentos para sustentar suas conclusões. A terceira é a articulação teórica propriamente dita, na qual material empírico e literatura técnica são mobilizados para sustentar as quatro contribuições desenvolvidas na seção seguinte.

Duas limitações merecem registro. A recente emergência do fenômeno impede análise estatística de larga escala, de modo que as conclusões aqui apresentadas têm caráter teórico e propositivo, a confirmar por pesquisa empírica futura. E a natureza despersonalizada do corpus, embora necessária por razões éticas, restringe a possibilidade de rastrear desdobramentos processuais posteriores às decisões examinadas, como eventuais recursos ou reformas

4 RESULTADOS E DISCUSSÃO

As quatro contribuições apresentadas nesta seção não constituem blocos isolados, mas etapas de um mesmo raciocínio, construído em ordem de dependência lógica. A caracterização da inteligência artificial como terceiro leitor e a identificação da dupla camada documental, tratadas em 4.1, fornecem o vocabulário conceitual sem o qual as duas contribuições seguintes perderiam precisão. A formulação da barreira de integridade algorítmica, em 4.2, situa esse vocabulário no interior da teoria do acesso à justiça, e a distinção entre tradução semiótica legítima e manipulação algorítmica clandestina, em 4.3, fornece o critério normativo que permite separar, na prática, a persuasão processual lícita da fraude. As seções 4.4 e 4.5 testam essas categorias contra o material empírico: a primeira, ao confrontar decisões judiciais brasileiras de 2026, e a segunda, ao examinar a resposta regulatória do Conselho Nacional de Justiça. O percurso, portanto, vai do conceito à norma e da norma à prática, e cada seção pressupõe a anterior.

4.1 O TERCEIRO LEITOR E A DUPLA CAMADA DOCUMENTAL

A primeira contribuição do trabalho consiste em caracterizar a inteligência artificial como destinatária operacional intermediária da comunicação processual. A petição tradicional dirige-se ao órgão jurisdicional, para provocar a tutela, e à parte contrária, para instaurar o contraditório, além de alcançar os demais sujeitos que dela devam tomar conhecimento. Essa estrutura pressupunha destinatários humanos. Com a incorporação da inteligência artificial, o sistema que classifica a peça, gera seu resumo, recupera informações ou auxilia na compreensão dos autos passa a receber e processar o documento antes que o humano o faça, ou ao lado dele. Esse sistema não é sujeito processual: não titulariza direitos nem deveres, não ocupa posição na relação jurídica processual e não decide juridicamente. Atribuir-lhe subjetividade seria antropomorfização indevida.

A categoria proposta é mais modesta e mais precisa. A inteligência artificial é destinatária porque recebe o documento; é operacional porque o processa tecnicamente; é intermediária porque se interpõe entre o documento e a percepção humana do decisor. Chamá-la de terceiro leitor busca nomear precisamente essa posição: um leitor que interpreta sem julgar, que produz interpretantes sem ser sujeito.

A relevância da categoria está em suas consequências práticas. Se a ferramenta intermediária produz um resumo enviesado ou uma classificação incorreta, ela condiciona a atenção do decisor, ainda que este mantenha a palavra final sobre o mérito. Um magistrado que, diante de elevado volume de processos, utiliza a triagem automatizada para hierarquizar sua análise, opera sobre um substrato que a máquina ajudou a construir. Se esse substrato foi manipulado, a supervisão humana permanece formalmente preservada, mas fica materialmente esvaziada: o supervisor não controla um viés cuja existência desconhece.

A segunda contribuição decorre da primeira e consiste em identificar a existência de uma dupla camada documental. A inserção de instruções ocultas produz um efeito que a teoria processual ainda não havia enfrentado com precisão: a duplicação do documento em duas camadas de leitura. A camada humana é visível, composta pelo texto que o juiz e a parte contrária efetivamente leem, e submete-se ao contraditório. A camada maquínica é invisível ao olhar humano, composta por comandos redigidos para serem lidos e processados pelo sistema de inteligência artificial, com o propósito de modificar o processamento automatizado. Há, em rigor, dois documentos num só suporte, um para os humanos e outro para a máquina, e apenas o primeiro está sujeito ao crivo do contraditório.

A camada maquínica configura, nesses termos, um canal privado aberto no interior de um ato processual público. É essa duplicação, e não a mera existência de tecnologia de apoio à decisão, que estabelece a assimetria comunicacional incompatível com o processo democrático: o documento

ostensivamente submetido ao juiz e à parte adversa deixa de coincidir com o conteúdo efetivamente dirigido à infraestrutura tecnológica que auxilia a decisão.

4.2 A BARREIRA DE INTEGRIDADE ALGORÍTMICA

A terceira contribuição consiste em formular a categoria da barreira de integridade algorítmica: a barreira que surge quando o acesso à infraestrutura de inteligência artificial do Judiciário, ou sua utilização, ocorre em condições assimétricas, clandestinas ou manipulativas, comprometendo a igualdade processual e a confiabilidade da mediação tecnológica. Diferentemente das barreiras econômicas ou digitais, situadas no acesso ao sistema, essa barreira é interna: localiza-se no interior da própria mediação que a tecnologia realiza.

Cabe justificar por que essa categoria não é apresentada como sétima onda renovatória. A tentação de proclamar nova onda a cada inovação tecnológica merece resistência. Para Cappelletti e Garth (1988), uma onda corresponde a um conjunto autônomo e historicamente identificável de reformas voltadas a superar um tipo específico de obstáculo, e um único fenômeno, ainda que relevante, não sustenta essa proclamação. A integridade informacional funciona, isto sim, como condição de possibilidade da sexta onda: sem ela, a automação não amplia o acesso, apenas desloca a corrupção do processo para um plano menos visível.

A nocividade dessa barreira revela-se num feixe de cinco lesões, que atingem pilares distintos do processo justo. A primeira é a lesão à igualdade de armas: a parte que insere comandos ocultos abre para si um canal de persuasão não declarado, indisponível à parte contrária, rompendo a paridade que o processo pressupõe. A segunda atinge o contraditório em sua dimensão substancial, porque a contraparte só pode reagir àquilo que conhece, e a camada maquínica é, por definição, o que ela desconhece; o contraditório converte-se, assim, em ilusão parcial, incidente apenas sobre a camada visível do documento. A terceira lesão recai sobre a autonomia decisória: a instrução oculta interfere no modo como a ferramenta intermediária organiza as informações que chegam ao decisor, contaminando o substrato de sua análise. A quarta compromete a confiança institucional, pois a generalização dessas práticas produziria suspeita permanente sobre a higidez dos autos.

A quinta lesão é a mais grave sob a perspectiva do acesso à justiça, e merece destaque em parágrafo próprio. Trata-se do agravamento das desigualdades entre litigantes. Quem domina a técnica pode explorar vulnerabilidades dos sistemas contra adversários vulneráveis, contra defensores com pouca infraestrutura e contra pequenos escritórios que não dispõem de conhecimento técnico equivalente. Nesse ponto, o prompt injection deixa de ser fraude documental comum e passa a

configurar problema de acesso à justiça em sentido próprio, porque transfere para o plano algorítmico a mesma assimetria de poder que as três ondas renovatórias clássicas se propuseram a combater.

4.3 TRADUÇÃO SEMIÓTICA LEGÍTIMA E MANIPULAÇÃO ALGORÍTMICA CLANDESTINA

A quarta contribuição consiste em distinguir a tradução semiótica legítima da manipulação algorítmica clandestina. As duas operações trabalham sobre o mesmo material, o signo jurídico, e ambas buscam produzir efeitos de sentido em um destinatário, mas seguem direções opostas. A tradução semiótica legítima, tal como descrita no referencial teórico, torna o conteúdo visível, explicita estruturas, reduz assimetrias e submete-se ao controle intersubjetivo de todos os participantes. É operação pública, dialógica e reversível, cujo propósito é ampliar o número de sujeitos capazes de compreender o ato processual. O prompt injection inverte cada um desses traços: oculta o conteúdo em vez de explicitá-lo, cria assimetria em vez de reduzi-la, orienta um leitor artificial às escondidas e busca alterar o resultado interpretativo escapando ao contraditório.

A síntese conceitual pode ser formulada em termos diretos. A tradução semiótica democratiza a comunicação processual; o prompt injection privatiza clandestinamente uma camada dessa comunicação. Essa oposição não é apenas descritiva, ela oferece um critério operativo para separar persuasão legítima de manipulação ilícita, distinção que a doutrina processual ainda não havia formulado com essa precisão. Toda petição busca persuadir. O advogado que narra os fatos de modo favorável ao cliente, seleciona os precedentes mais convenientes e enfatiza a urgência de sua pretensão exerce persuasão legítima, e ninguém questiona essa prática. O problema do prompt injection não está, portanto, na persuasão em si, mas no seu modo clandestino e assimétrico de operar.

A diferença entre as duas condutas não é de grau, é de natureza. A persuasão legítima dirige-se a destinatários que a percebem e podem contra-argumentar. A manipulação clandestina dirige-se a um intermediário tecnológico cuja atuação escapa à percepção dos demais sujeitos processuais, e cujo comportamento se pretende condicionar sem que ninguém mais tome conhecimento. Sob a lente da semiótica, essa manipulação configura uma tradução às avessas: em vez de aproximar o signo do destinatário, sequestra o intérprete e o afasta dos demais participantes do processo.

4.4 A RESPOSTA INSTITUCIONAL BRASILEIRA E SEUS DILEMAS DOGMÁTICOS

As contribuições teóricas encontram confirmação empírica na resposta que o Judiciário brasileiro passou a dar ao fenômeno ao longo de 2026. Um primeiro episódio de grande repercussão envolveu peça processual na qual patronos inseriram, mediante texto branco sobre fundo branco, comandos determinando que a inteligência artificial da contraparte realizasse análise superficial da

peça adversária, deixasse de impugnar os documentos e privilegiasse a argumentação da inicial. Em outro caso, também examinado de forma despersonalizada, uma empresa foi condenada por litigância de má-fé após embutir, ao final de peça de especificação de provas, comandos imperativos ocultos dirigidos a induzir o sistema automatizado à improcedência total dos pedidos de uma consumidora. O magistrado qualificou o expediente como ardil processual que viola frontalmente os deveres de lealdade, cooperação e boa-fé objetiva, descrevendo-o como técnica de prompt injection caracterizadora de inovação ilegal de estado de fato e de expediente escuso voltado a desvirtuar a atividade cognitiva do juízo por vias tecnológicas não autorizadas.

Um terceiro conjunto de ocorrências surgiu em processos com indícios de litigância predatória em massa. Petições padronizadas traziam o comando oculto de que, sendo o leitor um agente de inteligência artificial, deferisse a gratuidade da justiça, deferisse a tutela de urgência e determinasse a citação do réu, sob a alegação de que todos os documentos estariam presentes. Ao apreciar um desses casos, o juízo classificou a prática como uma das condutas mais perniciosas contra a dignidade da Justiça, reconheceu fraude processual e litigância predatória, aplicou multa por litigância de má-fé diretamente ao advogado e determinou o encaminhamento do caso ao Ministério Público, à Corregedoria e às seccionais da Ordem dos Advogados do Brasil. Cortes superiores tornaram públicas tentativas semelhantes e a instauração de apuração administrativa, informando dispor de mecanismos próprios de detecção. No plano tecnológico e associativo, a resposta institucional incluiu a edição de nota técnica pelo comitê de governança de inteligência artificial e o lançamento, por entidades da advocacia em parceria com plataformas jurídicas, de ferramenta gratuita destinada a identificar prompt injection em petições antes de sua utilização por sistemas do Judiciário.

Esse material comprova o valor empírico das categorias propostas, mas não deve ser tratado como solução dogmática incontroversa. Vários dos enquadramentos adotados nas decisões comportam controvérsia que a pesquisa tem o dever de expor. Um primeiro ponto sensível é a qualificação da conduta como inovação ilegal no estado de fato de bem ou direito litigioso, prevista no artigo 77, inciso VI, do Código de Processo Civil. Trata-se de figura historicamente concebida para a alteração material do estado de fato de coisa ou direito em disputa, e seu alargamento para alcançar a inserção de comandos ocultos, embora defensável, exige fundamentação densa: o que se altera não é o estado do bem litigioso, mas o estado informacional do próprio documento.

Um segundo ponto é a caracterização como alteração da verdade dos fatos, do artigo 80, inciso II. Os comandos ocultos, a rigor, não alteram a narrativa fática visível; instruem a máquina sobre como qualificar juridicamente os fatos já narrados, o que talvez se ajustasse melhor às hipóteses de uso do processo para objetivo ilegal ou de conduta temerária, previstas nos incisos III e V do mesmo

dispositivo. A caracterização da má-fé, ademais, reclama demonstração do elemento subjetivo; não é possível presumir ilicitude apenas pela existência de comandos dirigidos a sistemas de inteligência artificial.

Um terceiro ponto, talvez o mais delicado, diz respeito à individualização da responsabilidade. A técnica de prompt injection é de domínio tecnológico específico, e é plausível que seja executada pelos advogados sem conhecimento do cliente. Punir a parte por conduta que pode ter sido exclusivamente do patrono, sobretudo com o percentual máximo da multa, suscita dúvida quanto à individualização. A jurisprudência consolidada do Superior Tribunal de Justiça sustenta que as penas por litigância de má-fé são endereçadas às partes e não podem ser estendidas ao advogado, que responde em ação própria, na forma do artigo 32 da Lei 8.906 de 1994. Essa orientação convive, porém, com a perplexidade de que, em casos de litigância predatória, algumas decisões impuseram a multa diretamente ao advogado, o que evidencia a tensão do regime vigente diante de um ilícito tecnicamente praticado por quem domina a técnica. A essa tensão somam-se a exigência de motivação proporcional para a fixação da multa no teto legal, especialmente quando o dano concreto não se materializou, e a necessidade de contraditório prévio à sanção, sob pena de violação da vedação à decisão surpresa do artigo 10 do Código de Processo Civil.

O artigo perderia força científica se tratasse esses enquadramentos como perfeitos. Ganha, ao contrário, ao reconhecer sua importância institucional e submetê-los ao mesmo rigor crítico aplicado à tecnologia que os originou.

4.5 DA PROIBIÇÃO À GOVERNANÇA DA INTEGRIDADE

Resta enfrentar a tentação de responder ao fenômeno com o bloqueio da inteligência artificial. A reação seria inadequada por razões concretas. O bloqueio confundiria vulnerabilidade com inutilidade, eliminaria benefícios reais de organização e compreensão de autos extensos, não impediria o uso de ferramentas externas por partes e advogados, e apenas retiraria do Judiciário a possibilidade de governar institucionalmente a tecnologia. O resultado prático seria ampliar a própria assimetria que se pretendia conter. A questão relevante desloca-se, assim, de saber se a inteligência artificial deve ser usada para saber como assegurar a integridade de sua mediação.

A manipulação clandestina comporta-se, nesse terreno, como alvo móvel: cada salvaguarda conhecida convida à engenhosidade de novos ardis. A governança da integridade não se esgota, por isso, em um conjunto fixo de medidas a implementar de uma vez; ela exige acompanhamento contínuo e capacidade de antecipar manobras ainda não catalogadas.

O marco normativo brasileiro já oferece a moldura desse processo. A Resolução CNJ 615 de 2025 não adota proibição geral, mas política condicionada por participação e supervisão humana em todas as etapas, segurança, transparência, curadoria de fontes rastreáveis e auditáveis, avaliação de impacto e monitoramento contínuo, vedando soluções que não possibilitem a revisão humana dos resultados. Sobre essa moldura, a manifestação técnica do Comitê Nacional de Inteligência Artificial do Judiciário desenvolve uma arquitetura de defesa cujo eixo é decisivo para a tese aqui sustentada: no ambiente judicial, a segurança começa na ingestão segura dos documentos, e não apenas no controle da resposta. O texto extraído das peças, incluindo imagens e metadados, deve ser tratado como dado externo potencialmente não confiável, preservando-se atributos como tamanho e cor da fonte, contraste, posição e camada. Essa preservação permite distinguir o texto canônico visível, base legítima para sumarização e classificação, do conteúdo suspeito ou anômalo, como a fonte diminuta, o texto branco sobre fundo branco e as camadas ocultas. Esse conteúdo suspeito não deve ser descartado em silêncio, mas segregado, preservado e registrado em trilha de auditoria, para permitir revisão humana e eventual apuração de responsabilidades.

A filtragem determinística de saída complementa a ingestão segura, mas não a substitui: a própria manifestação técnica reconhece que ela é necessária e insuficiente, porque a manipulação semântica pode produzir um resumo aparentemente imparcial e, ainda assim, contaminado. O controle da entrada e a supervisão humana informada permanecem, por essa razão, insubstituíveis.

Dessa arquitetura extrai-se uma medida de especial interesse para o direito processual: a certificação nos autos da detecção de conteúdo adversarial. Identificada pelo sistema a presença de instrução oculta, recomenda-se que a ocorrência seja certificada de forma clara e objetiva, com indicação da peça, da natureza do conteúdo detectado e da data da detecção, preservada a autonomia do magistrado quanto às providências cabíveis. A certificação tem o mérito de reconverter em público aquilo que a manipulação tornara oculto, restaurando a simetria comunicacional e devolvendo à camada maquínica a visibilidade sem a qual não há contraditório. Nessa medida, ela é a resposta processual mais fiel ao diagnóstico teórico aqui proposto: se a nocividade do prompt injection está na privatização clandestina de uma camada da comunicação, o remédio começa por trazê-la de volta à luz.

5 CONSIDERAÇÕES FINAIS

A investigação confirma a hipótese formulada na introdução. A inteligência artificial amplia o acesso à justiça ao reduzir custos cognitivos, acelerar a organização das informações e melhorar a capacidade institucional de análise, integrando, nesse sentido, a sexta onda renovatória. Ao converter-

se em componente da infraestrutura cognitiva do Judiciário, contudo, ela cria uma vulnerabilidade nova: a exposição da mediação algorítmica a comunicações ocultas. O prompt injection em petições é a manifestação concreta dessa vulnerabilidade, e viola a boa-fé objetiva, a lealdade, o contraditório em sua dimensão substancial e a igualdade processual. A resposta adequada passa pela construção de governança fundada em ingestão segura de documentos, supervisão humana informada, transparência, auditabilidade e responsabilização proporcional, e não pelo bloqueio da tecnologia. A integridade algorítmica não constitui nova onda; funciona como pressuposto de possibilidade da sexta.

As quatro contribuições articulam-se em um mesmo argumento. A caracterização da inteligência artificial como terceiro leitor, destinatária operacional intermediária sem subjetividade, explica por que a manipulação da máquina afeta a comunicação entre humanos. A identificação da dupla camada documental esclarece a mecânica dessa assimetria, ao mostrar que o documento se cinde entre uma face pública e uma face clandestina. A formulação da barreira de integridade algorítmica situa o problema no interior da teoria do acesso à justiça, como obstáculo emergente da sexta onda. A distinção entre tradução semiótica legítima e manipulação algorítmica clandestina fornece o critério que separa a persuasão devida da fraude, ao contrastar a operação que ilumina o sentido com a que o sequestra. É nesta última contribuição que a pesquisa dialoga mais diretamente com a agenda da tradução semiótica dos atos judiciais como vetor de acesso à justiça: o prompt injection é o avesso dessa tradução, evidência, pela via negativa, de que a comunicação processual só é legítima quando pública.

A pesquisa sugere, como agenda institucional, cinco linhas de trabalho. A primeira é o aperfeiçoamento da regulamentação para disciplinar especificamente a integridade do conteúdo processado pelos sistemas, e não apenas as salvaguardas da resposta gerada. A segunda é a consolidação da certificação nos autos como prática processual capaz de restaurar a publicidade da camada oculta. A terceira é a revisão do regime sancionatório, de modo a distribuir com mais precisão a responsabilidade entre parte e advogado. A quarta é o fortalecimento da transparência quanto ao uso da inteligência artificial pelos órgãos julgadores. A quinta é a incorporação de formação ética e tecnológica na educação jurídica, de modo que operadores do direito compreendam tanto o potencial quanto os riscos da automação que já integra sua prática.

O estudo apresenta limitações que devem ser reconhecidas. A taxonomia das ondas renovatórias permanece controvertida na doutrina, e foi adotada aqui como opção funcional, não como classificação consensual. A noção de destinatária operacional exige vigilância permanente contra o deslize antropomórfico, sob pena de a categoria perder a precisão que a justifica. A barreira de integridade algorítmica demanda demonstração empírica ulterior, capaz de mensurar a frequência e o impacto real

das manipulações identificadas. E a análise das decisões judiciais, deliberadamente críticas, trata-as como sintoma de um problema estrutural, não como modelo dogmático acabado, o que significa que revisões e reformas dessas decisões em instâncias superiores não invalidaram as categorias propostas neste artigo.

A integridade da mediação algorítmica é, por fim, condição para que a promessa de acesso à justiça sobreviva à sua própria automação. Assegurá-la depende de esforço conjunto entre regulação, prática processual e formação de operadores do direito, e essa tarefa permanece em curso.

REFERÊNCIAS

BRASIL. Conselho Nacional de Justiça. Comitê Nacional de Inteligência Artificial do Judiciário. *Manifestação Técnica CNIAJ 1/2026*: injeção de comandos: considerações técnicas sobre mitigação de riscos de prompt injection em sistemas judiciais de inteligência artificial. Brasília, DF: CNJ, 27 maio 2026.

BRASIL. Conselho Nacional de Justiça. Recomendação n. 144, de 25 de agosto de 2023. Recomenda aos Tribunais que implementem o uso da linguagem simples nas comunicações e atos que editem. Brasília, DF: CNJ, 2023.

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 inteligência artificial no Poder Judiciário. Brasília, DF: CNJ, 2020.

BRASIL. Conselho Nacional de Justiça. Resolução n. 615, de 11 de março de 2025. Estabelece diretrizes para o desenvolvimento, a utilização e a governança de soluções desenvolvidas com recursos de inteligência artificial no Poder Judiciário. Brasília, DF: CNJ, 2025.

BRASIL. *Constituição da República Federativa do Brasil de 1988*. Brasília, DF: Presidência da República, 1988.

BRASIL. Lei n. 8.906, de 4 de julho de 1994. Dispõe sobre o Estatuto da Advocacia e a Ordem dos Advogados do Brasil (OAB). *Diário Oficial da União*: seção 1, Brasília, DF, 5 jul. 1994.

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

BRASIL. Superior Tribunal de Justiça. AgInt no AREsp 1.722.332/MT, Quarta Turma, julgado em 13 jun. 2022.

CAPPELLETTI, Mauro; GARTH, Bryant. *Acesso à justiça*. Tradução de Ellen Gracie Northfleet. Porto Alegre: Sérgio Antonio Fabris Editor, 1988.

DIDIER JR., Fredie. *Curso de direito processual civil*: introdução ao direito processual civil, parte geral e processo de conhecimento. 28. ed. Salvador: JusPodivm, 2026.

DINAMARCO, Cândido Rangel. Instituições de direito processual civil. 9. ed. São Paulo: Malheiros, 2023.

ECONOMIDES, Kim. Lendo as ondas do "Movimento de Acesso à Justiça": epistemologia versus metodologia? In: PANDOLFI, Dulce; CARVALHO, José Murilo de; CARNEIRO, Leandro Piquet; GRYNSZPAN, Mario (Orgs.). Cidadania, justiça e violência. Rio de Janeiro: Editora Fundação Getulio Vargas, 1999. p. 61-76.

GRESHAKE, Kai *et al*. Not what you've signed up for: compromising real-world LLM-integrated applications with indirect prompt injection. *In*: PROCEEDINGS OF THE 16TH ACM WORKSHOP ON ARTIFICIAL INTELLIGENCE AND SECURITY (AISec '23), 2023. New York: ACM, 2023.

JAKOBSON, Roman. Aspectos linguísticos da tradução. ***In***: JAKOBSON, Roman. Linguística e comunicação. Tradução de Izidoro Blikstein e José Paulo Paes. 24. ed. São Paulo: Cultrix, 2010. p. 63-72.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. *Fundamentos de metodologia científica*. 5. ed. São Paulo: Atlas, 2003.

MARTINS-COSTA, Judith. A boa-fé no direito privado: critérios para a sua aplicação. 3. ed. São Paulo: SaraivaJur, 2024.

MENEZES CORDEIRO, António Manuel da Rocha e. Da boa fé no direito civil. 4. reimpr. Coimbra: Almedina, 2011.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST). *Adversarial machine learning*: a taxonomy and terminology of attacks and mitigations. NIST AI 100-2e2025. Gaithersburg: NIST, 2025.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD). *Governing with artificial intelligence*: the state of play and way forward in core government functions. Paris: OECD Publishing, 2025.

OWASP GEN AI SECURITY PROJECT. *OWASP Top 10 for large language model applications*. LLM01: prompt injection. 2025.

PEIRCE, Charles Sanders. Semiótica. Tradução de José Teixeira Coelho Neto. 3. ed. 2. reimpr. São Paulo: Perspectiva, 2005.