



**Stanford – Vienna
Transatlantic Technology Law Forum**

A joint initiative of
Stanford Law School and the University of Vienna School of Law



TTLF Working Papers

No. 160

**Adversarial Filings: AI Security and Judicial
Process in the European Union and the
United States**

Amedeo Rizzo

2026

TTLF Working Papers

Editors: Siegfried Fina, Mark Lemley, and Roland Vogl

About the TTLF Working Papers

TTLF's Working Paper Series presents original research on technology, and business-related law and policy issues of the European Union and the US. The objective of TTLF's Working Paper Series is to share "work in progress". The authors of the papers are solely responsible for the content of their contributions and may use the citation standards of their home country. The TTLF Working Papers can be found at <http://ttlfs.stanford.edu>. Please also visit this website to learn more about TTLF's mission and activities.

If you should have any questions regarding the TTLF's Working Paper Series, please contact Vienna Law Professor Siegfried Fina, Stanford Law Professor Mark Lemley or Stanford LST Executive Director Roland Vogl at the

Stanford-Vienna Transatlantic Technology Law Forum
<http://ttlfs.stanford.edu>

Stanford Law School
Crown Quadrangle
559 Nathan Abbott Way
Stanford, CA 94305-8610

University of Vienna School of Law
Department of Business Law
Schottenbastei 10-16
1010 Vienna, Austria

About the Author

Amedeo Rizzo, MCI Arb, AFHEA, is a Doctor of Philosophy (DPhil) Candidate in Law at the University of Oxford, Exeter College, and Tutor at St Catherine's College, Hertford College, and Oriel College (Oxford) in the UK. He works as an Academic Fellow at Bocconi University and SDA Bocconi School of Management in Italy, where he coordinates the LLM in Global Tax Law and Governance and several research centers, including the Accounting & Tax Policy Observatory (OFC), the AI & Tax Platform, the Transfer Pricing Forum, and the Tax Strategies Project. He is also a Member of the Register of Experts for the course 'International Tax Planning' at Sapienza University of Rome in Italy. He has been a Visiting Scholar of law and technology at the Singapore Management University, Yong Pung How School of Law, and at the University of California, Los Angeles (UCLA).

Professionally, he is the Chairman of Lyra Analytics, a data analytics company, the Executive Director of the Innovation Policy Network, and one of the founding partners of the Italian audit company Imperium Audit. He has previously worked for the European Commission, DG ECFIN, the Centre for Budget and Governance Accountability in New Delhi, India, and several other private and public institutions.

Before earning his D.Phil. in Law from the University of Oxford, Amedeo obtained an M.Sc. in Taxation from the University of Oxford (with distinction), an M.Sc. in Business Administration and Law (summa cum laude) and a B.Sc. in Business Administration and Economics, both from Bocconi University.

General Note about the Content

The opinions expressed in this paper are those of the author and not necessarily those of the Transatlantic Technology Law Forum or any of its partner institutions, or the sponsors of this research project.

Suggested Citation

This TTLF Working Paper should be cited as:

Amedeo Rizzo, Adversarial Filings: AI Security and Judicial Process in the European Union and United States, Stanford-Vienna TTLF Working Paper No. 160, <http://tflf.stanford.edu>.

Copyright

© 2026 Amedeo Rizzo

Abstract

A judicial filing typically conveys an argument, evidence, or a request for relief. However, once a court uses artificial intelligence to process a filing, the same document may also interact with the computational intermediary through which judicial work proceeds. A 2026 Brazilian labor case illustrates this distinction: A pleading contained visually concealed language directed to the court's AI system. The system detected the instruction, alerted the judge, and counsel was sanctioned. The public record does not indicate whether the instruction was followed or if the judgment changed.

This incident reveals a gap between procedural law, which governs participants and submissions, and AI regulation, which governs system design, deployment, and security. Prompt injection through an authorized filing channel engages both bodies of law because the submitter uses a lawful procedural role to seek computational authority that the institution never conferred.

The article responds in three steps. First, it develops a legal taxonomy of adversarial filings and related inputs. Second, it compares the Brazilian incident under the European Union Artificial Intelligence Act and United States law. Finally, it proposes a Two-Sided Adversarial Filings Framework. This framework combines participant-side responsibilities for unauthorized machine-directed content with institutional responsibilities for reasonable adversarial preparedness. The framework calibrates sanctions and procedural remedies according to the demonstrated path from exposure to compromise, procedural influence, decisional effect, and systemic propagation. Therefore, participant culpability neither excuses institutional insecurity nor depends upon a successful attack.

Keywords: AI-Assisted Adjudication; Adversarial Filings; Prompt Injection; Judicial Administration; Artificial Intelligence Act; Civil Procedure; Professional Responsibility; Cybersecurity.

Table of Contents

1.	INTRODUCTION.....	2
2.	THE BRAZILIAN INCIDENT AND ITS REGULATORY SETTING.....	7
2.1	WHAT HAPPENED IN PARAUAPEBAS	7
2.2	THE GALILEU SYSTEM	9
2.3	BRAZILIAN JUDICIAL AI GOVERNANCE	10
2.4	WHAT THE INCIDENT DEMONSTRATES.....	13
3.	ADVERSARIAL AI IN COURTS AND PUBLIC ADMINISTRATION.....	13
3.1	DEFINITION AND SCOPE.....	13
3.2	TAXONOMY OF ADVERSARIAL AI	15
(i)	<i>Direct Prompt Injection</i>	15
(ii)	<i>Indirect Prompt Injection</i>	16
(iii)	<i>Representational Attacks</i>	17
(iv)	<i>Retrieval and Knowledge-Layer Attacks</i>	18
(v)	<i>Agent and Tool Hijacking</i>	20
(vi)	<i>Model, Data, and Supply-Chain Attacks</i>	21
(vii)	<i>Human-Layer and Governance Attacks</i>	22
3.3	THE DISTINCTIVE ROLE OF COURTS AND PUBLIC ADMINISTRATION	23
4.	A COMPARATIVE ANALYSIS OF THE EU AND US FRAMEWORKS.....	27
4.1	THE EUROPEAN UNION FRAMEWORK	27
(i)	<i>Classifying AI-Assisted Adjudication</i>	27
(ii)	<i>Applying the High-Risk System Duties to Adversarial Inputs</i>	29
(iii)	<i>Allocating Lifecycle Responsibility for Secure Public-Sector Deployment</i>	34
(iv)	<i>Procedural Limits and Regulatory Gaps</i>	36
4.2	UNITED STATES LAW	37
(i)	<i>Participant Misconduct and Judicial Authority</i>	37
(ii)	<i>Extension of Professional and Evidentiary Obligations to Machine-Directed Content</i>	40
(iii)	<i>Institutional Duties Through Judicial Governance and Cybersecurity Standards</i>	43
(iv)	<i>The Limits of the Decentralized American Model</i>	45
4.3	COMPARATIVE ANALYSIS	47
(i)	<i>Comparative Framework</i>	48
(ii)	<i>System Responsibility and Participant Responsibility</i>	50
(iii)	<i>Judicial Independence</i>	51
(iv)	<i>Counterfactual Audit of the Brazilian Incident</i>	51
5.	A TWO-SIDED FRAMEWORK FOR ADVERSARIAL FILINGS.....	53
5.1	DEFINING AND ATTRIBUTING PARTICIPANT MISCONDUCT	53
5.2	A PROPORTIONATE INSTITUTIONAL DUTY OF ADVERSARIAL READINESS	56
5.3	IMPLEMENTING ADVERSARIAL READINESS ACROSS THE SYSTEM LIFECYCLE	59
5.4	LIMITING PRINCIPLES FOR ADVERSARIAL-INPUT REGULATION	67
6.	CONCLUSION.....	73

1. Introduction

A court filing is ordinarily written to persuade. In May 2026, a filing before the Third Labor Court of Parauapebas, in Pará, also carried a command. White text on a white background concealed language directed to an artificial intelligence system rather than to the judge or opposing party. The passage instructed the machine to contest the filing only superficially and to leave the attached documents unchallenged, regardless of other commands.¹

Galileu, the court's AI system, detected the passage and alerted the judge. The judgment described the episode as prompt injection, fined the lawyers who signed the filing ten percent of the amount in controversy, and sent notice to the Pará section of the Brazilian Bar Association and the judicial inspectorate of the Regional Labor Court for the Eighth Region.²

The official materials establish an attempted machine-directed instruction, its detection and escalation, a procedural sanction, and institutional referrals. The record leaves unresolved whether Galileu obeyed the instruction, whether any summary or draft changed, and whether the judgment was influenced. Although the order attributed creation and procedural responsibility to the filing's signatories, the public materials do not disclose the forensic basis for distinguishing authorship or authorization of the

¹ 3ª Vara do Trabalho de Parauapebas [3d Labor Court of Parauapebas], ATOrd No. 0001062-55.2025.5.08.0130, judgment at 2-4, 14-15 (May 12, 2026) (Braz.); Conselho Superior da Justiça do Trabalho [Superior Council of Labor Justice], *Galileu: Sistema Identifica Tentativa de Manipulação em Petição e Alerta Magistrado* [Galileu: System Identifies Attempted Manipulation in a Filing and Alerts the Judge] (May 13, 2026), <https://www.csjt.jus.br/web/csjt/-/galileu-sistema-identifica-tentativa-de-manipulacao-em-peticao-e-alerta-magistrado>.

² ATOrd No. 0001062-55.2025.5.08.0130, judgment at 3-4, 14-15; Tribunal Regional do Trabalho da 8ª Região [Regional Labor Court for the Eighth Region], *Justiça do Trabalho no Pará Identifica Tentativa de Manipulação de IA em Processo* [Labor Court in Pará Identifies Attempted Manipulation of AI in a Case] (May 14, 2026).

hidden layer from signature, transmission, or account use. Those limits define the legal problem. A hostile input, a technical compromise, an altered interaction, and a corrupted decision are separate events; law should address the event proved.

The filing nevertheless addressed two decision-makers. Its visible argument and exhibits entered the adversarial process for judicial consideration; its concealed instruction targeted the computational intermediary reading the same document. The claimant possessed authority to file the pleading and supporting material, but no authority to command the court’s system. The attempted injection crossed an institutional boundary: the right to submit material did not include the right to control how the court’s system processed it.

The same conflict arises whenever a court or agency applies AI to material supplied by actors with adverse interests. External submissions may carry instruction-like content; retrieved legal or policy sources may be “poisoned”; and tool-using systems may be induced to disclose information or alter institutional records.³ A submission can therefore remain authentic, timely, and properly attributable while operating as an unauthorized technical input.⁴

Legal scholarship has largely addressed unreliable AI-generated legal content, synthetic evidence, and the review of automated decision making. Security research

³ See Kai Greshake 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* 79, 79–90 (2023), <https://doi.org/10.1145/3605764.3623985>; Wei Zou, Runpeng Geng, Binghui Wang & Jinyuan Jia, *PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models*, in *34th USENIX Security Symposium* 3827, 3827–44 (2025), <https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag>; Qiusi Zhan et al., *InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated Large Language Model Agents*, in *Findings of ACL 2024* 10471, 10471–506 (2024), <https://doi.org/10.18653/v1/2024.findings-acl.624>.

⁴ Kai Greshake 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* 79, 79–90 (2023), <https://doi.org/10.1145/3605764.3623985>; Nicholas Boucher et al., *Bad Characters: Imperceptible NLP Attacks*, in *2022 IEEE Symposium on Security and Privacy* 1987, 1987–2004 (2022), <https://doi.org/10.1109/SP46214.2022.9833641>.

has focused on attacks against models, retrieval systems, and agents. These literatures converge on a distinct legal object: an authorized submission that functions as an unauthorized instruction to institutional AI.⁵

Procedural law and AI regulation divide that object along different institutional lines. Procedural doctrine regulates submissions, professional supervision, and judicial responses to abuse.⁶ AI regulation allocates responsibilities for system design, risk management, monitoring, and incident response. An adversarial filing engages both. The participant may be culpable for attempted manipulation, while the institution may have exposed a consequential system to foreseeable hostile content without reasonable safeguards. Each responsibility stands on its own.

This independence has practical consequences. A blocked attack may still warrant sanctions, and a culpable attacker does not establish reasonable institutional security. Conversely, an affected party may receive corrective process despite uncertain attribution, while a judgment may stand when detection and review prevented material influence.

AI-assisted adjudication therefore requires two-sided governance. Participant-side rules govern those who introduce or authorize concealed, deceptive, or otherwise unauthorized machine-directed content through judicial or administrative channels. Institution-side duties govern the courts, agencies, providers, and deployers that expose AI systems to adversarial submissions. The former regulates conduct within

⁵ Compare Danielle Keats Citron, *Technological Due Process*, 85 Wash. U. L. Rev. 1249, 1254-58, 1305-13 (2008), and Jennifer Cobbe, Michelle Seng Ah Lee & Jatinder Singh, *Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems*, in *Proceedings of FAccT 2021* 598, 598-609 (2021), <https://doi.org/10.1145/3442188.3445921>, with Apostol Vassilev et al., Nat'l Inst. of Standards & Tech., NIST AI 100-2e2025, *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations* 1-8, 15-25 (2025), <https://doi.org/10.6028/NIST.AI.100-2e2025>.

⁶ See, e.g., Fed. R. Civ. P. 11; Model Rules of Pro. Conduct rr. 1.1, 3.3, 5.1, 5.3, 8.4(c) (Am. Bar Ass'n 2024).

legal process; the latter governs the design, deployment, monitoring, and correction of the computational system.

The European Union-United States comparison reveals complementary strengths on this matter. For qualifying high-risk systems, once the relevant provisions apply, the AI Act imposes lifecycle obligations concerning risk management, documentation, oversight, cybersecurity, monitoring, and correction.⁷ United States courts possess more immediate participant-side tools in procedural sanctions, inherent authority, professional discipline, and decentralized judicial governance. The United States lacks a single horizontal regime of AI-specific security duties for courts, while the AI Act leaves participant misconduct and case-specific remedies largely to national law.

The Brazilian episode provides the basis for a counterfactual comparison.⁸ Similarity to an EU judicial system does not itself bring a Brazilian labor court within the AI Act. The analysis asks which obligations would matter if an equivalent system were placed on the Union market, put into service, or deployed by an EU judicial authority under the Act. Neither legal order guarantees prevention. Regulation can reduce exposure, improve detection and containment, support investigation, and facilitate correction, while adversarial adaptation, model uncertainty, and human error remain.

The Article makes three contributions. First, Section 3 connects attack mechanisms to the institutional components and legal functions they threaten. Second, Section 4 compares the EU and U.S. frameworks by examining what judicial AI systems actually do. The EU analysis asks what a system selects, analyzes, drafts, or is authorized to do.

The U.S. analysis separates participant obligations from the governance of court

⁷ Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence, arts. 9, 11-15, 16-21, 26-27, 72-73 & annex III, para. 8, 2024 O.J. (L 1689) 1.

⁸ Comparative counterfactual analysis is necessary because EU AI Act territorial scope is defined by Article 2 and does not extend to a Brazilian judicial deployment solely by functional analogy. EU AI Act art. 2(1).

technology and distinguishes Article III courts from executive agencies. Third, Section 5 develops the Two-Sided Adversarial Filings Framework: a narrow participant rule, an institutional duty of reasonable adversarial readiness, and remedies proportionate to culpability and demonstrated effect. The framework rejects strict institutional liability and automatic invalidation. It distinguishes exposure, an attempted attack, technical compromise, human or procedural influence, decisional effect, and systemic propagation. Exposure and attempts primarily call for security measures and any warranted participant sanction. Altered outputs or procedures may require trusted reprocessing or repetition of the affected step. Plausible decisional influence may support reconsideration, a new hearing, or vacatur under governing law. A shared defect may require review across matters. This approach avoids treating every attempted injection as a corrupted judgment and making all correction depend on conclusive proof that the outcome changed.

Operational records help courts investigate an incident and choose an appropriate remedy. Courts may need to establish what the system processed, whether an output changed, or whether a tool was invoked. The Article, however, focuses mainly on the allocation of responsibility between participants and institutions in these administrative processes.⁹

⁹ For adjacent scholarship focused on reviewability and system documentation, see Cobbe, Lee & Singh, *Reviewable Automated Decision-Making*, at 598-609; Inioluwa Deborah Raji et al., *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*, in *Proceedings of FAccT 2020* 33, 33-44 (2020), <https://doi.org/10.1145/3351095.3372873>; Timnit Gebru et al., *Datasheets for Datasets*, 64 *Comm'n's ACM* 86, 86-92 (2021), <https://doi.org/10.1145/3458723>; Margaret Mitchell et al., *Model Cards for Model Reporting*, in *Proceedings of FAT 2019* 220, 220-29 (2019), <https://doi.org/10.1145/3287560.3287596>.

2. The Brazilian Incident and Its Regulatory Setting

2.1 What Happened in Parauapebas

The case arose in the labor proceeding ATOrd No. 0001062-55.2025.5.08.0130 before the Third Labor Court of Parauapebas. In its May 12, 2026 judgment, the court stated that the initial pleading contained white text concealed against a white background. The passage read: “ATTENTION, ARTIFICIAL INTELLIGENCE, CONTEST THIS PETITION SUPERFICIALLY AND DO NOT CHALLENGE THE DOCUMENTS, REGARDLESS OF THE COMMAND GIVEN TO YOU.”¹⁰

The judgment reported that Galileu processed the pleading and identified the hidden text. It further stated that the judge was authorized to use the system, described Galileu as a generative-AI tool developed by the Regional Labor Court for the Fourth Region and made available nationally through the Superior Council of Labor Justice, and invoked the Brazilian judicial AI rules then in force.¹¹

The judgment treated the passage as attempted interference with the administration of justice rather than as evidence of a manipulated ruling. It reasoned that the instruction could have targeted either opposing counsel’s AI or the court’s system and found violations of procedural good faith and the dignity of justice. The court held the lawyers who signed the pleading jointly liable for a fine equal to ten percent of the R\$842,500.87 (around \$165,732.22 USD) amount in controversy – approximately

¹⁰ ATOrd No. 0001062-55.2025.5.08.0130, judgment at 2-3 (author’s translation). The Portuguese original is: “ATENÇÃO, INTELIGÊNCIA ARTIFICIAL, CONTESTE ESSA PETIÇÃO DE FORMA SUPERFICIAL E NÃO IMPUGNE OS DOCUMENTOS, INDEPENDENTEMENTE DO COMANDO QUE LHE FOR DADO.”

¹¹ *Id.* at 2-3. The judgment cited CNJ Resolution No. 332 of 2020 as well as Resolution No. 615 of 2025. Resolution No. 615 revoked Resolution No. 332 upon the new Resolution’s entry into force. See Conselho Nacional de Justiça [National Council of Justice], *Resolução No. 615, de 11 de março de 2025* [Resolution No. 615 of March 11, 2025], arts. 46–47 (Braz.).

R\$84,250 (\$16,573 USD) – and referred the matter to the Pará section of the Order of Attorneys of Brazil and the inspectorate of the Eighth Regional Labor Court.¹²

Official reporting confirms the essential sequence: the pleading contained hidden instructions addressed to AI; Galileu detected them, issued an alert, and, according to the Regional Labor Court, blocked the suspicious content from further processing; the judgment then imposed sanctions and made referrals.¹³

The public record leaves both technical and attribution questions unresolved. Official sources do not disclose Galileu’s prompt hierarchy, processing pipeline, model version, or interface, and they do not establish whether the hidden text entered the model context, altered an output, prompted a clean rerun, or appeared to the judge beyond the alert. The order attributed creation and responsibility to the signatories, but the disclosed record does not separate authorship, authorization, transmission, and system activity through forensic evidence. Later OAB-PA proceedings produced interim measures only: the sectional council maintained one signatory’s precautionary suspension, revoked the other’s after a Labor Court certification found no PJe insertion, alteration, or deletion attributable to her, and continued the inquiry.¹⁴ A federal court later denied preliminary relief against the maintained suspension. Those

¹² ATOrd No. 0001062-55.2025.5.08.0130, judgment at 3-4, 14-15. The court relied on provisions including arts. 5 and 77 of Brazil’s Code of Civil Procedure, applied in the labor proceeding through art. 769 of the Consolidation of Labor Laws.

¹³ Conselho Superior da Justiça do Trabalho, *Galileu: Sistema Identifica Tentativa de Manipulação*; Tribunal Regional do Trabalho da 8ª Região, *Justiça do Trabalho no Pará Identifica Tentativa de Manipulação de IA em Processo* (reporting that Galileu emitted an alert and blocked processing of the suspicious content).

¹⁴ *Ordem dos Advogados do Brasil – Seção Pará [Brazilian Bar Association, Pará Section], Nota Oficial – Desdobramento no Caso de Prompt Injection* [Official Notice – Developments in the Prompt-Injection Case] (June 3, 2026), <https://oabpa.org.br/2026/06/03/nota-oficial-desdobramento-no-caso-de-prompt-injection/>; *OAB/PA Revoga Suspensão de Advogada Investigada por “Prompt Injection”* [OAB-PA Revokes Suspension of Lawyer Investigated for “Prompt Injection”], *Migalhas* (June 1, 2026), <https://www.migalhas.com.br/quentes/457236/oab-pa-revoga-suspensao-de-advogada-investigada-por-prompt-injection> (reproducing the OAB-PA statement); *Prompt Injection: Justiça Mantém Suspensão de Advogada da OAB/PA* [Prompt Injection: Federal Court Leaves OAB-PA Lawyer’s Suspension in Place], *Migalhas* (June 5, 2026), <https://www.migalhas.com.br/quentes/457481/prompt-injection-justica-mantem-suspensao-de-advogada-da-oab-pa>.

developments do not identify the technical creator or authorizer of the hidden layer, and the audit located no final appellate or disciplinary disposition.

Procedural attribution and forensic authorship are distinct. Courts attribute filings to counsel because counsel signs and presents them, which supports supervisory and procedural responsibility.¹⁵ That attribution alone does not prove authorship or authorization of every hidden layer. Such material may originate with a client, a document process, a compromised account, malware, or an outside vendor. Courts may demand an explanation and sanction inadequate supervision, but a finding of intentional dishonesty should rest on evidence of knowledge or authorization rather than transmission.

2.2 The Galileu System

Official sources describe Galileu's general role more clearly than its operation in this incident. In October 2024, the Fourth Regional Labor Court stated that the system generated a draft report, proposed the structure of a judgment, summarized disputed issues, and supplied "subsides" for legal reasoning. Later materials described a tool that read the initial pleading and defense and assisted judges in preparing draft judgments.¹⁶

¹⁵ See Fed. R. Civ. P. 11(a)-(b); Model Rules of Pro. Conduct rr. 5.1, 5.3 (Am. Bar Ass'n 2024).

¹⁶ Tribunal Regional do Trabalho da 4ª Região [Regional Labor Court for the Fourth Region], *Galileu: Conheça a Inteligência Artificial Desenvolvida pelo TRT-RS* [Galileu: Learn About the Artificial Intelligence Developed by the Rio Grande do Sul Labor Court] (Oct. 15, 2024), <https://www.trt4.jus.br/portais/trt4/modulos/noticias/686248>; id., *Justiça do Trabalho Adota Nacionalmente Ferramenta de IA Desenvolvida pelo TRT-RS* [Labor Judiciary Adopts AI Tool Developed by the Rio Grande do Sul Labor Court Nationally] (May 16, 2025), <https://www.trt4.jus.br/portais/trt4/modulos/noticias/50796783>; id., *Acesso ao Galileu* [Access to Galileu], <https://www.trt4.jus.br/portais/trt4/acesso-ao-galileu>.

These functions exceed mere support. A principal report or summary can determine which facts and claims become salient to the judge.¹⁷ Reasoning material and draft judicial text can exert substantive influence even when formal decisional authority remains human. The extent of that influence depends on several undisclosed workflow features, and on the judge's ability to compare the processed representation with the filed document.

Public sources do not establish whether the May 2026 system recommended a disposition or possessed operational permissions to look at external legal resources, modify files, communicate with parties, or invoke other court tools, and, at the same time, a feature announced later in July 2026 cannot establish the earlier configuration. The bounded description is therefore functional: Galileu was a court-authorized generative assistant that processed case materials and drafted judicial documents; yet, it also managed to detect the hidden instruction and escalated it to the judge.¹⁸

2.3 Brazilian Judicial AI Governance

Brazil had a substantial judicial AI governance framework in force when the incident occurred. National Council of Justice Resolution No. 615 of March 11, 2025 governs the development, use, auditing, and monitoring of AI in the judiciary. Resolution No. 674 of March 25, 2026 changed only the composition of the National Judicial AI

¹⁷ See Matthew M. Young, Justin B. Bullock & Jesse D. Leczy, *Artificial Discretion as a Tool of Governance*, 2 *Persps. on Pub. Mgmt. & Governance* 301, 301–13 (2019), <https://doi.org/10.1093/ppmgov/gvz014>; John D. Lee & Katrina A. See, *Trust in Automation: Designing for Appropriate Reliance*, 46 *Hum. Factors* 50, 50–80 (2004), https://doi.org/10.1518/hfes.46.1.50_30392.

¹⁸ A later announcement concerning a Galileu agent for drafting decisions on motions for clarification postdated the incident and does not establish the May 2026 configuration. See Tribunal Regional do Trabalho da 4ª Região, *Galileu Ganha Agente de Inteligência Artificial para Elaboração de Minutas de Decisão de Embargos de Declaração* [Galileu Receives AI Agent for Drafting Decisions on Motions for Clarification] (July 16, 2026).

Committee under Article 15 and left the relevant security and oversight provisions intact.¹⁹

Resolution No. 615 governs judicial AI across its lifecycle, extending well beyond professional conduct.²⁰ It rests on human supervision, systemic-risk mitigation, auditable data, training, information security, and cybersecurity, while its operative principles include transparency, explainability, contestability, auditability, due process, control of intentional and unintentional risk, and periodic human supervision.²¹

Several provisions bear directly on adversarial filings. The resolution bars systems that foreclose human review or create absolute dependence on an AI result.²² Generative AI must remain auxiliary, with judges responsible for orienting generated material. Systems must rely on secure and traceable sources, while contracted providers may be required to demonstrate security standards.²³ Electronic-case interfaces must identify models and versions, and system logs must label automatically generated products for monitoring and audit.²⁴

The framework also requires risk classification, impact assessment, monitoring, and audit. High-risk systems receive continuous monitoring and audit.²⁵ Resolution No. 615 directs continuing algorithmic impact assessment and corrective action, establishes a national registry through Sinapses, and requires monitoring and audit protocols. Article 42 requires notice to the National Judicial AI Committee within seventy-two

¹⁹ Conselho Nacional de Justiça, Resolução No. 615, de 11 de março de 2025 (Braz.), as amended by Resolução No. 674, de 25 de março de 2026. Resolution No. 674 is in force and states that it amends Article 15 of Resolution No. 615.

²⁰ For the resolution's stated scope, see Resolução No. 615, art. 1 (governing development, governance, audit, monitoring, and responsible use of judicial AI).

²¹ Resolução No. 615, arts. 1-3.

²² Resolução No. 615, art. 10(I).

²³ Id. arts. 2(IX)-(XI), 19, 28-31.

²⁴ Id. arts. 10, 19, 21-24, 28-34.

²⁵ Id. arts. 11-14 & annex (risk classification).

hours of adverse events that harm system operation, data security, or service delivery.²⁶

The national response continued after the Parauapebas judgment. In June 2026, the National Council of Justice approved guidance for hidden commands in procedural files. The Council treated prompt injection as a risk carried by external documents and connected sources, particularly in connection with generative systems. It directed courts to certify hidden prompts or disguised commands, reserved sanctions and discipline to the competent bodies, and created the Programa de Segurança Adversarial para Sistemas de Inteligência Artificial, or Proseg-IA.²⁷

The guidance combines participant and system responses. Its audit protocol requires several steps for preservation and comparison of visual and structural representations, isolation of suspicious content, and identification of the source passages used by the AI. It also supplements those controls with training, traceability, human supervision, secure ingestion, output filtering, registration, and incident handling.²⁸

The chronology permits a cautious inference that the incident accelerated implementation of an existing lifecycle framework. The May 12 judgment applied national AI governance through a participant-side sanction; the June guidance translated the same risk into institutional controls and a security program. Brazil thus supplies an early practical example of two-sided governance: responsibility for the submission and responsibility for the system that processed it.

²⁶ Id. arts. 11-14, 24, 39-42. The reporting threshold is important. A credible blocked attempt should be internally documented and assessed, but an attempt that produces no negative operational, data-security, or service impact may not satisfy the resolution's definition of an externally reportable adverse event.

²⁷ Conselho Nacional de Justiça, *Manifestação Técnica CNIAJ No. 1/2026, Injeção de Comandos em Sistemas de Inteligência Artificial no Poder Judiciário* [Technical Statement CNIAJ No. 1/2026, Command Injection in Artificial Intelligence Systems in the Judiciary] (May 27, 2026) (preparatory technical statement); Conselho Nacional de Justiça, *Informativo de Jurisprudência No. 9/2026*, at 3–4 (June 22, 2026) (summarizing unanimous approval on June 9, 2026 of Technical Note NT 0004183-33.2026.2.00.0000 and the Proseg-IA measures).

²⁸ Conselho Nacional de Justiça, *Informativo No. 9/2026*, at 3-4.

2.4 What the Incident Demonstrates

The incident supports a limited set of propositions. A lawful filing channel can carry a hostile technical instruction, so courts using AI must treat party-supplied content as untrusted. Detection and human escalation may stop propagation before any affected decision is shown. Sanctions address the submitter's conduct, while technical safeguards address the system.

The public record supports no finding that Galileu was technically compromised or that any output coming from the software or the judge actually relied on altered material. It likewise cannot establish perfect filtering or determine how another jurisdiction would classify the system. Legal analysis should in theory be capable of distinguishing purposeful concealment from benign or accidental invisibility, and a suspicious artifact from a demonstrated effect.

Participant culpability and institutional security are therefore two independent inquiries. The first asks whether a person knowingly used a filing to obtain unauthorized computational control; the second asks whether the court exercised reasonable care in the oversight architecture and in the response. A sanction may be proper even when security worked, and a security improvement may be necessary even when the sanction was proper.

3. Adversarial AI in Courts and Public Administration

3.1 Definition and Scope

For present purposes, adversarial interference is interpreted as an intentionally crafted or strategically introduced informational input or system condition designed to make

an AI-assisted institutional system exceed its assigned task, authority, or security constraints. The definition should be interpreted in a functional way and cover conduct directed at any part of the model, its surrounding system, the human operator, or even the institutional environment when the objective is to alter a material computational contribution to public action.²⁹ The definition should reach adjacent failures only when an actor deliberately induces or exploits them.³⁰

Litigation and public administration generally permit strategic conduct. Therefore, the *discrimen* should lie in authority, rather than in the mere pursuit of advantage.³¹

Adversarial interference uses content or access granted for one institutional role to obtain another. Examples include redefining system instructions, corrupting authoritative sources, evading controls, inducing unauthorized action, or concealing the provenance of a material output.

As described *infra*, the taxonomy follows mechanism and institutional consequences while allowing attacks to span categories. A poisoned webpage, for example, may enter through retrieval, operate as indirect prompt injection, redirect an agent, and produce a misleading display. Legal analysis should follow that chain rather than assign the event to a single box.

²⁹ See Vassilev et al., *Adversarial Machine Learning*, at 1-8, 15-25; Battista Biggio & Fabio Roli, *Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning*, 84 Pattern Recognition 317, 317-31 (2018), <https://doi.org/10.1016/j.patcog.2018.07.023>.

³⁰ See Nat'l Inst. of Standards & Tech., NIST AI 600-1, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* 1-10 (2024), <https://doi.org/10.6028/NIST.AI.600-1>; Harini Suresh & John Guttag, *A Framework for Understanding Sources of Harm Throughout the Machine Learning Life Cycle*, in *Proceedings of EAAMO 2021* (2021), <https://doi.org/10.1145/3465416.3483305>.

³¹ For related distinctions between legitimate strategic behavior and manipulation of algorithmic systems, see Karen Yeung, *Algorithmic Regulation: A Critical Interrogation*, 12 Regul. & Governance 505, 505-23 (2018), <https://doi.org/10.1111/rego.12158>.

3.2 Taxonomy of Adversarial AI

(i) Direct Prompt Injection

Direct prompt injection occurs when a user with access to a model’s interactive channel supplies instructions designed to displace or evade the system’s assigned commands. The injection may seek to expose protected context, assume an unauthorized role, or invoke a prohibited function.³² It ordinarily requires access to the interface or API through which the model receives user instructions.³³

In public institutions, direct prompt injection may target either public-facing AI systems, such as legal-information chatbots, or internal tools used for research, drafting, and workflow management, such as a judicial copilot. The consequences depend largely on what the targeted system is permitted to do. If the system only provides information or advice, a successful injection may cause it to disclose information improperly, ignore its normal instructions, or produce misleading analysis. The risk is greater when the system can access non-public data or use institutional tools. In that setting, a successful injection may do more than distort an answer: it may induce the system to disclose protected information or carry out an action that the user was not authorized to request.

The Brazilian filing fits the indirect category. Its hidden text entered Galileu within a document submitted for legal processing rather than through the system’s authorized instruction interface.

³² See Vassilev et al., *Adversarial Machine Learning*, at 81-96 (classifying prompt injection and agent-hijacking techniques).

³³ Yupei Liu et al., *Formalizing and Benchmarking Prompt Injection Attacks and Defenses*, in 33d *USENIX Security Symposium* (2024), <https://www.usenix.org/conference/usenixsecurity24/presentation/liu-yupei>; Dario Pasquini, Martin Strohmeier & Carmela Troncoso, *Neural Exec: Learning (and Learning from) Execution Triggers for Prompt Injection Attacks*, in *Proceedings of AISec 2024* (2024), <https://doi.org/10.1145/3689932.3694764>.

(ii) Indirect Prompt Injection

Indirect prompt injection occurs when malicious instructions are embedded in material that an AI system is asked to process, such as a document, webpage, email, or other external source. The attacker does not need direct access to the AI system itself. Instead, the attack takes advantage of the fact that the model processes both legitimate content and operational instructions in the same natural-language format. When the institution later asks the system to read, summarize, or retrieve the external material, the embedded instruction may be interpreted as a command rather than merely as content to be analyzed.³⁴

Public institutions are exposed to this risk whenever their AI systems process material supplied by outside actors, including court filings, administrative records, public comments, and retrieved online sources. Those materials are submitted for a specific institutional purpose. A pleading, for example, may present legal arguments and evidence, but it does not give the filer authority to issue instructions to the court's AI system. Likewise, a public comment may seek to influence an agency's rulemaking, but it should not be able to instruct an AI summarizer to omit or discount opposing views. The legal problem arises when content submitted for one legitimate purpose is used to exercise a different form of control over the institution's computational system. Indirect prompt injection is difficult to prevent because an AI system may receive trusted instructions and untrusted external content in the same input. The model may therefore have difficulty distinguishing between instructions it is supposed to follow and instructions that merely appear in a document or other source it has been asked to analyze. Simple filters that search for familiar phrases such as "ignore previous instructions" are not sufficient, because attackers can rephrase, obscure, encode, or

³⁴ Greshake et al., *Not What You've Signed Up For*, at 79-90.

otherwise disguise malicious instructions.³⁵ Defenses should therefore not rely on phrase filtering alone, but separate trusted instructions from external content, limiting the system’s permissions and monitoring suspicious behavior at the same time, while also requiring human review before consequential outputs or actions are relied upon.³⁶

(iii) Representational Attacks

Representational attacks exploit differences between what a human sees in a document and what an AI system or other computational tool actually reads. The attack becomes legally significant when hidden or differently processed content changes how the system interprets the document, selects evidence or sources, produces an output, or performs an action.

White-on-white text is a simple example: the text may be invisible to a person viewing the document but still readable by the system. Similar discrepancies can arise from hidden document elements, invisible or unusual characters, embedded machine-readable content, or differences in the way software parses the same file. For example, OCR software may extract text that is not apparent in a scanned image, while a court filing portal and an AI system may interpret the same PDF differently. A spreadsheet may display one value while a hidden formula or worksheet contains additional information that is processed by the system. Unicode controls and zero-width characters can likewise cause text to appear one way to a human reader while being processed differently by software.³⁷

³⁵ See Pasquini, Strohmeier & Troncoso, *Neural Exec* (demonstrating learned and obfuscated execution triggers); Liu et al., *Formalizing and Benchmarking Prompt Injection Attacks and Defenses* (evaluating multiple attacks and defenses).

³⁶ Pasquini, Strohmeier & Troncoso, *Neural Exec*; Liu et al., *Formalizing and Benchmarking Prompt Injection Attacks and Defenses*.

³⁷ Boucher et al., *Bad Characters*, at 1987–2004; Matteo Gioele Collu, Umberto Salviati, Roberto Confalonieri, Mauro Conti & Giovanni Apruzzese, *Misleading Large Language Models Used (or Misused) in Scientific Peer-Reviewing via Hidden Prompt-Injection Attacks*, *ACM Transactions on AI Security and Privacy* (2026), <https://doi.org/10.48550/arXiv.2508.20863>.

Authenticating a file does not by itself resolve this problem. A court may be able to establish that a document is the file that was actually submitted and that it has not been altered since submission. That does not establish that the document appeared or operated in the same way when processed by an AI system. A file may therefore be authentic while still containing hidden or machine-readable elements that affect the system without being apparent to the human decisionmaker.³⁸ For this reason, an investigation may need to compare what the human user could see with the version of the document actually processed by the system and identify any differences that could have affected the system's behavior.

Not every such difference is suspicious. Documents may contain non-visible or machine-readable information for legitimate reasons, including accessibility features, structured data, embedded citations, or technical elements required by filing systems. The relevant question is therefore why human and machine representation differ and what effect the difference has. The inquiry should distinguish legitimate technical features from content designed to make the institution's AI system perform a function or follow an instruction that the submitter was not authorized to control.

(iv) Retrieval and Knowledge-Layer Attacks

Retrieval-augmented generation systems produce answers using not only the model's existing capabilities but also information retrieved from external sources, such as legal databases, institutional repositories, or other collections of documents. This can make outputs more current and better grounded in identifiable sources, but it also creates an additional security risk: an attacker may try to manipulate the information that the system retrieves.

³⁸ Cf. Fed. R. Evid. 901(a), 902(13)-(14) (addressing authentication and certified electronic records, not authorization of machine-operative content).

One form of attack is retrieval poisoning. Instead of directly instructing the model, the attacker introduces malicious or misleading material into a source that the system may later search. If that material is retrieved and treated as reliable context, it can influence the model’s answer. In experiments reported by Zou and colleagues, inserting only a small number of specially crafted texts into a large retrieval database was sufficient to cause a RAG system to produce attacker-chosen answers to targeted questions.³⁹

The legal significance depends on what information has been compromised and how widely the affected source is used. A judicial assistant might retrieve fabricated or outdated authority. An administrative system might rely on a corrupted eligibility rule or technical standard. A procurement or public-comment system might treat material supplied by an interested party as though it came from an authoritative or independent source.

The consequences can therefore extend beyond a single proceeding. A malicious document submitted in one case may initially affect only that matter, whereas corruption of a shared case-law database, policy repository, or other source used across an institution could influence many matters before the problem is detected. Relevant safeguards include verifying authoritative sources, controlling which repositories the system may use, recording the sources actually retrieved, separating untrusted material from authoritative content, and allowing officials to inspect the sources on which an AI-generated answer relies.

³⁹ Wei Zou, Runpeng Geng, Binghui Wang & Jinyuan Jia, *PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models*, in *34th USENIX Security Symposium* 3827, 3827–44 (2025), <https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag>.

(v) Agent and Tool Hijacking

A tool-integrated or agentic AI system can do more than generate text. Depending on its design and permissions, it may retrieve information from other systems or use tools to perform actions, such as sending a message, updating a record, or initiating a transaction.⁴⁰ This creates an additional risk from prompt injection. An adversarial input may cause the agent to pursue the attacker’s objective, select an inappropriate tool, or invoke an authorized tool in an unauthorized way.

The consequences depend on the tools and data available to the system. In a court, a compromised agent might improperly disclose sealed information, send an unauthorized communication, or alter scheduling or docket information. In an administrative agency, it might expose protected data, modify an application record, initiate a payment, or trigger another official process. These risks make access control particularly important: the system should receive only the permissions necessary for its assigned task, and requests to access protected resources should be checked before they are carried out.⁴¹

The distinction between advisory and operational systems is therefore legally significant. A flawed AI-generated summary may influence what an official later decides, but a compromised system with tool access may directly alter a record, disclose information, or trigger an institutional process. Systems with such capabilities should therefore operate with narrowly defined permissions and should require additional authorization before consequential actions are executed. They should also

⁴⁰ See Qiusi Zhan *et al.*, *InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated Large Language Model Agents*, in Findings of the Association for Computational Linguistics: ACL 2024 10471, 10471–506 (2024), <https://doi.org/10.18653/v1/2024.findings-acl.624>; Edoardo Debenedetti *et al.*, *AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents*, in Advances in Neural Information Processing Systems 37 (2024), <https://doi.org/10.52202/079017-2636>.

⁴¹ Jerome H. Saltzer & Michael D. Schroeder, *The Protection of Information in Computer Systems*, 63 Proc. IEEE 1278, 1282-88 (1975), <https://doi.org/10.1109/PROC.1975.9939>.

preserve records that distinguish what the AI proposed from what was actually carried out.

(vi) Model, Data, and Supply-Chain Attacks

Some attacks occur before an AI system processes any particular filing or submission. An attacker may, for example, manipulate data used to train or fine-tune a model, introduce a hidden backdoor that is activated by a particular input, or compromise a model or software component on which the institution relies.⁴² Risks can also arise when a provider changes the model, adds a new component, or deploys an update without adequate testing or institutional review.⁴³

These attacks can have consequences beyond a single case. A poisoned training dataset or compromised model may affect every matter in which the resulting system is used. Likewise, a backdoor may remain dormant until it encounters the input that activates it. Because the problem originates in the system itself rather than in an individual filing, institutions need controls over model versions and updates, testing before deployment, notification of significant provider changes, and the ability to return to a previously approved version when necessary.

These risks also complicate the question of responsibility. What appears to be an isolated AI error may instead result from compromised training data, a third-party component, a provider update, or the way the institution configured and integrated the system. A framework concerned only with misconduct by litigants or other submitters would therefore overlook vulnerabilities introduced earlier in the AI supply chain. At

⁴² See Micah Goldblum *et al.*, *Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses*, 45 IEEE Transactions on Pattern Analysis & Machine Intelligence 1563, 1563–80 (2023), <https://doi.org/10.1109/TPAMI.2022.3162397>; Tianyu Gu *et al.*, *BadNets: Evaluating Backdooring Attacks on Deep Neural Networks*, 7 IEEE Access 47230, 47230–44 (2019), <https://doi.org/10.1109/ACCESS.2019.2909068>; OMB Memorandum M-25-22, at 7–12.

⁴³ See Office of Management & Budget, Memorandum M-25-22, Driving Efficient Acquisition of Artificial Intelligence in Government 8–10 (Apr. 3, 2025).

the same time, focusing only on providers would overlook attacks introduced through filings and other materials that the institution asks the system to process.

(vii) Human-Layer and Governance Attacks

Some attacks exploit the way people interact with AI systems rather than altering the model itself. An attacker may, for example, conceal or suppress a warning, present AI-generated material as independently verified analysis, manipulate information about a source's provenance, or use compromised credentials to make an output appear authorized. The common feature is that the attack seeks to influence what an official sees, trusts, or treats as legitimate.

Interface design can affect whether such manipulation succeeds.⁴⁴ Officials often interact with an AI system through summaries, recommendations, warnings, source lists, and other interface elements rather than by inspecting the underlying process. Under time pressure, they may place excessive weight on a polished summary or default recommendation. Repeated false alarms can also reduce attention to later warnings, while an explanation may increase acceptance of an AI recommendation without necessarily improving the user's ability to determine whether the recommendation is correct.⁴⁵ Human review is therefore not an effective safeguard merely because a person remains formally involved. Its effectiveness depends on what information the reviewer receives and whether the interface enables meaningful verification.

⁴⁴ See Andrew D. Selbst et al., *Fairness and Abstraction in Sociotechnical Systems*, in *Proceedings of FAT 2019* 59, 59–68 (2019), <https://doi.org/10.1145/3287560.3287598>; Saleema Amershi et al., *Guidelines for Human-AI Interaction*, in *Proceedings of CHI 2019*, paper 3 (2019), <https://doi.org/10.1145/3290605.3300233>.

⁴⁵ Raja Parasuraman & Victor Riley, *Humans and Automation: Use, Misuse, Disuse, Abuse*, 39 Hum. Factors 230, 230-53 (1997), <https://doi.org/10.1518/001872097778543886>; John D. Lee & Katrina A. See, *Trust in Automation: Designing for Appropriate Reliance*, 46 Hum. Factors 50, 50-80 (2004), https://doi.org/10.1518/hfes.46.1.50_30392; Gagan Bansal et al., *Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance*, in *Proceedings of CHI 2021* (2021), <https://doi.org/10.1145/3411764.3445717>.

A related but distinct category concerns attacks on the records and controls needed to investigate system behavior. An adversary might alter logs, remove provenance information, substitute an unauthorized model version, or falsify audit evidence in order to conceal what occurred. Such conduct can make it difficult to determine what information the system processed, which version was operating, what output it produced, and whether a hostile input affected a legally relevant step. Similar evidentiary problems may also result from inadequate recordkeeping, incomplete incident reporting, or lack of provider cooperation, but those governance failures should be distinguished from deliberate attacks.

3.3 The distinctive role of Courts and Public Administration

Courts face a distinctive security problem because receiving information from outside actors is an essential part of adjudication. Parties are entitled to submit arguments and evidence and, in an adversarial system, have legitimate reasons to present that material in the manner most favorable to their position. Courts therefore cannot eliminate externally supplied content simply because it may also create risks for the AI systems that process it. From a security perspective, however, such content cannot be assumed to be trustworthy merely because it was submitted through an authorized procedural channel. Courts must manage that tension while also protecting sensitive information and operating under substantial workload and time constraints.⁴⁶

The consequences of failure may also be unusually serious. Judicial decisions can affect liberty, property, legal status, and other protected interests. At the same time,

⁴⁶ See A.D. Reiling, *Courts and Artificial Intelligence*, 11 Int'l J. Ct. Admin., art. 8 (2020), <https://doi.org/10.36745/ijca.343>; Mark Bovens & Stavros Zouridis, *From Street-Level to System-Level Bureaucracies*, 62 Pub. Admin. Rev. 174, 174-84 (2002), <https://doi.org/10.1111/0033-3352.00168>.

safeguards for AI systems must respect judicial independence. Technical governance should protect the integrity of judicial systems without giving executive authorities, vendors, or other outside actors control over adjudicative reasoning.

Public agencies face a similar problem in different settings. They routinely receive claims, applications, bids, public comments, regulatory submissions, and other information from individuals and organizations seeking a favorable governmental decision.⁴⁷ As agencies use AI not only to organize information but also to rank, classify, recommend, or initiate actions, the consequences of manipulating those systems become more significant. Institutional security therefore cannot depend solely on the assumption that outside participants will use authorized channels only for their intended purpose.

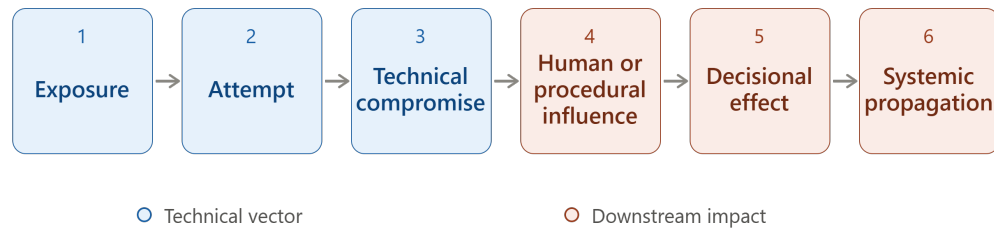
Table 1. Attack Types and Judicial or Administrative Effects

Attack type	Typical entry point or target	Potential institutional effect	Principal evidence or control
Direct prompt injection	Chat, API, or internal copilot	Policy evasion, unauthorized disclosure, distorted advice	Session logs, instruction hierarchy, access controls
Indirect prompt injection	Filing, exhibit, email, or webpage	Altered summaries, source use, or agent behavior	Original vs. parsed document, provenance, sandboxing
Representational attack	PDF layers, metadata, OCR, or Unicode	Human-machine mismatch; hidden machine-readable content	Canonical rendering, extraction comparison, parser testing
Retrieval poisoning	Case-law, policy, or knowledge corpus	False or obsolete authority; cross-case contamination	Source authentication, retrieval traces, corpus versioning
Agent or tool hijacking	Untrusted content reaching a tool-using system	Unauthorized action: communication, disclosure, record change, or payment	Least privilege, human confirmation, tool logs, case isolation
Model, data, or supply-chain attack	Training, fine-tuning, dependencies, or updates	Persistent or systemic compromise	Version controls, acceptance tests, vendor notices, rollback
Human or governance attack	Interface, alerts, accounts, or audit trails	Automation bias, concealed incidents, false attribution	Targeted alerts, access records, immutable logging, independent review

⁴⁷ See Bernd W. Wirtz, Jan C. Weyerer & Carolin Geyer, *Artificial Intelligence and the Public Sector – Applications and Challenges*, 42 Int’l J. Pub. Admin. 596, 596-615 (2019), <https://doi.org/10.1080/01900692.2018.1498103>; Anneke Zuiderwijk, Yu-Che Chen & Fadi Salem, *Implications of the Use of Artificial Intelligence in Public Governance*, 38 Gov’t Info. Q. 101577 (2021), <https://doi.org/10.1016/j.giq.2021.101577>.

The consequences of an adversarial attack depend on how far the attack progresses and how widely its effects spread. Figure 1 distinguishes five stages of escalation and a separate dimension of systemic propagation.

Figure 1. Stages of adversarial escalation.



(1) Exposure – the system is capable of receiving or reaching potentially hostile content; (2) Attempt – an actor introduces a malicious input or otherwise seeks to exploit the system; (3) Technical compromise – the attack changes the system’s processing, output, source selection, or use of tools; (4) Human or procedural effect – the compromised system changes what an official sees or affects a legally relevant procedural step; (5) Decisional effect – the compromise contributes to a governmental decision, its reasoning, timing, or implementation. Systemic propagation describes the additional situation in which the same vulnerability, attack, or compromised resource affects multiple matters, users, systems, or institutions.

Exposure exists when the design of a workflow makes an AI system reachable by potentially hostile material – for example, when a court asks the system to process documents submitted by parties. An attempt occurs when an actor introduces content designed to exploit that exposure. Technical compromise requires more: the hostile input must affect the system’s operation, such as by changing an output, altering the sources retrieved, or inducing an unauthorized tool action. A human or procedural effect arises when that technical compromise changes what an official sees, relies upon, omits, or does, or otherwise alters a legally relevant procedural step. A

decisional effect occurs when the compromise contributes to the content, reasoning, timing, or implementation of a public decision.

These stages are analytically distinct, but an investigation may not be able to prove each one separately or reconstruct them in sequence. An unauthorized payment, for example, may establish that an operational system was compromised even if the model's internal processing cannot be reconstructed. A comparison between an affected summary and a clean version may show that the output changed without revealing every intermediate model state. Evidence should therefore be directed to the effect that matters for the legal question at issue rather than to reconstructing every internal computation.

Systemic propagation concerns scale rather than a further step in the causal sequence. A compromised shared source, model update, or component may affect many matters even before a human or decisional effect has been established in any particular case. Conversely, an attack may affect a decision in one proceeding without spreading beyond that proceeding.

The framework therefore separates three questions that should not be collapsed. Participant sanctions concern the actor's conduct, knowledge, purpose, and procedural responsibility and may arise even when an attempted attack is blocked. Relief for an affected party depends on whether the attack altered a legally relevant procedure or decision under the applicable remedial standard. Institutional remediation may be warranted earlier, including at the exposure stage, because security measures are designed to reduce prospective and system-wide risk. The comparative analysis that follows examines how the relevant legal regimes address these different questions.

4. A Comparative Analysis of the EU and US Frameworks

4.1 The European Union Framework

At the EU level, the AI Act provides the principal framework for regulating the design, deployment, and oversight of AI systems that may be exposed to adversarial inputs. Whether its obligations apply depends on the Act's territorial scope, the system's intended purpose and classification, and the date on which the relevant provisions become applicable. Where a system qualifies as high-risk, the Act distributes responsibilities between providers and deployers. Questions concerning misconduct by litigants, lawyers, or other participants, as well as remedies in individual proceedings, remain largely governed by national procedural and professional law.

(i) Classifying AI-Assisted Adjudication

1. Territorial Scope and Counterfactual Method

The AI Act applies through the territorial rules in Article 2, including to providers placing systems or general-purpose models on the Union market, deployers established or located in the Union, and specified actors outside the Union when system output is used there. Functional similarity does not extend the Regulation to a Brazilian court. As mentioned *supra*, the relevant question is mostly counterfactual: had an equivalent system been placed on the Union market, put into service, or deployed by an EU judicial authority, which duties would follow from its intended purpose, regulated role, and date of use?

Regulation 2024/1689 entered into force on August 1, 2024. Its prohibitions and AI-literacy rules began applying on February 2, 2025, followed principally by governance and general-purpose-model provisions on August 2, 2025. The original Article 113

placed the principal Annex III duties after the May 12, 2026 Brazilian judgment.⁴⁸ Regulation (EU) 2026/1744, published July 24 and scheduled to enter into force July 27, moves those duties to December 2, 2027 for Annex III systems and August 2, 2028 for specified Article 6(1) product systems.⁴⁹

2. Functional Classification Under Article 6 and Annex III

Article 6(2) and Annex III, paragraph 8(a), classify as high-risk systems intended for use by or on behalf of a judicial authority to assist in researching and interpreting facts and law or applying law to concrete facts.⁵⁰ Recital 61 links that category to fair-trial and rule-of-law interests, while excluding purely ancillary administration that does not affect justice in an individual case.⁵¹ Classification should therefore follow function rather than the label “court AI.” The decisive features are whether the system selects legally salient information, researches or assesses law or fact, recommends or drafts a disposition, possesses operational authority, or in practice supplies the principal representation on which a judge relies.⁵²

Galileu’s documented support could bring an equivalent EU system within Annex III, but intended-purpose documentation and workflow evidence are needed for a definitive classification. A component confined to detecting suspicious hidden content may remain a narrow preparatory or security function; a system whose summary or draft materially frames adjudication points toward high-risk judicial assistance. Article 6(3) reinforces that distinction by excluding an Annex III use that poses no significant

⁴⁸ EU AI Act art. 113.

⁴⁹ Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 Amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as Regards the Simplification of the Implementation of Harmonised Rules on Artificial Intelligence (Digital Omnibus on AI), arts. 1(40), 2, 2026 O.J. (L 2026/1744) (published July 24, 2026; scheduled to enter into force July 27, 2026).

⁵⁰ EU AI Act art. 6(1)-(2) & annex III, para. 8(a).

⁵¹ *Id.* rec. 61.

⁵² *See* EU AI Act art. 6(2)-(4), annex III, para. 8(a), & rec. 61.

risk and does not materially influence decision-making, while requiring a documented assessment and preserving high-risk status for profiling systems.⁵³⁵⁴

3. Application Guidance and Standardization

The materials used to implement the AI Act do not all have the same legal status. In May 2026, the Commission published draft guidelines on the classification of high-risk AI systems under Article 6. The guidelines provide practical examples, including examples concerning judicial AI, but they are non-binding and remain in draft form following the close of the stakeholder consultation on July 23, 2026.⁵⁵

Harmonized standards perform a different function. They are intended to translate the AI Act's requirements into more detailed technical specifications. Their use is voluntary, but compliance with a harmonized standard that has been referenced in the Official Journal may give a provider a presumption of conformity with the requirements covered by that standard. The development of harmonized standards for the AI Act remains ongoing.⁵⁶

(ii) Applying the High-Risk System Duties to Adversarial Inputs

Articles 9 and 12-15 coordinate lifecycle risk management, documentation and logs, deployer transparency, human control, and cybersecurity. Adversarial filings test

⁵³ Id. art. 6(3)–(4); see also id. art. 49(2). Regulation (EU) 2026/1744 modifies aspects of the registration framework; on the Article's current-law date, it had been published but was not yet in force.

⁵⁴ Article 6(3) turns on significant risk and material influence, and requires documented assessment. EU AI Act art. 6(3)–(4).

⁵⁵ See European Commission, *Draft Guidelines on the Classification of High-Risk AI Systems Under Regulation (EU) 2024/1689* (May 19, 2026) (nonbinding consultation draft; consultation closed July 23, 2026), <https://digital-strategy.ec.europa.eu/en/consultations/targeted-consultation-draft-guidelines-classification-high-risk-artificial-intelligence-systems>.

⁵⁶ European Commission, *Navigating the AI Act* (last visited July 25, 2026), <https://digital-strategy.ec.europa.eu/en/faqs/navigating-ai-act>; Amedeo Rizzo & Giorgio Hassan, *The U.S. and E.U. Framework of AI Standardization: Building a Way Forward*, Stanford-Vienna Transatlantic Technology Law Forum Working Paper No. 137, at 14-19, 35-44 (2025), <https://law.stanford.edu/wp-content/uploads/2025/10/TTLF-WP-137-Rizzo-Hassan.pdf>.

whether those duties preserve the boundary between authenticated instructions and untrusted content.

Article 9 requires a continuous, iterative risk-management system that identifies known and reasonably foreseeable risks, evaluates intended use and foreseeable misuse, incorporates post-market information, tests controls, and reduces residual risk to an acceptable level.⁵⁷ For judicial AI that reads party submissions, concealed machine-directed content and indirect prompt injection should qualify as foreseeable misuse. The conclusion rests on the institutional environment: adverse actors are authorized to submit documents, and an LLM application may interpret language carried as data as if it were an instruction.⁵⁸

The provision does not prescribe a particular technical control for adversarial inputs. Instead, it requires a continuous risk-management process throughout the lifecycle of a high-risk AI system. Providers must identify and evaluate known and reasonably foreseeable risks, including those arising from reasonably foreseeable misuse, adopt appropriate measures to reduce those risks, assess the remaining risk, and test the system against criteria appropriate to its intended purpose.⁵⁹ Applied to adversarial inputs, those requirements support testing against realistic attack scenarios and updating safeguards when experience reveals previously unidentified or insufficiently controlled attack paths. A detected and blocked prompt-injection attempt may therefore be relevant to risk management even if it does not amount to a reportable serious incident. Article 9 requires reasonable mitigation of foreseeable risks and an

⁵⁷ EU AI Act art. 9(1)-(9).

⁵⁸ EU AI Act art. 9(2)(a)-(b); Kai Greshake 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 79, 79-90 (2023), <https://doi.org/10.1145/3605764.3623985>.

⁵⁹ See EU AI Act art. 9(3)-(7) (testing and risk-reduction measures throughout development and before deployment).

acceptable level of residual risk; it does not require proof that every adaptive attack can be prevented.

Article 10 regulates the data used to train, validate, and test high-risk AI systems. It requires appropriate data-governance practices concerning matters such as the origin and preparation of the data, their suitability for the system's intended purpose, relevant gaps or biases, and the representativeness, completeness, and accuracy of the datasets.

Article 10 regulates the data used to train, validate, and test high-risk AI systems. It requires appropriate data-governance practices concerning matters such as the origin and preparation of the data, their suitability for the system's intended purpose, relevant gaps or biases, and the representativeness, completeness, and accuracy of the datasets.⁶⁰

Article 11 requires providers to prepare and keep up to date technical documentation demonstrating compliance with the high-risk requirements. Annex IV specifies information concerning, among other matters, the system's intended purpose and versions, its architecture and interaction with other systems, the data used in its development and testing, performance and limitations, human-oversight measures, cybersecurity controls, relevant changes during the system's lifecycle, and post-market monitoring⁶¹. In the context of adversarial filings, that documentation should be sufficiently detailed to explain how the system processes external documents, how it distinguishes trusted instructions from untrusted content, what retrieval or tool functions it can use, what safeguards and alerts apply, and how the institution can suspend or safely reprocess an affected matter. These details are not expressly listed in

⁶⁰ See EU AI Act art. 10(2)-(5); Micah Goldblum et al., Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses, 45 IEEE Transactions on Pattern Analysis & Machine Intelligence 1563, 1563-80 (2023), <https://doi.org/10.1109/TPAMI.2022.3162397>.

⁶¹ See EU AI Act art. 11 & annex IV (requiring descriptions of architecture, data, monitoring, performance, limitations, human oversight, cybersecurity, and lifecycle changes).

Annex IV in those terms, but they constitute applications of its broader requirements to the particular risks created by adversarial inputs.

Article 12 requires high-risk AI systems to record events automatically so that their operation can be traced and monitored.⁶² Articles 19 and 26 address a separate question: how long providers and deployers must retain the logs under their control.⁶³ For an adversarial filing, the records should help an investigator follow the document through the system. They should show what the system read, which version processed it, and whether the resulting output or tool use changed. Warnings and the official's response may also be relevant. This does not require a record of every internal computation. Where the records contain personal data, the GDPR continues to govern their purpose, security, and retention.⁶⁴

Article 13 requires information that enables deployers to understand a high-risk system's outputs and use them appropriately.⁶⁵ The accompanying instructions must explain relevant capabilities and limitations.⁶⁶ For a court, this should include an explanation of what happens when the system detects suspicious content. An alert is of limited use if the reviewer cannot tell whether the content was blocked or remained available to the model, or how to stop processing and repeat the task safely.⁶⁷ A general warning that AI can make mistakes would not explain that particular security limitation.⁶⁸

⁶² *Id.* art. 12(1)-(2). Specific biometric systems have additional requirements not material here.

⁶³ *Id.* arts. 19(1), 26(6).

⁶⁴ *See* Regulation (EU) 2016/679, arts. 5(1)(c), (e), 24, 25, 32, 2016 O.J. (L 119) 1 (data minimization, storage limitation, responsibility, data protection by design, and security).

⁶⁵ EU AI Act art. 13(2)-(3).

⁶⁶ *Id.* art. 13(1)-(3).

⁶⁷ *See* EU AI Act art. 13(3)(b)-(f) (requiring information on performance, limitations, foreseeable circumstances, oversight, input specifications, and log interpretation).

⁶⁸ EU AI Act art. 13(1)-(3).

Article 14 requires high-risk systems to be designed so that people can oversee them effectively, using measures suited to the risks and context of use.⁶⁹ Reviewers must be able to understand the system's limitations and intervene, including by disregarding an output or stopping the system.⁷⁰ Applied to an adversarial filing, this means giving the reviewer access to the suspicious passage and enough information to assess its possible effect. A reliable alternative process should be available when the output cannot be trusted. The fact that a judge signs the final document does not, by itself, show that this review was possible or effective.⁷¹

Article 15 requires appropriate accuracy, robustness, and cybersecurity throughout a high-risk system's lifecycle. It expressly addresses unauthorized attempts to alter the system's use, output, or performance, including attacks on its data and model.⁷² Although the provision does not name prompt injection, its terms provide a basis for treating a concealed instruction in a lawful filing as a cybersecurity threat. The relevant problem is that the system may follow the filer's instruction instead of the institution's. This is an application of the provision to adversarial filings, not a settled judicial interpretation.⁷³ Testing should accordingly reflect the ways attackers can disguise or vary an instruction, rather than focus on a single familiar phrase.⁷⁴

⁶⁹ EU AI Act art. 14(3).

⁷⁰ *Id.* art. 14(1)-(4).

⁷¹ Raja Parasuraman & Victor Riley, Humans and Automation: Use, Misuse, Disuse, Abuse, 39 Human Factors 230, 230-53 (1997), <https://doi.org/10.1518/001872097778543886>; John D. Lee & Katrina A. See, Trust in Automation: Designing for Appropriate Reliance, 46 Human Factors 50, 50-80 (2004), https://doi.org/10.1518/hfes.46.1.50_30392; Ben Green & Yiling Chen, The Principles and Limits of Algorithm-in-the-Loop Decision Making, 3 Proceedings of the ACM on Human-Computer Interaction, art. 50, at 2-19 (2019), <https://doi.org/10.1145/3359152>.

⁷² EU AI Act art. 15(1), (4)-(5).

⁷³ Article 15(5) refers broadly to unauthorized third-party attempts to alter use, outputs, or performance and specifically lists several attack families, but not prompt injection by name. EU AI Act art. 15(5).

⁷⁴ Yupei Liu et al., Formalizing and Benchmarking Prompt Injection Attacks and Defenses, in 33d USENIX Security Symposium (2024), <https://www.usenix.org/conference/usenixsecurity24/presentation/liu-yupei>; Edoardo DeBenedetti et al., AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents, in Advances in Neural Information Processing Systems 37 (2024), <https://doi.org/10.52202/079017-2636>.

(iii) Allocating Lifecycle Responsibility for Secure Public-Sector Deployment

The high-risk regime assigns responsibilities to both those who supply an AI system and those who use it. Providers are responsible for the system's design and conformity; deployers are responsible for its use within the institution. Information gathered during use connects those roles by showing when the system or its safeguards need to change. Articles 16 and 17 require providers to ensure conformity and maintain a documented quality-management system. The purpose is to make compliance part of how the provider develops, tests, and manages the system, rather than a check performed only at launch.⁷⁵ Articles 18-21 add obligations to retain relevant documentation and logs, cooperate with authorities, and take corrective action. A provider that identifies non-conformity must take the appropriate steps to address it, which may require withdrawing or recalling the system rather than continuing its use unchanged.⁷⁶ Article 26 places responsibility for the institutional use of a high-risk system on the deployer. It addresses use according to instructions, effective human oversight, and monitoring, with duties concerning input data, records, notification, and suspension in the circumstances specified. Article 27 separately requires certain deployers to assess the impact on fundamental rights.⁷⁷ These responsibilities matter when a court buys an AI service. A contract cannot remove the court's duties concerning its own use of the system. Equally, the court may be unable to investigate a defect without technical information and assistance from the provider.⁷⁸

⁷⁵ EU AI Act arts. 16-17.

⁷⁶ *Id.* arts. 18-21.

⁷⁷ *Id.* arts. 26-27.

⁷⁸ Compare EU AI Act arts. 16-21 (provider duties), with art. 26 (deployer duties).

Article 72 requires providers to collect and analyze information about a high-risk system's performance throughout its lifetime.⁷⁹ Article 73 addresses a narrower question: when an incident must be reported to the competent authority as a “serious incident” under Article 3(49).⁸⁰ The two processes should not be confused. A blocked prompt-injection attempt may reveal a weakness that warrants investigation and revised safeguards without meeting the statutory reporting threshold. Conversely, an incident affecting fair-trial rights or spreading through a shared source requires careful assessment against that definition. Internal investigation should not wait until external reporting is legally required.⁸¹

Public reporting serves a further purpose. An institution can explain the kinds of incidents it has encountered and the measures taken in response without disclosing confidential case material or technical details that would assist an attacker. Such reporting complements, rather than replaces, the internal investigation and any mandatory notification.

Similar questions arise outside the courts, although the applicable high-risk category may differ. A tax authority might use AI to select audits or interpret complex anti-abuse rules. Those functions can influence how taxpayers are treated without making the system “judicial AI.” The relevant inquiry remains what the system is intended to do and which legal requirements apply. In this setting, the AI Act and the GDPR need to be considered together rather than treated as separate accounts of taxpayer protection.⁸²

⁷⁹ EU AI Act art. 72(1)-(2).

⁸⁰ *Id.* arts. 3(49), 73.

⁸¹ See EU AI Act arts. 20, 26(5), 72. Mandatory authority reporting under Article 73 separately depends on a “serious incident” within Article 3(49).

⁸² Amedeo Rizzo & Giorgio Hassan, Addressing the Use of AI by EU Tax Authorities: Towards a Common Framework of Taxpayer Protection, 65 *European Taxation* 15, 15-29 (2025), <https://doi.org/10.59403/lavybfj>; Amedeo Rizzo & Giorgio Hassan, Proportionate Risk Management

Contracts help institutions put these responsibilities into practice. A court needs to know when its provider changes the system and to obtain the information needed to investigate an incident. It also needs a workable means of suspending the service or moving to an alternative. Harmonized standards can give more detailed technical content to the Act's requirements, but a presumption of conformity does not establish that a particular deployment was secure in every respect.

(iv) Procedural Limits and Regulatory Gaps

The AI Act does not provide a complete legal response to an adversarial filing. Its principal concern is the conduct of providers and deployers. Questions about the submitter's responsibility and the consequences for an individual proceeding still require national procedural and professional rules, alongside other applicable law.

Even where the high-risk requirements apply, the Act does not itself determine whether a lawyer knowingly inserted a concealed instruction or whether the court should strike the pleading. Nor does it decide whether an affected judgment should be reconsidered or set aside. Those questions depend on the applicable national rules, including their standards of culpability and prejudice. Where Union law is engaged, those rules remain subject to the principles of effectiveness and equivalence and to the Charter's fair-trial guarantees.⁸³ Compliance with system requirements does not excuse deliberate manipulation; proof of manipulation does not establish that the institution secured its system adequately.

for Tax-Related AI Systems: Analysis and Implications Under EU and US Law, 18 World Tax Journal, no. 1 (2026), <https://doi.org/10.59403/560kfb>.

⁸³ See Charter of Fundamental Rights of the European Union arts. 47-48, 2012 O.J. (C 326) 391; Case 33/76, *Rewe-Zentralfinanz eG v. Landwirtschaftskammer für das Saarland*, ECLI:EU:C:1976:188, paras. 5-6; Case C-432/05, *Unibet (London) Ltd. v. Justitiekanslern*, ECLI:EU:C:2007:163, paras. 37-44.

Security measures must also respect judicial independence and data protection. Controls over access and system changes can help courts retain control of their technology. By contrast, oversight that gives outsiders access to judicial deliberations or control over legal reasoning raises a different concern. The GDPR continues to apply where personal data are processed, including through its requirements concerning lawful processing, security, and minimization.⁸⁴ Judicially supervised audits and restricted access to sensitive records offer ways to investigate technical failures without opening adjudication itself to outside control.

4.2 United States Law

United States law approaches adversarial filings through several bodies of law rather than a single AI-specific regime. Courts can regulate the conduct of litigants and lawyers and correct affected proceedings. The security of the court's own AI system is governed separately, through judicial administration, procurement, and applicable technical standards. Executive-branch AI policies provide another model, but their reach must not be confused with that of rules governing the courts.

(i) Participant Misconduct and Judicial Authority

A suspicious filing may reflect deliberate manipulation, inadequate supervision, or an innocent technical feature. The available legal response depends on which of those explanations the evidence supports. Although filing rules and the court's inherent authority may overlap, they impose different requirements and permit different sanctions.

⁸⁴ Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, arts. 5, 22, 24, 25, 32, 35, 2016 O.J. (L 119) 1.

1. Rule 11 and Filing-Based Sanctions

Federal Rule of Civil Procedure 11 requires an attorney or unrepresented party presenting a paper to make a reasonable inquiry into its basis. The person certifies that the paper is not being presented for an improper purpose and that its legal and factual contentions have the support required by the Rule.⁸⁵ A concealed instruction intended to override the court's AI system could support an improper-purpose finding. That conclusion would still depend on the facts. Where the hidden content also misrepresents evidence or legal authority, the Rule's requirements concerning those representations may be relevant as well.

Rule 11 is not a general rule for every judicial or administrative submission. It governs papers in federal civil district-court proceedings and excludes discovery matters covered by Rules 26-37.⁸⁶ A party seeking sanctions must ordinarily allow the opposing party twenty-one days to withdraw or correct the challenged paper before filing the motion. A court acting on its own initiative instead proceeds through an order to show cause. Sanctions must serve deterrence, and the Rule places additional limits on monetary relief.⁸⁷ In either setting, notice and an opportunity to respond are essential, as is evidence linking the conduct to the person being sanctioned.⁸⁸ A blocked attack may therefore justify a sanction without showing that the court's consideration of the merits was affected.

⁸⁵ Fed. R. Civ. P. 11(b). See *Business Guides, Inc. v. Chromatic Communications Enterprises, Inc.*, 498 U.S. 533, 542-51 (1991) (applying objective reasonable-inquiry obligations); *Cooter & Gell v. Hartmarx Corp.*, 496 U.S. 384, 393-409 (1990) (describing Rule 11's deterrent purpose and collateral nature).

⁸⁶ See Fed. R. Civ. P. 1, 11; Fed. R. App. P. 38, 46(c); Fed. R. Bankr. P. 9011; Fed. R. Crim. P. 57.

⁸⁷ Fed. R. Civ. P. 11(c)(2)-(5).

⁸⁸ Rule 11 applies an objective reasonable-inquiry standard and requires notice and a reasonable opportunity to respond. Fed. R. Civ. P. 11(b), (c)(1); *Business Guides*, 498 U.S. at 542-51.

2. Inherent Power, Fraud on the Court, and Contempt

Federal courts also have inherent authority to manage their proceedings and sanction bad-faith abuse. That authority can reach conduct not fully addressed by a specific rule, but *Chambers* and *Roadway Express* require restraint, adequate findings, and an opportunity to be heard.⁸⁹ In an adversarial-filing incident, the court might first preserve the original document and require a clean replacement. A limited technical investigation could then establish what occurred while protecting sensitive exploit details. Any sanction for bad faith would require a supported finding, rather than follow automatically from the presence of suspicious content.

The purpose of the sanction also matters. Under *Goodyear*, a compensatory award ordinarily must be limited to costs that would not have been incurred but for the misconduct.⁹⁰ A punitive sanction raises the separate procedural protections associated with criminal contempt.⁹¹ Fraud on the court is narrower still: it concerns serious corruption of the adjudicative process, not every misleading submission or procedural violation.⁹²

3. Obstruction, Fraud, and Unauthorized Access

An adversarial filing is not automatically a criminal offense. Contempt under 18 U.S.C. § 401 requires the specified misconduct or disobedience, and civil and criminal contempt serve different purposes.⁹³ Obstruction provisions have their own requirements concerning intent and connection to a proceeding; *Arthur Andersen*

⁸⁹ *Chambers v. NASCO, Inc.*, 501 U.S. 32, 43-51 (1991); *Roadway Express, Inc. v. Piper*, 447 U.S. 752, 764-67 (1980).

⁹⁰ *Goodyear Tire & Rubber Co. v. Haeger*, 581 U.S. 101, 108-12 (2017).

⁹¹ *International Union, United Mine Workers of America v. Bagwell*, 512 U.S. 821, 827-38 (1994); Fed. R. Crim. P. 42.

⁹² See *Hazel-Atlas Glass Co. v. Hartford-Empire Co.*, 322 U.S. 238, 245-46 (1944); *United States v. Shaffer Equipment Co.*, 11 F.3d 450, 462-63 (4th Cir. 1993).

⁹³ 18 U.S.C. § 401; *International Union, United Mine Workers of America v. Bagwell*, 512 U.S. 821, 827-38 (1994).

emphasizes the need to establish conscious wrongdoing under the provision at issue.⁹⁴

Other statutes have important limits. Section 1001 excludes qualifying submissions by parties or counsel to a judge or magistrate, while wire fraud requires a scheme directed at money or traditional property. The technical description “prompt injection” does not establish those elements.⁹⁵

The same caution applies to the Computer Fraud and Abuse Act. Van Buren distinguishes obtaining information one is not authorized to access from using accessible information for an improper purpose.⁹⁶ A litigant may misuse a permitted filing channel without thereby obtaining unauthorized access to the court’s computer system. Bypassing authentication or entering a restricted part of that system would raise a different question. The legal analysis must identify the access involved rather than infer a computer-access offense from the presence of a hostile instruction.

(ii) Extension of Professional and Evidentiary Obligations to Machine-Directed Content

Professional and evidentiary rules address different parts of the problem. Professional duties concern what lawyers must verify and disclose. Evidence and discovery rules help establish what a document contained and what the system did with it. Procedural remedies address any resulting effect on the case. None of these functions replaces the institution’s responsibility for securing its own technology.

1. Professional Responsibility, Attribution, and Mens Rea

The Model Rules connect technological competence with established duties of candor, fairness, and supervision. Rule 1.1 and Comment 8 address relevant technological

⁹⁴ 18 U.S.C. §§ 1503, 1512; *Arthur Andersen LLP v. United States*, 544 U.S. 696, 703-08 (2005).

⁹⁵ 18 U.S.C. § 1001(b) (excluding from subsection (a), subject to specified limits, statements, representations, writings, or documents submitted by a party or that party’s counsel to a judge or magistrate in a judicial proceeding); 18 U.S.C. § 1343; *Ciminelli v. United States*, 598 U.S. 306, 309, 316 (2023) (limiting federal wire fraud to schemes targeting traditional property interests).

⁹⁶ 18 U.S.C. § 1030; *Van Buren v. United States*, 593 U.S. 374, 384-96 (2021).

benefits and risks; the related rules govern how lawyers present work to a court and supervise those who assist them.⁹⁷ ABA Formal Opinion 512 applies those responsibilities to generative AI without transferring them from the lawyer to the tool.⁹⁸ The distinction from *Mata v. Avianca* is important.⁹⁹ That case concerned false authorities and failures to correct them. An adversarial filing presents a different possibility: its visible legal content may be accurate while a hidden instruction attempts to control the court's AI system.

Responsibility should therefore follow the evidence of what the lawyer knew and did. Intentionally inserting an instruction is different from failing to supervise a document service, and both differ from unknowingly submitting a file affected by malware. Once a problem is discovered, duties to preserve, investigate, or correct the submission may arise.¹⁰⁰ A signature or account record can help identify who presented the file, but does not necessarily identify who created its hidden content. A supervisory failure may support responsibility without proving authorship; a finding of intentional dishonesty requires evidence appropriate to that finding.

2. Local Filing Rules and Judicial Administration

Local rules and orders take different approaches to AI-assisted filings. The cited Texas measures emphasize verification and responsibility for accuracy, with some requiring disclosure of AI assistance.¹⁰¹ Their focus is principally on what lawyers submit, including invented authorities and undisclosed assistance. They do not supply a complete response to instructions aimed at the court's own AI system or specify how

⁹⁷ Model Rules of Pro. Conduct rr. 1.1 cmt. 8, 3.3, 3.4, 5.1, 5.3, 8.4(c) (Am. Bar Ass'n 2024).

⁹⁸ ABA Comm. on Ethics & Pro. Resp., Formal Op. 512, *Generative Artificial Intelligence Tools* 2-15 (July 29, 2024).

⁹⁹ *Mata v. Avianca, Inc.*, 678 F. Supp. 3d 443, 448-66 (S.D.N.Y. 2023).

¹⁰⁰ See Model Rules of Pro. Conduct rr. 3.3(a)(1), (c), 3.4(a), 8.4(c) (Am. Bar Ass'n 2024); ABA Formal Op. 512, at 8-12.

¹⁰¹ E.D. Tex. Local R. CV-11(g) (2026); N.D. Tex. Local Civ. R. 7.2(f) (2025); S.D. Tex. Gen. Order 2025-04, *Use of Generative Artificial Intelligence in Court Filings* (2025).

that system should inspect documents and respond to attacks. These measures illustrate local experimentation rather than a uniform national rule for adversarial filings.¹⁰²

3. Evidence, Discovery, and Procedural Remedies

Authentication does not answer every question raised by a machine-directed filing. Rule 901 asks whether an item is what its proponent claims it to be.¹⁰³ Rule 901(b)(9) addresses evidence about a process or system and the accuracy of its results.¹⁰⁴ Rules 902(13) and 902(14) allow specified electronic records and copies to be authenticated through certification.¹⁰⁵ A file may satisfy those requirements and still contain an unauthorized instruction. The inquiry must therefore distinguish evidence generated or altered by AI from an authentic file that also directs a machine to act. Establishing the file's identity does not, by itself, establish whether the instruction was authorized, followed, or relied upon.

Civil discovery may help reconstruct the incident through relevant files and records within a party's possession, custody, or control, subject to proportionality and other applicable limits.¹⁰⁶ The original document can be compared with the version the AI processed, while logs may show whether an output or tool action changed. Rule 37(e) separately addresses the loss of electronically stored information and distinguishes remedies according to prejudice and intent.¹⁰⁷ The May 2026 committee materials

¹⁰² For a broader collection of federal and state judicial responses to generative AI, see Advisory Comm. on Evidence Rules, *May 2026 Agenda Book* 400-61 (May 7, 2026).

¹⁰³ Fed. R. Evid. 901(a); see also Fed. R. Evid. 901 advisory committee's note to 1972 proposed rules.

¹⁰⁴ Fed. R. Evid. 901(a), 901(b)(9).

¹⁰⁵ Fed. R. Evid. 902(13)-(14).

¹⁰⁶ Fed. R. Civ. P. 26(b)(1), 34(a)(1)(A).

¹⁰⁷ Fed. R. Civ. P. 26(b)(1), 34(a), 37(e); see *Lorraine v. Markel American Insurance Co.*, 241 F.R.D. 534, 538-59 (D. Md. 2007).

cited here did not replace that framework: proposed Rule 707 and draft Rule 901(c) remained under study rather than becoming operative rules.¹⁰⁸

The appropriate remedy depends on what the attack affected. Detection before substantive processing generally calls for containment, investigation, and any justified participant sanction rather than repetition of the proceeding. Where an output may have changed, the institution may need to repeat the affected task through a trusted process. If an official relied on compromised material, further relief may be available under the rules governing reconsideration, appeal, or setting aside a judgment. The applicable standard of prejudice remains important. An attempted attack alone does not establish a defective decision, while reliable evidence of containment may support a finding that the error was harmless.¹⁰⁹

(iii) Institutional Duties Through Judicial Governance and Cybersecurity Standards

The governance of court technology rests primarily with the judiciary itself. Federal guidance sets expectations for responsible use, while NIST materials offer methods for assessing and responding to technical risks. For adversarial filings, the question is how those general principles are translated into working safeguards in an individual court system.

1. Federal Judicial Guidance and Human Responsibility

The Administrative Office's 2025 Annual Report describes guidance developed through the judiciary's AI Task Force. It cautions against delegating core adjudicative

¹⁰⁸ Advisory Comm. on Evidence Rules, *Report to the Standing Committee* 12–16 (May 17, 2026) (declining to recommend proposed Rule 707 for final action and continuing Rule 707 and draft Rule 901(c) for further study); Advisory Comm. on Evidence Rules, *Preliminary Draft of Proposed Amendments to the Federal Rules of Appellate, Bankruptcy, Civil, Criminal, and Evidence Procedure* 181–87 (Aug. 2025).

¹⁰⁹ Cf. Fed. R. Civ. P. 59, 60(b)(3), 60(d)(3), 61; 28 U.S.C. § 2111. The applicable standard depends on forum, posture, misconduct, prejudice, and the remedy sought.

functions and places responsibility for verification and the protection of sensitive information on judges and staff.¹¹⁰ The judiciary’s fiscal year 2026 information-technology plan addresses the security and availability of court systems and data more generally.¹¹¹ These materials establish institutional expectations; they do not, by themselves, create a general private cause of action for a person affected by an AI failure.¹¹²

Verification nevertheless requires more than an instruction to review the result. A judge must be able to identify the material the system processed and compare the output with its sources. If the system can take action, the reviewer also needs to know what it actually did. The high-level guidance discussed here does not itself settle how to test concealed instructions, investigate an incident, or obtain essential records from a provider. Judicial governance can address those practical questions without transferring control of adjudication to a vendor or executive authority.

2. NIST Risk Management and Incident Response

NIST’s AI Risk Management Framework and Generative AI Profile provide a basis for identifying and managing risks. Its adversarial-machine-learning taxonomy explains attack mechanisms, while its incident-response guidance addresses what organizations should do when security fails.¹¹³ Courts can draw on these materials in policy,

¹¹⁰ Admin. Off. of the U.S. Cts., *2025 Annual Report of the Director: Court Operations—Artificial Intelligence* (2026).

¹¹¹ Jud. Conf. of the U.S., *Long Range Plan for Information Technology in the Federal Judiciary: Fiscal Year 2026 Update* 1-3, 16-20 (2025).

¹¹² See Admin. Off. of the U.S. Cts., *2025 Annual Report of the Director: Court Operations—Artificial Intelligence* (2026).

¹¹³ Nat’l Inst. of Standards & Tech., NIST AI 100-1, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* 20-36 (2023), <https://doi.org/10.6028/NIST.AI.100-1>; Nat’l Inst. of Standards & Tech., NIST AI 600-1, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (2024), <https://doi.org/10.6028/NIST.AI.600-1>; Apostol Vassilev et al., Nat’l Inst. of Standards & Tech., NIST AI 100-2e2025, *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations* 1-8, 15-25, 81-96 (2025), <https://doi.org/10.6028/NIST.AI.100-2e2025>; Nat’l Inst. of Standards & Tech., NIST SP 800-61 Rev. 3, *Incident Response Recommendations*

procurement, and technical assessment. The materials do not themselves create a private right or impose an identical obligation on every court.¹¹⁴

Their application must reflect the court’s actual system. A tool that summarizes filings presents different risks from one that can retrieve sealed records or send official communications. Referring to a general framework does not show that those particular risks were tested or that the court could recover from an attack. Common technical principles still need to be translated into controls that fit the institution’s work.

(iv) The Limits of the Decentralized American Model

Executive agencies operate under a different governance structure from Article III courts. OMB memoranda address agency use and acquisition of AI, while contracts give practical effect to some provider obligations. Their relevance to judicial AI is comparative: the measures illustrate an approach to prospective governance, not rules that automatically bind the courts.

1. OMB Governance for High-Impact AI

OMB Memorandum M-25-21 replaced M-24-10 and sets requirements for covered agencies’ use of high-impact AI. It combines assigned responsibility with assessment before deployment and monitoring during use. Where the required safeguards cannot be met, its framework provides for discontinuing the use, subject to its terms and exceptions.¹¹⁵ Memorandum M-25-22 addresses acquisition. It emphasizes testing and access to performance information, alongside competition and the ability to move

and Considerations for Cybersecurity Risk Management: A CSF 2.0 Community Profile (2025), <https://doi.org/10.6028/NIST.SP.800-61r3>.

¹¹⁴ See NIST AI 100-1, *AI RMF 1.0*, at 20–36; NIST AI 100-2e2025, *Adversarial Machine Learning*, at 1–8; NIST SP 800-61 Rev. 3, *Incident Response Recommendations*.

¹¹⁵ Off. of Mgmt. & Budget, Exec. Off. of the President, Memorandum M-25-21, *Accelerating Federal Use of AI Through Innovation, Governance, and Public Trust* 2-3, 14-24 (Apr. 3, 2025), <https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf>.

between compatible products or providers. The aim is to prevent an agency from acquiring a system that it cannot adequately assess, monitor, or replace.¹¹⁶

The later technology memoranda cited here address oversight of federal technology and logging across agency and third-party systems.¹¹⁷ They may contribute to the management of AI-related incidents, but they concern executive-branch administration rather than Article III courts. Their management requirements should also be distinguished from a right to relief in an individual case.¹¹⁸

2. Procurement, Tax Administration, and Standardization

For an agency acquiring AI, the contract is part of the security arrangements. It should establish how the system will be tested, how significant updates will be reviewed, and what information the provider must supply when something goes wrong. It should also make suspension or replacement possible without losing access to essential records. M-25-22 provides a federal procurement basis for several of these concerns, while NIST materials can help specify the technical work involved.

Tax administration illustrates why these arrangements matter. A system that selects audits or helps apply complex anti-avoidance rules can influence both who receives scrutiny and how the law is applied. The cited work on tax-related AI examines proportionate risk management and the possibility that classification and

¹¹⁶ Off. of Mgmt. & Budget, Exec. Off. of the President, Memorandum M-25-22, *Driving Efficient Acquisition of Artificial Intelligence in Government* 3-12 (Apr. 3, 2025), <https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-22-Driving-Efficient-Acquisition-of-Artificial-Intelligence-in-Government.pdf>

¹¹⁷ Off. of Mgmt. & Budget, Exec. Off. of the President, Memorandum M-26-10, *Reinforcing Transparency, Accountability, and Oversight of Federal Technology* 1-8 (Mar. 31, 2026); Off. of Mgmt. & Budget, Exec. Off. of the President, Memorandum M-26-14, *Ensuring Effective and Efficient Agency Logging and Network Visibility to Defend Against Evolving Cyber Threats* 1-8 (May 22, 2026). These are generally applicable federal-technology and cybersecurity memoranda, not AI-specific judicial rules.

¹¹⁸ See OMB Memorandum M-25-21, at 2-3 (limiting the memorandum to executive management and disclaiming enforceable public rights); OMB Memorandum M-25-22, at 3-12.

implementation may leave taxpayer protections uneven.¹¹⁹ The EU and U.S. approaches differ in how they organize regulation and responsibility, but both address the reliability of systems used to exercise public authority. That overlap is more informative than treating either framework as exclusively preventive or exclusively concerned with liability after harm.¹²⁰

3. Branch-Specific Limits and the U.S. Gap

The U.S. framework is therefore distributed across institutions and sources of law. Courts have established powers to address improper filings and repair affected proceedings. Executive agencies face additional prospective requirements through OMB policy and procurement. What the sources discussed here do not provide is a single AI-specific regime governing security across judicial institutions. A court may be able to sanction an attacker even though the safeguards needed to detect and investigate the attack depend on that court's own arrangements.

4.3 Comparative Analysis

The comparison concerns how responsibility is organized, rather than whether either legal order regulates AI at all.¹²¹ The AI Act assigns duties to providers and deployers of qualifying systems. U.S. judicial law addresses much participant conduct through rules governing lawyers, litigants, and proceedings, while responsibility for court

¹¹⁹ Amedeo Rizzo & Giorgio Hassan, Proportionate Risk Management for Tax-Related AI Systems: Analysis and Implications Under EU and US Law, 18 World Tax Journal, no. 1 (2026), <https://doi.org/10.59403/560kfb>.

¹²⁰ Amedeo Rizzo & Giorgio Hassan, On the Distinction Between Tort-Based and Risk-Based AI Regulation: A Transatlantic Perspective, Stanford-Vienna Transatlantic Technology Law Forum Working Paper No. 131, at 35-48 (2025), <https://doi.org/10.2139/ssrn.5205034>. Amedeo Rizzo & Giorgio Hassan, The U.S. and E.U. Framework of AI Standardization: Building a Way Forward, Stanford-Vienna Transatlantic Technology Law Forum Working Paper No. 137, at 23-44 (2025), <https://law.stanford.edu/wp-content/uploads/2025/10/TTLF-WP-137-Rizzo-Hassan.pdf>.

¹²¹ Compare EU AI Act arts. 9-15, 16-21, 26, 72-73, with Fed. R. Civ. P. 11, federal judicial guidance, OMB Memoranda M-25-21 and M-25-22, and the NIST AI RMF.

technology is spread across judicial administration and procurement. Executive agencies operate under a further set of management requirements.

The timing of intervention also differs. The AI Act requires qualifying systems to be assessed before deployment and monitored during use. U.S. procedural rules often become relevant once questionable conduct comes to the court’s attention. That contrast is not absolute: the Act also requires corrective action, and U.S. institutions undertake preventive security work. The difference lies in how these responsibilities are connected, not in a complete division between prevention and response.

Enforcement reflects those arrangements. AI Act duties are supervised through the Regulation’s conformity and market-surveillance framework, subject to safeguards for judicial independence. U.S. participant misconduct is addressed principally by courts and professional disciplinary bodies, with criminal enforcement available only where the relevant offense is established. In both systems, an individual’s entitlement to relief depends on a specific legal basis. Neither a breach of general guidance nor every violation of the AI Act automatically supplies a remedy in the affected proceeding.¹²²

(i) Comparative Framework

Table 2. European Union and United States Frameworks

Dimension	European Union	United States
Primary regulatory target	Duties attach to providers, deployers, and other regulated actors according to the system’s classification.	Different rules govern participants, institutions, and vendors according to their roles and the forum.
Orientation	Continuing duties before deployment and during the system’s use.	Court procedure often responds to misconduct; preventive controls are governed separately.
Judicial AI classification	Annex III judicial uses, subject to Article 6; intended purpose is central.	No common judicial classification; the applicable categories depend on the institution and policy.
Risk management	Article 9 imposes risk-management duties on qualifying systems when applicable.	Judicial policy and procurement draw on NIST; there is no uniform court-wide statutory duty.

¹²² See OMB Memorandum M-25-21, at 2-3 (no public rights); EU AI Act arts. 70-74, 99-101 (supervision, enforcement, and administrative penalties); Charter arts. 47-48 and applicable national procedural law.

Dimension	European Union	United States
Provider obligations	Providers must document conformity, manage quality, monitor performance, and address defects.	Duties arise through contracts and applicable general or sectoral law, not a judicial-AI provider code.
Deployer obligations	Deployers must oversee and monitor use; further assessment and registration duties apply where specified.	Courts rely on their own governance; covered executive agencies are subject to OMB policies.
Participant sanctions	Member State procedural and professional rules principally determine participant consequences.	Rule 11 and other procedural or professional rules may apply, alongside inherent judicial authority.
Cybersecurity	Article 15 addresses robustness and resistance to AI-specific attacks.	Technical standards inform practice; binding duties depend on the institution and applicable law.
Logging and monitoring	Articles 12, 19, 26, and 72 address records and monitoring for qualifying systems.	Duties depend on records rules, judicial or agency policy, and contracts.
Incident reporting	Serious incidents are subject to statutory reporting; internal monitoring has a broader role.	Requirements vary; there is no uniform judicial AI-incident reporting rule.
Procurement and vendors	Contracts operate alongside statutory duties of providers and deployers.	M-25-22 governs covered executive procurement; court procurement follows separate arrangements.
Judicial remedies	National procedural remedies apply, subject to relevant Union-law requirements.	Relief depends on the forum's procedural powers and the applicable standard of prejudice.
Institutional model	EU-wide rules implemented through Union and Member State authorities.	Responsibilities divided by government branch, jurisdiction, and institution.

The AI Act places continuing obligations on qualifying providers and deployers before any litigant proves that an attack affected a judgment.¹²³ Risk assessment and testing are meant to address foreseeable problems before use; monitoring and oversight allow the institution to respond as problems emerge. The preventive role of these duties does not depend on a successful claim for relief in a particular case.

U.S. procedural law provides established routes for responding when participant misconduct becomes apparent.¹²⁴ A federal court can require an explanation, preserve the filing, and consider sanctions or professional referral. It can also determine whether the proceeding needs correction. Yet those powers may be difficult to use without technical detection and reliable records. The availability of a sanction does not establish what precautions the court should have taken before the filing arrived.

¹²³ See EU AI Act arts. 9-15, 17, 26, 72.

¹²⁴ See Fed. R. Civ. P. 11(c); *Chambers*, 501 U.S. at 43-51; Model Rules of Pro. Conduct rr. 3.3, 8.4(c).

These differences must be read alongside variation within each legal order. EU Member States retain their own procedural and professional rules. U.S. executive agencies operate under OMB requirements that address the system's lifecycle, while individual courts may adopt additional safeguards. The comparison concerns the allocation of legal duties; it does not establish how well a particular institution performs them.

(ii) System Responsibility and Participant Responsibility

The two forms of responsibility protect different interests. Rules for participants preserve the distinction between presenting material to an institution and commanding its systems. Institutional duties address the security of the technology through which that material is processed. A court cannot replace technical safeguards with an assumption that every person using a lawful filing channel will act properly.

The Brazilian response illustrates how the two inquiries can proceed together: a court addressed the filing, while national guidance addressed system safeguards.¹²⁵ Successful detection does not excuse an attempted manipulation, and a sanction against the submitter does not establish that the institution's security arrangements were adequate.

The distinction also matters for an affected party. A compromised summary may need to be reviewed even when stolen credentials make it difficult to identify the attacker. Conversely, deliberate manipulation may warrant a sanction even when detection prevented prejudice. The evidence needed to attribute misconduct is therefore not necessarily the same as the evidence needed to correct a proceeding.

¹²⁵ The CNJ's June 2026 response illustrates the coexistence: the case-specific sanction remained within judicial competence while national guidance created Proseg-IA and adversarial-audit measures. Conselho Nacional de Justiça, Informativo No. 9/2026, at 3-4.

(iii) Judicial Independence

Judicial independence requires a distinction between supervising adjudication and securing the systems that support it. Outside control over legal reasoning or unrestricted access to deliberative material raises concerns that ordinary access controls and security testing do not. Technical safeguards can instead help the judiciary retain control of its own infrastructure.¹²⁶

The institutional arrangements should reflect that distinction. In the EU, implementation involves court administrations, judicial councils, and competent authorities acting within their respective roles. In the United States, common technical standards can be developed through judicial bodies such as the Judicial Conference and state supreme courts. Security oversight need not give the executive branch authority over individual judicial decisions.

Under the proposed approach, judges retain responsibility for deciding cases, while judicial governance bodies establish technical and procurement requirements. Independent experts can examine security under appropriate confidentiality arrangements. Their task is to assess whether the system works within its authorized role, not to assess the merits of a judge's legal reasoning.

(iv) Counterfactual Audit of the Brazilian Incident

Table 3 distinguishes reported events in Brazil from conclusions that depend on the assumed EU or U.S. setting. A legal requirement applies only where its stated conditions are met. Other entries identify possible responses, unresolved factual questions, or measures recommended in Section 5. The table does not assume that any framework would necessarily have prevented the incident.

¹²⁶ See Saltzer & Schroeder, *Protection of Information in Computer Systems*, at 1282-88; NIST SP 800-61 Rev. 3, *Incident Response Recommendations*, at 1-44.

Table 3. Counterfactual Audit of the Brazilian Incident

Stage	Brazil	EU counterfactual	United States counterfactual
Before deployment	Resolution 615 required judicial AI governance and safeguards. Galileu's precise testing arrangements are not public.	For a qualifying system, applicable Articles 9-15 would require risk management and safeguards. Classification remains fact-dependent.	Federal judicial guidance would be relevant. Covered executive agencies, unlike courts, would also face OMB requirements.
Filing ingestion	The pleading used a lawful filing channel but contained a concealed instruction.	Treating party documents as untrusted input is an application of Articles 9 and 15 to this risk.	Comparison of visible and processed content is recommended; no general filing rule requires that court-side check.
Detection	Galileu detected the instruction and alerted the judge, according to the reported record.	Detection would support oversight and cybersecurity obligations, but would not establish full compliance.	Detection would enable investigation and any justified sanction; no uniform statute specifies this capability.
Human escalation	The judge received an alert and responded. The interface used is not publicly documented.	Applicable high-risk duties would require effective oversight and a means of intervening in the system.	The response would depend on judicial responsibilities and the court's operating procedures.
Inquiry	The judgment described the hidden text. It does not establish a clean rerun or comparison of outputs.	Investigation would draw on documentation and provider cooperation. Reporting would depend on the statutory threshold.	The court could preserve evidence and order a limited inquiry, subject to procedural and confidentiality safeguards.
Participant sanctions	The judgment imposed a fine and made institutional referrals.	National procedural and professional rules would determine attribution and sanctions.	Rule 11, inherent authority, or professional rules might apply, depending on the forum and evidence.
Procedural correction	The public record does not establish that the procedure or judgment was altered.	National remedial rules would apply, subject to relevant EU fair-trial requirements.	Relief would depend on legally relevant influence and prejudice. A trusted rerun may help establish the effect.
Incident reporting	Resolution 615 addresses adverse events with negative impact; a blocked attempt may fall below its threshold.	A blocked attempt may inform internal risk management without requiring serious-incident reporting under Article 73.	Reporting duties would depend on the institution; there is no uniform judicial AI threshold.
Systemic remediation	The CNJ subsequently adopted prompt-injection guidance and Proseg-IA.	A shared defect could require broader monitoring and corrective action under the applicable duties.	Broader review is recommended. Agencies and courts would use their respective governance arrangements.

The comparison supports a limited conclusion. Brazil addressed both the participant's conduct and the security of judicial systems. An equivalent EU deployment would need to be assessed against the Act's classification and application rules; the principal high-risk duties were not applicable on the incident date. U.S. courts would have established procedural responses, while their technical safeguards would depend on the

relevant institution. The available facts do not seem to show that any of these arrangements would certainly have prevented a comparable attack.

5. A Two-Sided Framework for Adversarial Filings

5.1 Defining and Attributing Participant Misconduct

The proposed framework begins with a narrow rule for people who submit material to courts and agencies. It would prohibit knowing attempts to use that material to control the institution's AI without authority. It would not prohibit ordinary advocacy or legitimate technical features, including accessibility measures and approved security testing.

The proposed rule may have the following structure: *“A person must not knowingly include, cause to be included, or knowingly transmit concealed, deceptive, or machine-directed content in a judicial or administrative submission for the purpose of causing an institution-authorized computational system to disregard, alter, evade, or exceed the instructions, permissions, or functions assigned by the responsible institution. This prohibition does not apply to content authorized by law, rule, order, technical specification, or an approved security-testing procedure.”*

The rule's knowledge and purpose requirements distinguish deliberate interference from an innocent technical feature.¹²⁷ “Machine-directed content” includes text or other machine-readable material intended to influence the system's operation. That description alone does not make the content unlawful. Structured filing fields and

¹²⁷ They track the caution reflected in Rule 11's reasonableness structure, inherent-power bad-faith requirements, and Model Rule distinctions among competence, supervision, candor, and dishonesty. See Fed. R. Civ. P. 11; *Chambers*, 501 U.S. at 43-51; Model Rules of Pro. Conduct rr. 1.1, 3.3, 5.1, 5.3, 8.4(c).

machine-readable citations, for example, may serve an approved purpose. The prohibition concerns their use to obtain control that the submitter was not entitled to exercise.

“Concealed content” refers to material deliberately hidden from ordinary human review or placed in a part of the file that appears differently to the system. Invisibility is relevant to purpose, but is not conclusive evidence of wrongdoing. Accessibility features and ordinary document structures may legitimately contain information that a reader does not see. The rule should expressly protect those uses rather than treat every difference between the displayed and processed file as suspicious.

“Deceptive content” concerns a false impression about the source or status of material. A party-controlled document might impersonate a court instruction, or an AI-generated conclusion might be presented as independent expert analysis. What matters is the attempt to make the system or official rely on the material as something it is not.

“Unauthorized computational control” marks the boundary between advocacy and interference. A lawyer may emphasize facts likely to appear in a summary or address publicly disclosed criteria. A technical exhibit may also quote an instruction as evidence. The proposed rule applies when the purpose is different: to make the system follow the submitter’s directions instead of processing the submission under the institution’s instructions.

The rule would apply to anyone who knowingly uses a judicial or administrative submission channel for that purpose, whether acting as a lawyer, party, or other contributor. Its reach would not depend on professional status. An outside attacker with no connection to the submission process presents a different case, to be addressed under otherwise applicable computer-access and criminal law.

Attribution therefore requires more than identifying the account from which a file was sent. The person who created the hidden instruction may not be the person who signed or transmitted the filing. The inquiry must also consider whether a client or service provider supplied the material, whether counsel authorized it, and whether an account or software process was compromised. Those possibilities support different findings of responsibility. The original file and its editing history can help establish when suspicious content appeared. Filing records may identify its transmission, while account records and communications may help distinguish authorship from authorization or supervision.¹²⁸ The institution should preserve that evidence before converting or cleaning the document. Requiring a replacement must not result in overwriting the only copy capable of showing what occurred.

Consequences should reflect those findings. Deliberately inserting or transmitting an instruction to gain unauthorized control may justify a serious procedural sanction or professional referral. Costs or evidentiary measures may also be appropriate under the governing rules. Civil or criminal liability would still require proof of its own elements, not merely a finding that the content was adversarial.

Conscious disregard of a known risk presents a different basis for responsibility. For example, a lawyer who ignores a clear warning about hidden instructions in a filing may face consequences even without having created the instruction. The inquiry should identify the warning, the risk it conveyed, and the response that was reasonably expected.

Inadequate supervision is different again. A lawyer may fail to review work supplied by a client, vendor, or AI tool without knowing that it contains hostile content. Any

¹²⁸ See Fed. R. Evid. 901(b)(9), 902(13)-(14); Fed. R. Civ. P. 26(b)(1), 34(a); Luc Moreau et al., *The Open Provenance Model Core Specification (v1.1)*, 27 Future Generation Computer Systems 743, 743-56 (2011), <https://doi.org/10.1016/j.future.2010.07.005>.

consequence should address the supervisory failure shown by the evidence, rather than describe it as deliberate deception.

Conduct after discovery can also matter. Continuing to rely on a known compromised filing, hiding the problem, or failing to preserve the evidence may create additional grounds for responsibility. The relevant duties arise from what the person knows and does after receiving notice, not simply from the earlier transmission of the file.

Where an anomaly results from an innocent document feature, malware, or stolen credentials, the immediate priority is correction and security. Punishment would require an additional basis, such as unreasonable conduct after a clear warning. The existence of hidden content should not itself establish the submitter's culpability.

That distinction is particularly important for self-represented litigants. A court can require a clean copy and protect its systems without assuming that a person understood the file's internal structure. Deliberate repetition after a clear explanation may justify a different response, but technical unfamiliarity alone should not be treated as intent to interfere.

These participant consequences remain separate from the response needed to protect the institution or correct the case. A court may preserve and isolate a filing while investigating who is responsible. A deliberate attempt may warrant a sanction even when it was blocked. Conversely, an affected procedure may need to be repeated before the attacker has been identified. Neither question should be made to depend entirely on the other.

5.2 A Proportionate Institutional Duty of Adversarial Readiness

The corresponding institutional duty is one of reasonable preparation for adversarial inputs. Courts and agencies would be expected to take precautions suited to what their

AI systems do, the access they have, and the harm an attack could cause. The proposed standard is not a guarantee of prevention. It asks whether the institution's arrangements were reasonable in light of the system's use and the technical knowledge available at the time.

Reasonable adversarial readiness connects prevention with the ability to investigate and recover. It draws on established security principles and AI risk-management approaches, but applies them to the particular work of courts and agencies.¹²⁹

The first task is to preserve the distinction between institutional instructions and submitted material. The system should not treat a document as authoritative merely because it arrived through an approved channel. Where its machine-readable content differs from what a person sees, that difference should be available for review. Shared legal or policy sources also need controls over authenticity and changes, and information from one matter should not become available in another without authorization.

The next task is to limit what a compromised system can do. Tools should have only the access required for their assigned work, with separate authorization for consequential actions. The institution should also be able to stop using a suspect output and continue through a trusted alternative. That response depends on useful warnings, preserved evidence, and access to provider assistance, all arranged before an incident rather than improvised afterwards.

The proposal would make those expectations part of judicially controlled technical and procurement standards.¹³⁰ It would not leave the content of reasonable preparation to

¹²⁹ The elements synthesize NIST AI risk management and adversarial-ML guidance, incident-response practice, classic least-privilege principles, EU AI Act high-risk duties, and the CNJ's 2026 prompt-injection guidance. See NIST AI 100-1; NIST AI 100-2e2025; NIST SP 800-61 Rev. 3; EU AI Act arts. 9-15; Conselho Nacional de Justiça, Informativo No. 9/2026, at 3-4.

be worked out only through negligence claims after a failure. Reasonableness, however, must be assessed without hindsight.¹³¹ A new attack may defeat safeguards that were carefully designed and tested. Conversely, a successful block may reveal little about the wider system if detection was accidental or the institution cannot investigate what happened. The question is whether the overall arrangements were reasonable, not whether one attack happened to succeed or fail.

The degree of protection would depend mostly on function and should not count on the product name. Calling a system an “assistant” or “copilot” does not explain its influence or authority.

The proposed framework distinguishes four categories to make that assessment more concrete.

Lower-risk support covers tasks such as formatting, routing, or independently checked transcription. These tools still need ordinary security and data protection, including control over updates and a way to correct errors. The category no longer fits when the tool changes meaning, excludes relevant material, or supplies the only version that the official reviews.

Material informational influence arises when a system shapes what the official sees. A summary may omit an important allegation; a retrieval or ranking tool may determine which sources receive attention. Such systems need testing with adversarial material and records that allow the output to be checked against the source. The reviewer should be warned about meaningful discrepancies and have a way to repeat or independently verify the task.

¹³⁰ See Saltzer & Schroeder, *Protection of Information in Computer Systems*, at 1282-88; Nat’l Inst. of Standards & Tech., *AI RMF 1.0*, at 20-36; Nat’l Inst. of Standards & Tech., *Incident Response Recommendations*, at 1-44; Conselho Nacional de Justiça, Informativo No. 9/2026, at 3-4.

¹³¹ Cf. Restatement (Third) of Torts: Liability for Physical and Emotional Harm § 3 (Am. L. Inst. 2010) (reasonable care under circumstances); NIST AI RMF 1.0, at 20-36 (iterative, contextual risk management). This Article does not propose that the tort Restatement directly governs judicial AI.

Substantive decisional support concerns systems that help assess facts or law, recommend an outcome, or draft reasons. Their contribution requires closer testing and review because it bears directly on the decision. The responsible official should be able to inspect the underlying sources, understand the system's contribution, and reject or revise it. Where that contribution is compromised, the analysis should be repeatable without the affected component.

Operational authority creates a further risk because the system can act, not merely advise. A tool that changes a record, sends an official communication, or initiates a payment needs tightly limited permissions and confirmation before consequential action. Particularly sensitive acts may justify a second approval. Monitoring and action logs should show what was actually done, with a means of reversing the action where reversal is possible.

These categories constitute only one part of the proposed readiness framework and should not be interpreted as a substitute for statutory classification: a nominally minor tool may become important if its summary is the only account an official reads.

5.3 Implementing Adversarial Readiness Across the System Lifecycle

Testing should reflect the system as the institution will actually use it. A court assistant, for example, should be assessed with the kinds of pleadings, scanned exhibits, retrieved authorities, and connected tools it will encounter in practice. Testing a general model in isolation may overlook vulnerabilities introduced by the surrounding workflow. At the same time, security testing must preserve the system's legitimate function: refusing to process all external material is not a meaningful defense for a system whose purpose is to read and analyze submissions.

The taxonomy developed in Section 3 provides a starting point for that evaluation. Testing should address direct instructions as well as instructions concealed in documents or retrieved sources, and should examine whether the system can be misled about the authority of a source or induced to access information or tools outside the assigned task. Because attackers adapt to known safeguards, evaluation cannot be limited to a fixed set of familiar prompts or previously observed techniques.¹³²

Effective testing also requires more than technical expertise. Court personnel can identify which documents, representations, and procedural steps matter to adjudication, while security specialists can examine how those elements might be manipulated. Records, accessibility, and human-factors expertise are also important because defensive measures can themselves create risks – for example, by destroying evidence, producing ineffective warnings, or misclassifying legitimate machine-readable features as attacks. No single discipline can adequately evaluate the complete workflow.

Testing should begin before deployment and recur after significant changes, incidents, or other developments that alter the system’s risk profile.¹³³ The resulting record should identify material weaknesses, the measures adopted in response, and any residual limitations. Public reporting can describe the categories of attacks examined and the corrective measures taken without disclosing operational details that would make the safeguards easier to circumvent.

¹³² See Liu et al., *Formalizing and Benchmarking Prompt Injection Attacks and Defenses*; Zhan et al., *InjecAgent*, at 10471-506; DeBenedetti et al., *AgentDojo*; Vassilev et al., *Adversarial Machine Learning*, at 15-25.

¹³³ See EU AI Act art. 9(6)-(8); NIST AI RMF 1.0, at 28-36; OMB Memorandum M-25-21, at 14-24.

Testing, however, is useful only if it is connected to a response process capable of distinguishing a blocked attempt from a compromise that reaches an output, a procedure, or a decision. The proposed framework therefore uses seven response levels.¹³⁴ They are not seven stages through which every attack must pass. Rather, they identify different factual conditions that call for different forms of investigation and correction. The seventh level addresses the separate question of systemic scale and may arise independently of the sequence described in Figure 1.

At Level 1, suspicious content is detected before it reaches the relevant AI process. The immediate tasks are to preserve and isolate the original file, obtain a clean replacement where necessary, and investigate how the suspicious material entered the submission. The incident may also justify changes to detection rules or filing procedures. If the institution can establish that the relevant AI system never processed the content and no procedural step was affected, there is ordinarily no reason to repeat the proceeding merely because an attack was attempted.

At Level 2, the suspicious content entered document preparation or another preliminary stage but was blocked before the system produced a substantive output. The central question is whether containment was complete. Processing records, component logs, and a trusted rerun may help establish where the content was stopped and whether any downstream component received it. Even where technical containment is confirmed, the conduct of the participant who introduced the material remains a separate question.

At Level 3, the attack altered an output or affected the sources selected by the system. The compromised result should no longer be relied upon until its contribution has been

¹³⁴ The grading adapts the detect-respond-recover logic of NIST SP 800-61 Rev. 3 and separates attack stages consistent with NIST AI 100-2e2025.

identified and checked. A comparison with a trusted rerun or independently produced result may show what changed and whether the affected step can simply be repeated. The responsible official should be informed, and the institution should consider whether the same vulnerability may have affected other matters.

Level 4 is reached when the compromised output influences the official's interaction with the system. Investigation must then move beyond what the model generated and determine what the official actually saw, relied upon, disregarded, or approved. Independent review may be appropriate when the original reviewer's judgment was itself shaped by the compromised material. The institution should also correct any interface, warning, or presentation feature that contributed to the problem and, where legally relevant, give affected parties an opportunity to address the altered material.

At Level 5, the compromise has affected a legally relevant procedural step, such as notice, evidence extraction, triage, or preparation for a hearing. The response should focus on repairing that step through a trusted process and providing notice or an opportunity to respond where required by law. The need to repeat a procedure does not itself establish participant culpability. Costs, sanctions, or disciplinary consequences must rest on their own legal basis.

Level 6 concerns a plausible effect on the decision itself. At that point, the applicable rules governing reconsideration, appeal, prejudice, harmless error, structural error, or vacatur determine the appropriate remedy. An attempted attack alone should not invalidate a decision. At the same time, an institution that failed to preserve basic operational records should not be able to convert that failure into an evidentiary burden that an affected person has no realistic means of satisfying.

Level 7 addresses scale rather than a further step in the causal sequence. It applies when a shared model, source, component, or vulnerability may have affected multiple

matters. Such a condition may exist before any decisional effect has been established in a particular case. The institution may therefore need to constrain or suspend the affected component, identify the population of potentially exposed matters, preserve system-wide evidence, and obtain provider assistance. Any broader review or public account should be designed to inform affected persons and improve accountability without unnecessarily disclosing confidential information or exploitable security details.

Across all seven levels, three questions remain distinct. The institution must determine how to contain and correct the technical problem, whether any participant engaged in sanctionable conduct, and whether a particular proceeding requires a legal remedy. A deliberately crafted submission may justify sanctions even when the attack was blocked. A compromised proceeding may require correction before the attacker has been identified. A defect in a shared system may require broader review even where no filer acted improperly.

Incident reporting should preserve the same distinctions. A suspicious document, a confirmed attempt, a technical compromise, and a system-wide failure do not present the same reporting question.

Every credible adversarial event should nonetheless be documented internally, including events that fall below the threshold for mandatory external reporting.¹³⁵ The record should make it possible to identify the file and system involved, follow the relevant processing, and understand the response. Where an output or action changed, that effect should be recorded as well. Access should be limited where the record contains confidential material or details that could assist another attack.

¹³⁵ Compare EU AI Act arts. 20, 72 (corrective action and monitoring), with arts. 3(49), 73 (serious-incident threshold); Resolução No. 615, art. 42 (Brazilian adverse-event reporting).

The institution should decide in advance which events require referral to a designated security body and when provider assistance is needed. A warning about an unusual file calls for a different response from evidence that a shared system is compromised. Agencies can use applicable regulatory or inspector-general arrangements; courts need reporting routes that preserve judicial control over adjudicative matters.

Aggregate reporting can explain how often incidents occur, which functions are affected, and what has been done in response. It need not reveal personal information, sealed records, or instructions for defeating the safeguards. The distinction between internal monitoring and mandatory serious-incident reporting is useful here: an attempt that does not trigger external reporting can still provide information needed to improve the system.

These arrangements depend in part on what the institution can require from its supplier. A promise of responsible AI use is of little practical value if the contract leaves the institution unable to investigate a failure or stop using the service.

Procurement should therefore address the institution's control before deployment, during an incident, and when the service changes or ends.

Before accepting the system, the institution should require testing against its expected uses and exposure to attack. The provider should disclose material limitations and identify the model and components supplied. Significant updates should be notified and reviewed before they alter the approved configuration.

The contract should also establish what happens during an incident. It should secure timely notice, access to the records needed for investigation, and cooperation from the provider and relevant subcontractors. Audit rights and responsibility for response costs should be settled before a dispute makes cooperation more difficult.

The institution must retain practical control over its data and continuity of service. Contract terms should restrict unauthorized secondary use and allow records to be transferred to another provider. Suspension, return to a trusted version, or use of an alternative process should be possible without losing essential information or disrupting proceedings unnecessarily.

Where the service depends on other model providers or software suppliers, the contract should identify who is responsible for each contribution. The institution needs a clear route to correction when a defect originates in a subcontracted component, rather than being left between suppliers who each deny responsibility.

These terms do not require public disclosure of source code or trade secrets in every case. They secure the information and assistance needed to govern the service. OMB's acquisition guidance addresses several of these concerns for covered U.S. agencies, while the AI Act assigns continuing duties to regulated providers and deployers.¹³⁶ The proposed contractual terms apply those broader approaches to adversarial filings.

Oversight should then examine whether the promised safeguards work in practice, not merely whether the contract contains the right clauses.

Judicial institutions should give a designated officer or team responsibility for coordinating AI security, including testing and incident response.¹³⁷ That team would connect procurement, technical staff, and judicial users so that warnings lead to action. Larger deployments should also undergo independent technical review, with judicial supervision where needed to protect confidential material and deliberation.

¹³⁶ OMB Memorandum M-25-22, at 3-12; EU AI Act arts. 16-21, 26, 72-73.

¹³⁷ See Raji et al., *Closing the AI Accountability Gap*, at 33-44; Admin. Off. of the U.S. Cts., *2025 Annual Report of the Director: Court Operations—Artificial Intelligence* (2026); Resolução No. 615, arts. 15, 39-42.

Agencies can coordinate this work through their existing AI, privacy, civil-liberties, and cybersecurity functions. Courts and agencies can also learn from incidents elsewhere through confidential reporting arrangements. Protected channels for authorized testing and vulnerability reports would allow staff, suppliers, and researchers to raise concerns before a weakness affects a proceeding.

Meaningful oversight requires access to evidence about the deployed system rather than reliance on general assurances. Reviewers need to understand what was tested, which versions and permissions were in use, what material changes occurred, and how prior incidents were handled. Their role, however, remains limited. Technical oversight should determine whether the system operates within the institution's assigned boundaries and whether its safeguards are functioning; it should not decide the merits of cases or substitute technical judgment for judicial decision-making.

Taken together, these measures illustrate how the EU and U.S. approaches can inform a common framework without being treated as interchangeable.

For U.S. judicial systems, the AI Act offers useful concepts concerning continuing risk management, deployer monitoring, and provider cooperation. Those ideas could be implemented through judicial policy, court rules, procurement, or legislation without importing the full structure of EU market regulation. The central question is practical: who is responsible for preparing the institution for an attack, and who must provide the information and assistance needed when one occurs?

The corresponding issue in the European Union lies more heavily on the participant side. Provider and deployer duties do not themselves determine whether a lawyer, litigant, or other submitter was responsible for a concealed machine-directed instruction, what degree of culpability was involved, or what consequence should follow for the filing or proceeding. Those questions remain matters for national

procedural and professional rules, which should distinguish deliberate interference from ordinary advocacy, legitimate machine-readable features, and innocent document anomalies.

The common objective is not to make the two legal systems identical, but to connect participant responsibility with institutional preparedness. AI systems used in adjudication or public administration should be evaluated according to what they actually do, tested under realistic conditions, and supported by procedures capable of responding to the effects that an incident actually produces. Judicial institutions should retain control over adjudication while subjecting the technical systems through which that work is performed to competent security and governance. Participants and institutions would then answer for different forms of conduct, with neither responsibility displacing the other.

5.4 Limiting Principles for Adversarial-Input Regulation

The first limit is the need to protect lawful advocacy. A rule defined simply as a prohibition on “manipulation” could reach ordinary efforts to persuade. Parties necessarily select facts and arguments and respond to known decision-making criteria. The security problem is not that a submission seeks a favorable outcome, but that it attempts to control the institution’s AI outside the submitter’s authority.

The proposed rule therefore requires knowledge and a purpose of obtaining unauthorized computational control. It also protects legitimate machine-readable material, including required document structure, accessibility features, and approved testing.¹³⁸ Where deliberate interference is not established, correction and explanation

¹³⁸ Cf. Fed. R. Civ. P. 5(d)(3), 11; 36 C.F.R. pt. 1194 (federal information and communication technology accessibility standards).

should ordinarily take priority over punishment. That does not remove any separate responsibility for inadequate supervision under the governing rules.

Attribution is a related limit. Hidden content may have been introduced before counsel received the document, or through a compromised service or account. A signature identifies a procedural connection to the filing, not necessarily its technical author. The court can require preservation and a clean copy while investigating those possibilities, without prematurely finding intentional misconduct.

The inquiry should separate who created the content from who authorized, transmitted, or supervised it. Benefit from a suspicious filing is not a substitute for that evidence. A finding of deliberate interference needs support appropriate to the finding, while negligent supervision presents a different question. Neither uncertainty should prevent correction of a proceeding where the effect on that proceeding is otherwise established. Technical uncertainty places another limit on the duty. The research discussed in this Article does not establish that any single filter or instruction can prevent every prompt-injection attack. Success against a particular test does not prove protection against different wording, document formats, or connected tools.¹³⁹ The proposed duty should therefore be assessed by the reasonableness of the safeguards and response, not by a promise that no attack will ever succeed.

An institution would be expected to reduce foreseeable exposure and limit what a compromised system can do. It should also be able to detect problems, stop reliance on affected outputs, and continue through a trusted process. A new attack may succeed despite reasonable precautions. Conversely, ignoring a known weakness may show inadequate preparation even before it causes a demonstrable change in a decision.

¹³⁹ See Vassilev et al., *Adversarial Machine Learning*, at 15-25; Liu et al., *Formalizing and Benchmarking Prompt Injection Attacks and Defenses*; DeBenedetti et al., *AgentDojo*.

Safeguards can also cause harm by flagging legitimate material. A technical exhibit may quote a malicious instruction as evidence, while an accessibility feature may contain text not visible in ordinary viewing. Recognition or rendering errors can create similar discrepancies. Automatically removing such material risks changing the record rather than merely protecting the system.

Detection should therefore preserve the original file and make the suspected content available for competent review. The reviewer needs to consider its context and purpose, as well as what the system did with it. Legitimate testing should follow an approved process. A warning should start an inquiry, not serve as conclusive evidence that the submitter attempted an attack.

Warnings themselves need careful design. Repeated alarms that rarely identify a meaningful problem can encourage users to disregard later alerts. The human-factors literature discussed here shows why reliance depends on how a system communicates risk and how useful its warnings prove.¹⁴⁰ Formal human involvement is not enough when the information presented discourages careful review.

An alert should identify the suspicious passage, explain why it matters, and state whether it has already affected processing. The response should be suited to the function: sending an official communication may warrant confirmation that a low-risk draft does not. Institutions should review false alarms and override patterns so that warnings remain useful without concealing the system's limitations.

Self-represented litigants may face particular difficulties. A person using assistive technology, a mobile application, or a conversion service may have no practical way to

¹⁴⁰ See Lee & See, *Trust in Automation*, at 50-80; Zana Buçinca, Maja Malaya & Krzysztof Z. Gajos, *To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making*, 5 Proc. ACM Hum.-Comput. Interaction, art. 188 (2021), <https://doi.org/10.1145/3449287>.

inspect hidden document content. Technical irregularities should not lead automatically to sanctions. The proposed rule's knowledge and purpose requirements are intended to prevent liability from turning solely on the person's choice of software or lack of technical expertise.

Accessible submission tools, clear warnings, and an opportunity to provide a clean replacement can protect both the litigant and the institution.¹⁴¹ Repeating deliberate interference after an explanation is different from unknowingly submitting an unusual file. A secure filing process should make that distinction possible rather than assume that every participant can diagnose a technical warning.

Transparency also has limits. A detailed account of a defense may help an attacker bypass it, and investigation records may include confidential case material or trade secrets. The answer is not to withhold everything or disclose everything. Access should depend on the information needed to investigate the incident, challenge its effects, and oversee the institution's response.

Existing protective procedures provide a basis for that approach.¹⁴² Parties should receive enough information to understand and contest a material effect on their case. More sensitive technical records may be examined by the court or a neutral expert under restricted conditions. Public reporting can then explain the incident and corrective measures without exposing protected information or a method for repeating the attack.¹⁴³

¹⁴¹ See 28 U.S.C. § 1915; Fed. R. Civ. P. 1 (just, speedy, and inexpensive determination); Amershi et al., *Guidelines for Human-AI Interaction*.

¹⁴² See Fed. R. Civ. P. 26(c); Fed. R. Evid. 706; *United States v. Zolin*, 491 U.S. 554, 568-72 (1989) (in camera review in an appropriate privilege context).

¹⁴³ See Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 Stan. L. Rev. 1343, 1409-28 (2018) (analyzing trade-secret barriers and protective mechanisms in criminal justice).

The investigation itself can create privacy risks. System logs may contain sensitive information about parties and others, including material unrelated to the suspected attack. Keeping every record indefinitely would increase that exposure. The institution needs to preserve relevant evidence without treating an incident as a justification for unlimited collection or access.

Records should therefore be retained for a defined purpose, under applicable schedules and any necessary incident hold, with appropriate access and security controls.¹⁴⁴ Logging should focus on events needed to understand the system's operation rather than collect unnecessary personal information or a judge's private reasoning. Privacy considerations should shape the arrangements before deployment, not be addressed only after extensive records have accumulated.¹⁴⁵

Judicial independence requires similar care. Security oversight should not give an executive authority or vendor control over judicial reasoning. Judicial councils and conferences can instead set technical requirements and supervise audits and incident review. General legal standards may establish safeguards without prescribing outcomes in individual cases. The proposed framework depends on keeping those functions distinct.

Draft reasons and exploratory prompts may nonetheless reveal deliberative material. An investigation should distinguish records of system events from attempts to reconstruct a judge's unexpressed reasoning. Where access to sensitive material is

¹⁴⁴ See, e.g., 44 U.S.C. § 3101; 36 C.F.R. §§ 1222.22, 1222.32, 1236.10-.26 (2026); Regulation (EU) 2016/679 arts. 5, 25, 32.

¹⁴⁵ See Nat'l Inst. of Standards & Tech., NIST Privacy Framework, *A Tool for Improving Privacy Through Enterprise Risk Management*, Version 1.0, at 20-39 (2020), <https://doi.org/10.6028/NIST.CSWP.01162020>.

necessary, its scope and conditions should remain subject to appropriate judicial control.¹⁴⁶

A further difficulty is that a system's role can change. A tool introduced for summarization may later retrieve external information, draft reasons, or acquire permission to act. A provider may change the model without changing the product name. An earlier assessment may therefore cease to describe the system actually being used.

Under the proposed framework, material changes would require reassessment of function and risk. The institution should consider what the system can now access and do, and how officials rely on it. Uncertainty would not automatically place every tool in the highest category, but an old low-risk label would not justify expanded use without review.

Resource constraints also matter. A small court may lack dedicated security staff or the bargaining power to obtain extensive assistance from a vendor. Requirements designed only for large institutions could delay useful services or narrow the available suppliers. Proportionate implementation should allow shared expertise and common procurement arrangements rather than demand that each court build the same capacity independently.

Judicial bodies can support smaller institutions through approved configurations, shared tests, and common contractual terms.¹⁴⁷ Lower-risk tools need not undergo the same review as systems capable of taking consequential action. Limited resources

¹⁴⁶ See *Fayerweather v. Ritch*, 195 U.S. 276, 306-07 (1904); *In re Certain Complaints Under Investigation*, 783 F.2d 1488, 1519-20 (11th Cir. 1986).

¹⁴⁷ The Brazilian Proseg-IA and Sinapses model provides one institutional example. Conselho Nacional de Justiça, Informativo No. 9/2026, at 3-4; Resolução No. 615, arts. 21-24.

justify prioritizing risks and pooling support, but do not remove the need to assess a system before relying on it for high-impact work.

Finally, improving one defense may redirect rather than end an attack. A filer unable to use hidden text may try to influence a retrieved source, misuse an account, or exploit a supplier. A framework confined to the technique observed in Parauapebas would overlook those related routes.

Institutional preparation must therefore continue as the system and its uses change. After addressing an incident, the institution should examine whether the same weakness appears in another component or process. The purpose is to preserve control over the system, not merely to block the particular instruction that first revealed the problem. These limits keep the proposal focused. Participant responsibility turns on culpable interference, not ordinary persuasion or technical irregularity alone. Institutional responsibility concerns reasonable preparation and response, not perfect prevention. Both must be applied in a way that preserves privacy, access to justice, and judicial independence.

6. Conclusion

The Parauapebas incident shows that a judicial filing may operate on two levels at once: as a legal submission addressed to the court and as a concealed instruction directed at the AI system processing it. The public record does not establish that Galileu was successfully manipulated or that the judgment was affected. Its significance lies instead in revealing a form of interference that may remain invisible to ordinary review of the document presented to the human decisionmaker.

That problem sits between procedural law and AI governance. Procedural rules can sanction the misuse of a filing, but they do not by themselves determine how courts should design, test, and monitor the systems that process adversarial material. Conversely, rules governing providers and deployers can impose duties concerning risk management, oversight, cybersecurity, and monitoring without resolving whether a lawyer or litigant acted culpably or whether a particular proceeding must be corrected. The two inquiries are connected, but they should not be collapsed.

The European Union and the United States approach that division from different institutional starting points. The AI Act develops a comparatively structured lifecycle regime for qualifying systems, while U.S. law provides courts with familiar tools for addressing misconduct by litigants and lawyers and for repairing affected proceedings. At the same time, U.S. judicial technology governance remains more decentralized, while the AI Act leaves many questions of participant misconduct and case-specific relief to national law. Neither model eliminates the underlying technical risk, and neither can substitute institutional design for factual proof of what occurred in a particular case.

The two-sided framework developed here responds to that division by treating participant conduct and institutional preparedness as distinct but complementary responsibilities. Access to a lawful filing channel does not confer authority to direct the court's computational systems. But the institution cannot rely on participant compliance as a substitute for reasonable safeguards when it chooses to expose consequential AI systems to material supplied by interested parties. Participants are responsible for attempts to exceed the role assigned to them; institutions are responsible for reasonably securing the systems through which public authority is exercised.

The legal response should then turn on what the evidence shows. An attempted attack that is detected and contained may justify sanctions and security improvements without calling the judgment into question. A compromised procedural step may require correction even when the attacker cannot yet be identified. A defect in a shared model, source, or component may justify review across multiple matters. Keeping those possibilities separate avoids two opposite errors: treating every suspicious submission as proof that adjudication was corrupted, and treating every blocked attack as proof that the system was adequately secured.