

FROM ARTIFICIAL INTELLIGENCE TO INSTITUTIONAL INTELLIGENCE: THE REORGANIZATION OF JUDICIAL COGNITIVE SPACE AND THE GOVERNANCE OF AGENTIC SYSTEMS IN THE JUDICIARY

João Sergio dos Santos Soares Pereira *

Abstract: Generative artificial intelligence and agentic systems can redistribute reading, research, synthesis, criticism, drafting, and coordination within courts. This article asks how that redistribution changes judicial cognition and which institutional safeguards are needed to preserve legitimate adjudication. Through an interdisciplinary narrative review and doctrinal analysis, it develops a framework of institutional intelligence anchored in Brazilian National Council of Justice Resolution No. 615/2025, as amended by Resolution No. 674/2026. Five categories organize the argument: institutional cognitive space describes the allocation of cognitive work; artificial adversariality identifies structured internal challenge among agents; the institutional rein integrates controls over delegated autonomy; meaningful human initiative emphasizes the continuing human authority to frame problems and determine purposes; and institutional AI curation describes the competence required to select, validate, contextualize, and reject artificial contributions. Recent research on human–AI collaboration and multi-agent failure qualifies the framework: neither human review nor artificial disagreement guarantees reliability, and short-term performance gains do not establish long-term professional competence. The article distinguishes technical auditability from legal reason-giving and internal agent debate from the parties’ right to be heard. It proposes workflow-level evaluation and an illustrative protocol connecting permissions, source verification, escalation, and responsibility. Its contribution is a judicial governance synthesis, not an empirically validated architecture. Legitimate delegation depends on preserving institutional competence and contestability throughout the production of decisions, rather than relying exclusively on human approval of a final text.

Keywords: judicial artificial intelligence; agentic AI; distributed cognition; human oversight; institutional governance.

INTRODUCTION

The incorporation of artificial intelligence into the judiciary is often described through a narrative centered on automation. Computational systems are expected to replace manual tasks, accelerate legal research,

* Ph.D. in Law, State, and Constitution, University of Brasília (UnB); LL.M. in Constitutional Law, Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa (IDP). Professor at FGV Rio and judicial and public prosecution training institutions; researcher in artificial intelligence and the justice system. Email: josh.sergio@uol.com.br / ORCID iD: <https://orcid.org/0000-0001-6271-4242>.

organize case files, assist in document drafting, and reduce the time required to perform repetitive activities. Although this description captures an important part of the ongoing technological transformation, it is becoming increasingly insufficient to explain the institutional effects of generative artificial intelligence, large language model-based agents, and multi-agent architectures.¹

The current transformation goes beyond the replacement of tools or the acceleration of previously existing procedures. Artificial systems are beginning to participate in cognitive activities that have historically structured judicial work: reading case files, identifying legally relevant facts, synthesizing documents, conducting statutory and case-law research, formulating hypotheses, comparing arguments, drafting opinions, reviewing texts, and organizing decision-making workflows. With the emergence of agentic systems, this participation is no longer limited to the production of responses. It increasingly encompasses planning, tool use, memory retention, coordination of intermediate steps, and the execution of goal-oriented tasks. In multi-agent architectures, different specialized agents may further divide functions, challenge intermediate outputs, review conclusions, and collectively produce legal artifacts.²

This evolution changes the nature of the legal problem. The relevant question is no longer merely whether a system produces accurate answers, hallucinates information, or reproduces biases embedded in its training data. These concerns remain indispensable, but they now form part of a broader institutional problem: the redistribution of cognitive capacities among judges, judicial teams, and artificial systems.

This article refers to this phenomenon as the reorganization of institutional cognitive space. The expression denotes the process through which cognitive activities previously concentrated in human actors become shared with, mediated by, or delegated to artificial systems, thereby shifting human activity toward the definition of purposes, coordination, curation, supervision, interruption, accountability, and legitimation. The issue is therefore not simply the automation of isolated tasks. It concerns a

¹ Richard Susskind, *Online Courts and the Future of Justice* (Oxford University Press 2019); David Uriel Socol de la Osa & Nydia Remolina, *Artificial Intelligence at the Bench: Legal and Ethical Challenges of Informing—or Misinforming—Judicial Decision-Making Through Generative AI*, 6 *Data & Policy* e59 (2024), <https://doi.org/10.1017/dap.2024.53>.

² Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei & Jirong Wen, A Survey on Large Language Model Based Autonomous Agents, 18 *Frontiers Comput. Sci.* 186345 (2024), <https://doi.org/10.1007/s11704-024-40231-1>; Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest & Xiangliang Zhang, Large Language Model Based Multi-Agents: A Survey of Progress and Challenges, in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* 8048, 8048–57 (2024), <https://doi.org/10.24963/ijcai.2024/890>.

transformation in the institutional location in which reading, interpretation, criticism, revision, and decision-making are produced.

This shift challenges the traditional narrative of professional replacement. Asking whether artificial intelligence will replace judges, judicial clerks, or court staff may obscure a transformation that is both more immediate and more consequential: the decomposition of judicial work into sets of tasks characterized by different degrees of delegability, risk, autonomy, and need for human control. Some activities may become extensively automated; others may be performed collaboratively; and certain functions may remain predominantly human not because artificial systems lack the technical capacity to perform them, but because those functions possess constitutional, political, and institutional significance.³

The distinction is crucial. The legitimacy of adjudication does not derive exclusively from the apparent quality of its outcome. It depends on the constitutional authority of the decision-maker, the procedure through which the decision is reached, the possibility of contestation, publicity, reasoning, the allocation of responsibility, and the preservation of institutional conditions for review and correction.

A system may produce linguistically sophisticated and legally plausible text without answering more fundamental questions: Who defined the objectives of the task? Which sources were selected? Which agents participated in the workflow? Which conclusions were discarded? Who had the authority to interrupt the process? And who remains accountable for its effects? Scholarship on AI-assisted adjudication has already shown that the legal significance of generative systems cannot be reduced to accuracy alone because judicial independence, procedural fairness, transparency, and public trust are implicated by the way such systems enter the reasoning process.⁴

Agentic systems deepen these difficulties because they transform language models into executors of workflows. Generative AI predominantly produces cognitive inputs. Agentic AI may instead receive an objective, decompose it into tasks, select tools, access databases, preserve contextual information, adapt its plan, and carry out successive actions under limited human supervision. Governance in this environment cannot therefore be restricted to reviewing the final textual output. It must extend to the

³ Sebastian Raisch & Sebastian Krakowski, *Artificial Intelligence and Management: The Automation–Augmentation Paradox*, 46 *Acad. Mgmt. Rev.* 192, 192–210 (2021), <https://doi.org/10.5465/amr.2018.0072>; Ethan Mollick, *Co-Intelligence: Living and Working with AI* (Portfolio 2024).

⁴ David Uriel Socol de la Osa & Nydia Remolina, *Artificial Intelligence at the Bench: Legal and Ethical Challenges of Informing—or Misinforming—Judicial Decision-Making Through Generative AI*, 6 *Data & Policy* e59 (2024), <https://doi.org/10.1017/dap.2024.53>; Richard M. Re & Alicia Solow-Niederman, *Developing Artificially Intelligent Justice*, 22 *Stan. Tech. L. Rev.* 242 (2019).

architecture of the system itself: its permissions, memory, execution cycles, communication mechanisms, escalation points, and possibilities for interruption.⁵

Artificial adversariality must not be confused with the constitutional principle of adversarial proceedings and the parties' right to be heard. It does not replace participation, defense, influence over the decision, or protection against surprise. Rather, it is an analytical category intended to describe artificial mechanisms of epistemic friction designed to reduce errors, expose weak assumptions, and improve the quality of internal deliberation. Its relevance derives from a deeply rooted legal insight: decisions tend to improve when they are exposed to disagreement, criticism, and revision. Yet the same architecture also makes it more difficult to reconstruct the decisional chain, identify the contribution made by individual agents, and allocate responsibility.

A paradox therefore emerges. The more distributed and sophisticated artificial intelligence becomes, the greater its potential institutional capacity; at the same time, the processes leading to its outputs may become more opaque, complex, and difficult to govern. Artificial collective intelligence may mitigate the weaknesses of isolated agents while simultaneously generating new forms of cascading error, goal drift, emergent behavior, dependency on shared memories, and diluted accountability.⁶

Against this background, the article formulates the following research question: How does the incorporation of agentic and multi-agent systems reorganize the cognitive space of judges and their teams, and which governance mechanisms are required to preserve responsibility, legitimacy, and meaningful human initiative?

The working proposition is that, when incorporated into substantive judicial workflows, agentic systems can redistribute cognitive activities rather than merely accelerate isolated tasks. Human work may consequently shift toward defining purposes, verifying sources, coordinating contributions, interrupting execution, and adopting legally defensible grounds for decision. Whether this redistribution preserves expertise and improves institutional

⁵ Yonadav Shavit et al., *Practices for Governing Agentic AI Systems* (2023), <https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf>; Lei Wang et al., *A Survey on Large Language Model Based Autonomous Agents*, 18 *Frontiers Comput. Sci.* 186345 (2024), <https://doi.org/10.1007/s11704-024-40231-1>.

⁶ Alan Chan et al., *Harms from Increasingly Agentic Algorithmic Systems*, in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* 651, 651–66 (2023), <https://doi.org/10.1145/3593013.3594033>; Taicheng Guo et al., *Large Language Model Based Multi-Agents: A Survey of Progress and Challenges*, in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* 8048, 8048–57 (2024), <https://doi.org/10.24963/ijcai.2024/890>.

performance is an empirical question, not a result established by the conceptual framework.

The central thesis of the article is that the most significant transformation produced by agentic artificial intelligence in the judiciary does not consist in the replacement of judges or court personnel, but in the reorganization of institutional cognitive space. Such reorganization requires moving beyond governance models based on the mere presence of a human actor at the end of an automated process and toward institutional architectures capable of preserving normative agency, critical capacity, and genuine power of intervention.

The formal presence of a human being within a workflow does not guarantee effective human control. Supervision may become little more than a ritual of validation when the responsible actor lacks sufficient time, knowledge, attention, access to system logs, capacity to challenge an output, or authority to halt the system's operation. This asymmetry becomes particularly acute in high-volume environments in which a judge or judicial team is presented with hundreds of artificial outputs that appear coherent and linguistically well structured. The fluency of the final text may conceal weaknesses in the process through which it was produced. Empirical research on overreliance and cognitive forcing functions supports the broader proposition that the effectiveness of human oversight depends on interaction design and preserved human competence, not merely on nominal human presence.⁷

To address this problem, the article develops five interconnected analytical categories. The first is the reorganization of institutional cognitive space, which describes the changing distribution of cognitive activities between human actors and artificial systems. The second is artificial adversariality, which identifies deliberative architectures in which specialized agents subject intermediate results to criticism, revision, and validation. The third is the institutional rein, defined as the integrated set of normative, technological, and organizational mechanisms designed to constrain, monitor, audit, interrupt, and redirect artificial systems endowed with operational autonomy. The fourth is meaningful human initiative, understood as the institutionally preserved capacity to define objectives, establish limits, challenge outputs, interrupt workflows, and assume responsibility for the effects produced by artificial systems. The fifth is institutional AI curation, conceived not as a new professional title but as a

⁷ Zana Buçinca, Maja Barbara 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*, no. CSCW1, art. 188, at 1–21 (2021), <https://doi.org/10.1145/3449287>; Ben Green, The Flaws of Policies Requiring Human Oversight of Government Algorithms, 45 *Comput. L. & Sec. Rev.* 105681 (2022), <https://doi.org/10.1016/j.clsr.2022.105681>.

distributed institutional competence involving the selection, organization, validation, rejection, coordination, and supervision of artificial capabilities incorporated into judicial activity.

These categories operate at distinct but complementary levels. The reorganization of institutional cognitive space describes the phenomenon. Artificial adversariality identifies one possible deliberative architecture. The institutional rein organizes mechanisms for governing autonomy. Meaningful human initiative provides the normative principle that should guide such governance. Institutional AI curation describes the transformation of human functions in hybrid institutional environments.

The contribution is integrative and jurisdiction-specific. Distributed cognition, institutional oversight, and meaningful human control are established research traditions; the article does not claim to originate them. Recent work by Follesdal and Langford likewise connects different forms of AI delegation to the legitimacy of adjudication. The additional step proposed here is to connect these concerns to the intermediate handoffs of an agentic judicial workflow: who frames the issue, which source representations travel between agents, when an objection survives orchestration, and how delegated functions can be reclaimed. The five categories are therefore analytical tools for identifying failures and responsibilities, rather than claims that every component is conceptually unprecedented.⁸

Methodologically, the article combines an interdisciplinary narrative review with doctrinal and normative analysis. The review was updated through October 21, 2026, using targeted web searches for scholarly publications and official materials concerning agentic AI, multi-agent failure, distributed cognition, human oversight, cognitive offloading, and judicial legitimacy. Publisher pages, research repositories, and official legal sources were prioritized; searches also followed references and related work identified in those sources. Recent studies from 2024–2026 supplement foundational organizational and cognitive scholarship. This is a purposive synthesis, not a systematic review: it does not claim exhaustive coverage or provide a pooled estimate of effects. Peer-reviewed studies, preprints, institutional reports, and professional commentary are distinguished according to their evidential status. Brazilian law supplies the principal normative setting; international sources serve as analytical comparators rather than directly applicable Brazilian law.⁹

⁸ Andreas Follesdal & Malcolm Langford, *Trustworthy AI in Courts: A Legitimacy Analysis*, 2 Cambridge Forum on AI: Law and Governance e20 (2026), <https://doi.org/10.1017/cfl.2026.10061>.

⁹ Conselho Nacional de Justiça [CNJ], Resolução No. 615, de 11 de março de 2025, DJe/CNJ No. 54/2025, at 2–17 (Mar. 14, 2025), as amended by Resolução No. 674, de 25 de março de 2026, DJe/CNJ No. 71/2026, at 2 (Mar. 26, 2026), <https://atos.cnj.jus.br/atos/detalhar/6001>.

The analysis separates three kinds of claim. Descriptive claims concern capabilities or findings documented in the cited literature. Interpretive claims concern the meaning and application of existing legal duties. Design proposals concern institutional arrangements that require subsequent testing. No new experiment, interview study, or court deployment is reported. Hypothetical workflows illustrate how the framework could be operationalized; they do not establish that Brazilian courts already use the proposed architectures or that those architectures improve adjudication.

The Brazilian framework is especially useful because it emerged after the rapid diffusion of generative AI within courts and was expressly designed to update the governance model applicable to judicial AI. The National Council of Justice's own mapping shows that AI is no longer hypothetical in the Brazilian judiciary: a 2025 mapping identified 157 AI projects in forty-three courts, with most initiated in 2024 and 2025 and a substantial share already in regular production.¹⁰ Resolution No. 615/2025 therefore operates in an institutional environment in which AI governance is not anticipatory regulation of a remote technology, but governance of systems already entering ordinary judicial work.

The Resolution represents a significant institutional advance because it recognizes that AI governance requires more than authorization for the use of artificial systems. It requires classification, documentation, assessment, monitoring, and the definition of responsibilities. This article nevertheless argues that the transition from assistive systems to agentic and multi-agent architectures introduces underexplored grey areas, including the classification of hybrid functionalities, emergent behavior, goal drift, distributed intermediate decisions, shared memories, agent orchestration, and the possibility of cascading errors.

The argument is not that existing regulatory frameworks are inherently or entirely inadequate. Rather, it is that regulation is not the same as governance. Rules initiate governance by assigning competences and establishing parameters. Effective governance, however, depends on continuing institutional capacities to observe systems in operation, learn from incidents, revise classifications, adjust permissions, train users, detect deviations, and adapt organizational structures.

The article proceeds in five sections. Section 1 distinguishes digitalization from the redistribution of judicial cognition. Section 2 examines cognitive offloading, expertise, and the limits of human–AI collaboration. Section 3 analyzes multi-agent systems and artificial adversariality. Section 4 develops

¹⁰ Conselho Nacional de Justiça [CNJ], *Mapeamento de IA no Poder Judiciário 2* (2025), <https://www.cnj.jus.br/wp-content/uploads/2025/12/relatorio-mapeamento-projetos-iniciativas-inteligencia-artificial.pdf> (identifying 157 AI projects across forty-three courts, including eighty-two generative-AI projects and ninety-two projects in regular production).

the governance framework in dialogue with Brazilian law. Section 5 addresses institutional AI curation and proposes an illustrative evaluation protocol.

The article thus examines the conditions under which artificial capabilities can be incorporated into judicial cognition while preserving lawful authority, professional competence, and the ability to contest and correct the resulting work.

I. FROM DIGITALIZATION TO INSTITUTIONAL INTELLIGENCE

A. Digital transformation and the continuity of judicial cognition

For much of the last three decades, technological change in courts has been understood through the language of digital transformation. Paper files became electronic records; procedural communications migrated to digital platforms; case management systems reorganized judicial workflows; search engines expanded access to legislation and precedents; and automation progressively absorbed repetitive administrative tasks. These developments profoundly altered the material infrastructure of adjudication. They changed how information was stored, transmitted, retrieved, and processed, and in many jurisdictions they also reshaped how citizens interacted with courts.¹¹

Earlier digitalization already influenced attention, access, and the organization of judicial work; it was not cognitively neutral. An electronic record could nevertheless leave much of the substantive sequence of reading, identifying issues, comparing authorities, and drafting reasons to human professionals. The distinction drawn here is therefore one of degree and function, not a clean historical rupture: generative and agentic systems can take a more active role in producing the representations and intermediate judgments on which later legal work depends.

This distinction matters because technological modernization and cognitive transformation are not synonymous. Digitizing a document changes its format. Automating a clerical routine changes the means by which a predefined operation is performed. By contrast, delegating the synthesis of evidence, the identification of precedents, the construction of argumentative alternatives, or the review of legal reasoning alters where cognition occurs within the institution.

Richard Susskind anticipated part of this distinction in his account of online courts. His central insight was not simply that courts would adopt digital tools, but that technology could force legal institutions to reconsider what courts are for, which services they provide, and how those services

¹¹ Richard Susskind, *Online Courts and the Future of Justice* (Oxford University Press 2019).

should be designed.¹² The contemporary development of generative and agentic AI extends that intuition. The relevant question is no longer only whether an existing judicial service can be delivered digitally. It is increasingly whether the intellectual operations through which that service is produced remain concentrated in human professionals or become distributed across human and artificial components.

Organizational research helps clarify the significance of this transition. Herbert Simon famously treated organizations as structures for dealing with bounded rationality: they allocate information, attention, authority, and decision-making capacity under conditions in which no individual can process everything that might be relevant.¹³ March and Olsen similarly showed that institutions shape not only formal competences but also the categories, routines, and interpretive structures through which actors understand situations and decide how to respond.¹⁴ Long before contemporary AI, organizational theory therefore suggested that cognition is never purely individual. Institutions structure what is noticed, who processes it, what counts as relevant information, and which procedures transform information into authorized decisions.

Research on distributed cognition makes the point even more explicit. Hutchins demonstrated that cognitive performance may be distributed across people, representational systems, artifacts, and organizational routines rather than located exclusively within an individual mind.¹⁵ This perspective is particularly useful for judicial institutions. Courts have always depended on distributed cognition: judges rely on clerks and judicial staff; files preserve institutional memory; procedural rules structure attention; precedents stabilize legal knowledge; databases organize retrieval; and collegial courts divide, contest, and aggregate interpretive work.

Artificial intelligence does not create distributed judicial cognition. It changes its composition.

The crucial transformation begins when computational systems cease to operate merely as repositories or transmission mechanisms and participate directly in activities through which legally relevant knowledge is constructed. A system that stores a judgment is part of judicial information infrastructure. A system that compares hundreds of judgments, identifies lines of precedent, summarizes conflicting authorities, and proposes how they bear on a pending case occupies a different institutional position. It contributes to the cognitive pathway through which the decision becomes possible.

¹² Richard Susskind, *supra* note 11.

¹³ Herbert A. Simon, *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations* (4th ed., Free Press 1997).

¹⁴ James G. March & Johan P. Olsen, *Rediscovering Institutions: The Organizational Basis of Politics* (Free Press 1989).

¹⁵ Edwin Hutchins, *Cognition in the Wild* (MIT Press 1995).

This is why the distinction between digital transformation and what this article calls cognitive-institutional transformation is necessary. The former changes the technological means through which institutional functions are performed. The latter changes the distribution of the cognitive capacities on which those functions depend.

The distinction also prevents an overly simple opposition between automation and human work. Organizational scholarship has shown that automation and augmentation are often interdependent rather than mutually exclusive. Raisch and Krakowski describe an automation–augmentation paradox in which AI may simultaneously take over some tasks while changing and intensifying the human capabilities required around them.¹⁶ In judicial environments, this implies that automating research, synthesis, classification, or drafting does not simply remove human activity. It can relocate human effort toward reviewing, coordinating, challenging, contextualizing, and taking responsibility for machine-produced outputs.

The empirical literature on generative AI in professional work reinforces the need to focus on redistribution rather than replacement. Brynjolfsson, Li, and Raymond, studying the deployment of a generative AI assistant among more than five thousand customer-support workers, found substantial average productivity gains but also significant heterogeneity across workers, including different effects on learning and performance according to prior experience.¹⁷ Although judicial work is normatively and professionally distinct from customer support, the study illustrates a broader proposition relevant here: AI does not enter an organization as a neutral productivity layer. It changes how expertise is exercised, transferred, and valued.

Performance gains also depend on the comparator. Vaccaro, Almaatouq, and Malone’s meta-analysis found that human–AI combinations did not, on average, outperform the better of humans or AI working alone, with different results for decision and content-creation tasks. This challenges any presumption that combining a judge with an AI assistant necessarily creates epistemic synergy. In courts, technical superiority over a comparator would still be insufficient to establish legal authority; nevertheless, meaningful evaluation must specify both the institutional safeguards and the performance baseline against which assistance is assessed.¹⁸

¹⁶ Sebastian Raisch & Sebastian Krakowski, *Artificial Intelligence and Management: The Automation–Augmentation Paradox*, 46 *Acad. Mgmt. Rev.* 192, 192–210 (2021), <https://doi.org/10.5465/amr.2018.0072>.

¹⁷ Erik Brynjolfsson, Danielle Li & Lindsey R. Raymond, *Generative AI at Work*, 140 *Q.J. Econ.* 889, 889–942 (2025), <https://doi.org/10.1093/qje/qjae044>.

¹⁸ Michelle Vaccaro, Abdullah Almaatouq & Thomas Malone, *When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis*, 8 *Nature Hum. Behav.* 2293–2303 (2024), <https://doi.org/10.1038/s41562-024-02024-1>.

The same point has been developed in organizational studies of learning algorithms. Faraj, Pachidi, and Sayegh argue that algorithmic systems can reshape expertise, occupational boundaries, coordination, and organizational control.¹⁹ Shrestha, Ben-Menahem, and von Krogh similarly distinguish different structures through which human and algorithmic decision-making may be combined, ranging from delegation to sequential and aggregated forms of human–AI decision-making.²⁰ These studies predate the current generation of agentic systems, but they reveal why institutional design becomes increasingly important as artificial systems move closer to the production of consequential decisions.

B. From content generation to the execution of cognitive workflows

Generative AI accelerated this transformation because it entered domains previously understood as knowledge work rather than merely information processing. Large language models can summarize extensive records, reorganize arguments, extract issues, compare texts, generate drafts, explain doctrinal distinctions, and propose alternative formulations. Their relevance to legal institutions follows precisely from the textual and interpretive character of legal practice.

Still, a generative model used as a conversational assistant ordinarily preserves a recognizable allocation of initiative. A human formulates a prompt, the system produces an output, and a human decides what to do next. Even when the model performs cognitively sophisticated operations, the interaction remains largely episodic.

Agentic systems alter that structure. The contemporary literature describes LLM-based autonomous agents as systems that combine language-model capabilities with components such as profiling or role specification, memory, planning, and action.²¹ Rather than responding only to discrete prompts, an agent may pursue an objective across multiple steps, maintain relevant state, select tools, retrieve information, revise a plan, and execute actions in an external environment. The distinction is therefore not simply one of greater computational performance. It concerns the organization of action.

Terminology remains unsettled. Li, Reinhard, and Leimeister’s 2026 account distinguishes individual agentic systems from configurations of

¹⁹ Samer Faraj, Stella Pachidi & Karla Sayegh, *Working and Organizing in the Age of the Learning Algorithm*, 28 *Info. & Org.* 62, 62–70 (2018), <https://doi.org/10.1016/j.infoandorg.2018.02.005>.

²⁰ Yash Raj Shrestha, Shiko M. Ben-Menahem & Georg von Krogh, *Organizational Decision-Making Structures in the Age of Artificial Intelligence*, 61 *Cal. Mgmt. Rev.* 66, 66–83 (2019), <https://doi.org/10.1177/0008125619862257>.

²¹ Lei Wang et al., *A Survey on Large Language Model Based Autonomous Agents*, 18 *Frontiers Comput. Sci.* 186345 (2024), <https://doi.org/10.1007/s11704-024-40231-1>.

interacting agents and emphasizes capabilities rather than a rigid sequence of technological stages. This article accordingly does not treat “generative,” “agentic,” and “multi-agent” as mutually exclusive classes. A generative model can be a component of an agent; an agent can operate alone; and a multi-agent arrangement need not possess extensive autonomy. Delegation should be assessed by permitted actions, persistence, dependencies, and the consequences of execution, not by a vendor’s label.²²

A useful way to express this transition is to distinguish content delegation from workflow delegation. When a court professional asks a generative model to summarize a document, a bounded cognitive product is delegated. When an agent is instructed to identify the relevant legal issue, retrieve authorities, compare competing lines of precedent, produce an initial synthesis, verify citations, and deliver a structured memorandum, the delegation increasingly encompasses the pathway through which the product is generated.

The shift becomes more consequential when agents operate together. Research on LLM-based multi-agent systems describes architectures in which specialized agents communicate, divide roles, coordinate actions, review one another’s work, or engage in cooperative and competitive interactions.²³ In such environments, the output no longer necessarily results from a single human–model exchange. It may emerge from an artificial division of cognitive labor.

That possibility introduces a qualitatively different institutional problem. Traditional digital systems largely allowed courts to ask whether a technological tool functioned as designed. Generative systems added questions about the reliability and provenance of generated content. Agentic systems require an additional layer of inquiry: how was the workflow itself structured, and where within that workflow were meaningful choices made?

Governance scholarship on agentic AI already recognizes this shift. Shavit and co-authors define agentic systems in terms of their capacity to pursue complex goals with limited direct supervision and emphasize the need to govern not only model outputs but the responsibilities, permissions, actions, monitoring arrangements, and accountability structures surrounding agents.²⁴ Attempts to distinguish AI agents from broader forms of agentic AI similarly emphasize the progression from task-specific agent components

²² Mahei Manhai Li, Philipp Reinhard & Jan Marco Leimeister, *Agentic AI: From Autonomous Agents to Enterprise Agentic AI Systems*, 68 *Bus. & Info. Sys. Eng.* 1191–1204 (2026), <https://doi.org/10.1007/s12599-026-01009-w>.

²³ Taicheng Guo et al., *Large Language Model Based Multi-Agents: A Survey of Progress and Challenges*, in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* 8048, 8048–57 (2024), <https://doi.org/10.24963/ijcai.2024/890>.

²⁴ Yonadav Shavit et al., *Practices for Governing Agentic AI Systems* (2023), <https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf>.

toward systems capable of orchestrating multi-step, goal-directed processes.²⁵

For judicial institutions, the implications are substantial. A hallucinated citation in a generated paragraph is a serious problem, but it is comparatively visible and conceptually familiar. A flawed decision embedded upstream in an agentic workflow may be harder to detect: an agent may formulate an incomplete search strategy, another may rely on the resulting subset of authorities, a third may synthesize them coherently, and a final agent may produce a persuasive draft. Linguistic quality at the end of the chain can obscure epistemic weakness at its beginning.

This transition is central to the argument of this article. Once AI systems participate in planning, selection, coordination, revision, and execution, judicial institutions are no longer merely adopting tools. They are redesigning the internal distribution of cognitive functions.

C. Institutional intelligence as distributed and governed cognition

The concept of institutional intelligence proposed here is intended to capture this emerging configuration. It should not be understood as suggesting that an institution possesses consciousness, intentionality, or intelligence in the same sense as an individual person. Nor is the term used as a synonym for the aggregate technical sophistication of a court.

Institutional intelligence refers instead to an organizational condition in which the cognitive work required for institutional action is distributed across human actors, artificial systems, informational infrastructures, procedural rules, and governance mechanisms, while remaining integrated into an institutionally accountable decision-making architecture.

Two aspects of this definition are particularly important.

First, institutional intelligence is relational rather than technological. A powerful model does not make an institution intelligent. What matters is how different capacities are connected. A comparatively modest system embedded in a carefully designed architecture of expertise, verification, contestation, traceability, and responsibility may contribute more to institutional intelligence than a technically superior model deployed without such structures.

Second, institutional intelligence is governed cognition. Distribution alone is insufficient. Cognitive work may be efficiently distributed and still produce an institutionally illegitimate arrangement. Courts exercise public authority under legal and constitutional constraints. The capacity to generate or optimize a result therefore cannot be separated from questions concerning

²⁵ Ranjan Sapkota, Konstantinos I. Roumeliotis & Manoj Karkee, *AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges*, 2(3) *SuperIntelligence – Robotics – Safety & Alignment* (2025), <https://doi.org/10.70777/si.v2i3.15161>.

who is authorized to define objectives, which functions may be delegated, how reasoning can be reconstructed, when human intervention is required, and who remains answerable for the final exercise of public power.

This point separates institutional intelligence from an efficiency-centered understanding of AI adoption. An institution does not become more intelligent merely because it decides more cases, generates documents more rapidly, or reduces the time required for legal research. Those outcomes may reflect greater computational capacity without greater institutional capacity. Indeed, an organization may become operationally faster while becoming epistemically more dependent, less capable of detecting errors, or less able to explain how relevant conclusions were reached.

Institutional intelligence therefore contains an irreducibly normative dimension. In the judicial context, its quality must be assessed not only through productivity but through the preservation of capabilities necessary for lawful and legitimate adjudication: critical scrutiny, reason-giving, contestability, responsibility, institutional memory, capacity for correction, and the possibility of meaningful human intervention.

This framework also changes how the familiar debate over “human versus machine” should be formulated. The relevant question is not which of the two is more intelligent in the abstract. Judicial institutions have never depended on isolated intelligence. They depend on organized combinations of expertise, procedure, records, institutional roles, review mechanisms, and authority. AI adds new cognitive components to that pre-existing architecture.

Which cognitive operations should remain predominantly human? Which may be shared? Which may be delegated under specific constraints? Which forms of artificial collaboration improve epistemic quality? At what points should institutional friction deliberately slow automated execution? What forms of expertise must humans retain to supervise systems meaningfully? And how should responsibility be structured when a legally relevant result emerges from a sequence of interactions among multiple artificial and human actors?

These questions cannot be answered at the level of technological performance alone. They require a theory of how judicial cognition is allocated, preserved, and governed. The transition from artificial intelligence to institutional intelligence therefore describes more than the adoption of increasingly capable systems. It marks a shift in the object of governance itself. The task is no longer simply to regulate a model that produces content. It is to govern a hybrid cognitive architecture in which people and artificial systems perform interdependent functions within the production of judicial knowledge.

If this diagnosis is correct, the next question follows directly. Once artificial systems begin to occupy parts of the cognitive pathway of adjudication, what happens to the cognitive space previously occupied by judges and their teams?

II. THE REORGANIZATION OF JUDICIAL COGNITIVE SPACE

A. Judicial cognition as institutional infrastructure

Legal institutions are ordinarily described through jurisdiction, competence, procedure, hierarchy, and authority. Less visible is the cognitive infrastructure that enables those categories to operate in practice. Judicial decisions are not produced by legal rules alone. They depend on sustained acts of attention: reading, remembering, distinguishing, comparing, prioritizing, interpreting, and justifying. These activities are often treated as background features of professional expertise, but they are also institutional resources. A court that lacks the capacity to identify the legally relevant facts in a record, distinguish controlling from persuasive authority, recognize exceptions, reconstruct procedural history, or test the coherence of an argument cannot convert formal authority into legitimate adjudication.

For that reason, the introduction of AI into judicial work should not be evaluated exclusively in terms of output quality. It should also be assessed in terms of what happens to the cognitive practices through which professional judgment is formed and maintained. The issue is not romantic attachment to labor-intensive methods. Many forms of legal work are unnecessarily repetitive and should be simplified. The institutional concern is more precise: when a task is delegated, the organization may also be delegating an opportunity for learning, error detection, professional calibration, or the maintenance of expertise.

This distinction becomes especially relevant in environments of chronic workload pressure. Courts have strong incentives to use AI where the technology promises faster synthesis, shorter drafting time, or improved retrieval. Those gains may be real. Yet the same organizational pressures that make automation attractive may also reduce the time available for meaningful review of automated outputs. Efficiency can therefore generate a feedback loop: the institution delegates more because workload is high, then restructures production around the new speed, and eventually lacks the temporal slack required for the human scrutiny on which the legitimacy of delegation was initially conditioned.

The problem is not that judges or staff necessarily become less capable whenever they use AI. The problem is that capability becomes path dependent. Professional competence is partly maintained through repeated practice. If certain forms of reasoning are progressively outsourced, the

ability to perform them independently may weaken, especially among less experienced professionals whose expertise has not yet been consolidated. This possibility is now supported by a growing empirical literature on human-AI collaboration and cognitive dependency.²⁶

B. The paradox of cognitive assistance

AI assistance presents a genuine paradox. In the short term, it can increase performance. In the longer term, the same assistance may alter how users allocate effort, develop expertise, and persist when the system is unavailable or wrong.

The paradox is not unique to law. Human-computer interaction research has repeatedly shown that people can over-rely on automated recommendations, particularly when outputs are fluent, confident, or accompanied by explanations. Bućinca, Malaya, and Gajos found that cognitive forcing interventions could reduce overreliance on incorrect AI advice, while ordinary explanatory interfaces did not reliably produce the same effect.²⁷ Their work is significant for legal institutions because it shifts the problem from a simplistic “transparency versus opacity” frame toward interaction design. A user may receive an explanation and still fail to engage critically with the recommendation. Conversely, an interface that requires an independent judgment before revealing AI advice can preserve a degree of cognitive resistance.

Lee and co-authors’ 2025 survey of 319 knowledge workers, covering 936 reported examples, provides complementary evidence. Greater confidence in generative AI was associated with less reported critical thinking, while work shifted toward verification, integration, and task stewardship. Because the study measures self-reported practices rather than longitudinal cognitive decline, it cannot establish that AI causes professional deskilling. Its relevance is organizational: the activities grouped here under institutional AI curation already have counterparts in observed knowledge-work practices, although their judicial form and legal consequences require separate investigation.²⁸

²⁶ Grace Liu, Brian Christian, Tsvetomira Dumbalska, Michiel A. Bakker & Rachit Dubey, *AI Assistance Reduces Persistence and Hurts Independent Performance* (arXiv preprint, version 6, Oct. 13, 2026), <https://doi.org/10.48550/arXiv.2604.04721>; Nicholas Bednar, David R. Cleveland, Allan Erbsen & Daniel Schwarcz, *Artificial Intelligence and Human Legal Reasoning*, 121 Nw. U. L. Rev. (forthcoming 2027), SSRN No. 6525800 (working paper, 2026), <https://doi.org/10.2139/ssrn.6525800>.

²⁷ Zana Bućinca, Maja Barbara 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, no. CSCW1, art. 188, at 1–21 (2021), <https://doi.org/10.1145/3449287>.

²⁸ Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks & Nicholas Wilson, *The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in*

Recent experimental evidence makes the competence question more explicit. In a 2026 preprint, updated on October 13, Liu and co-authors report randomized experiments involving 1,222 participants in mathematical-reasoning and reading-comprehension tasks. Assistance improved immediate performance but was followed by reduced persistence and poorer unassisted performance under the study conditions. These findings support concern about the design of assistance, but do not establish lasting loss of judicial expertise. Transfer to judges, longer periods, and professional environments remains to be tested; the proposed explanation involving expectations of immediate answers should also be distinguished from the experimentally observed outcomes.²⁹

Early legal-specific research complicates any simple deskilling narrative while reinforcing the importance of interaction design. Bednar, Cleveland, Erbsen, and Schwarcz found no general decline in downstream legal comprehension after short-term AI use; participants who used AI during an earlier synthesis task performed better on a later unassisted application task. Yet when participants subsequently used AI to revise legal-reasoning memoranda, those who began with weaker memoranda improved while those who began with stronger memoranda regressed.³⁰ The result is more nuanced than a claim that AI necessarily weakens legal reasoning. It instead suggests that the cognitive effects of assistance depend on when, how, and by whom the system is used. “Better assisted output” and “stronger independent professional capacity” therefore remain analytically distinct variables.

The legal study concerns upper-level law students and remains forthcoming; it should not be read as direct evidence about sitting judges or long-term professional trajectories. Taken together, these studies support a conditional governance claim: institutions should evaluate when assistance is introduced, which skills users already possess, and what they can still do without it. They do not support either a universal deskilling thesis or a universal learning benefit.

The Brazilian debate has begun to identify this issue in explicitly juridical terms. Luis Felipe Salomão and Dierle Nunes have argued that AI design should preserve calibrated forms of cognitive friction because immediate, frictionless assistance may encourage dependence and weaken the active

Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers, in Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, art. 1121, 1–22 (2025), <https://doi.org/10.1145/3706598.3713778>.

²⁹ Grace Liu, Brian Christian, Tsvetomira Dumbalska, Michiel A. Bakker & Rachit Dubey, *AI Assistance Reduces Persistence and Hurts Independent Performance* (arXiv preprint, version 6, Oct. 13, 2026), <https://doi.org/10.48550/arXiv.2604.04721>.

³⁰ Nicholas Bednar, David R. Cleveland, Allan Erbsen & Daniel Schwarcz, *Artificial Intelligence and Human Legal Reasoning*, 121 Nw. U. L. Rev. (forthcoming 2027), SSRN No. 6525800 (working paper, 2026), <https://doi.org/10.2139/ssrn.6525800>.

construction of legal reasoning.³¹ Their intervention is especially useful because it reframes friction as an institutional good under certain conditions. In digital transformation discourse, friction is normally treated as waste. In adjudication, however, some friction is epistemic: the effort of articulating an argument, testing a premise, comparing competing authorities, or writing a justification can itself improve understanding.

This does not mean that every traditional practice should be preserved. It means that the relevant design question is not simply “Can AI perform this task?” but “What institutional function is lost if humans cease to perform this task at all?” That question is central to the reorganization of judicial cognitive space.

C. Cognitive offloading, deskilling, and the supervision paradox

The movement from assistance to delegation may generate what can be called a supervision paradox. The more an institution relies on AI to perform a cognitive activity, the more it needs competent humans to verify the resulting work; yet the same reliance may reduce the frequency with which humans practice the underlying skill required for meaningful verification.

This paradox exposes a weakness in many “human-in-the-loop” formulations. Human oversight is often treated as a binary safeguard: either a person is present or not. But effective supervision is itself a demanding cognitive practice. It requires domain knowledge, time, situational awareness, confidence to disagree, access to relevant evidence, and institutional authority to reverse the system. A human who merely approves an output after the fact may function as a signature mechanism rather than a genuine source of judgment.

Ben Green’s analysis of public-sector algorithmic governance is instructive. Surveying policies that require human oversight, Green argues that such requirements can provide false reassurance when they assume, without evidence, that people will reliably detect and correct algorithmic errors.³² His proposed shift from individual human oversight toward institutional oversight is closely aligned with the argument developed here: the legitimacy of AI-assisted public decision-making cannot rest on the unsupported assumption that a human reviewer will compensate for defects generated elsewhere in the system.

The emerging European regulatory approach similarly recognizes that oversight depends on capacity, not mere presence. Article 14 of the EU AI

³¹ Luis Felipe Salomão & Dierle Nunes, O Jurista que Parou de Tentar: IA, Dependência Cognitiva e Como o Design Importa, ConJur (May 8, 2026), <https://www.conjur.com.br/2026-mai-08/o-jurista-que-parou-de-tentar-ia-dependencia-cognitiva-e-como-o-design-importa-2/>.

³² Ben Green, The Flaws of Policies Requiring Human Oversight of Government Algorithms, 45 Comput. L. & Sec. Rev. 105681 (2022), <https://doi.org/10.1016/j.clsr.2022.105681>.

Act requires high-risk AI systems to be designed so that human overseers can understand relevant capabilities and limitations, remain aware of automation bias, interpret outputs, disregard or reverse them, and intervene or stop the system where appropriate.³³ This is a more demanding conception of oversight than a symbolic human checkpoint. Yet even such requirements leave open a deeper institutional question: how can organizations preserve the expertise that makes these capacities real over time?

In judicial settings, that question is particularly acute. Consider a court team in which AI routinely performs first-pass reading, issue extraction, precedent search, argument mapping, and initial drafting. The judge may gain time for higher-level evaluation. But the team may also gradually lose opportunities to learn how weak issues appear in an unstructured record, how misleading patterns emerge in precedent search, or how seemingly minor factual details change the applicable legal frame. Experienced judges may retain those abilities because they developed them before widespread AI adoption. Junior judges and staff may not acquire them in the same way.

The resulting risk is not only individual deskilling. It is institutional deskilling: the erosion of distributed capacities required to understand, challenge, and reconstruct the system's work. An organization may remain highly productive while becoming increasingly dependent on infrastructures it cannot independently audit in substance.

This is why the reorganization of cognitive space must be evaluated longitudinally. An AI system can appear beneficial when assessed through immediate metrics—time saved, cases processed, documents generated—while producing slower changes in expertise, attention, and organizational memory. Governance must therefore ask not only whether the system improves today's output, but whether its continued use preserves tomorrow's capacity for independent judgment.

D. Cognitive friction as a governance resource

If cognitive capacity can be weakened by poorly designed assistance, then some forms of friction should be understood as governance resources rather than inefficiencies.

A cognitive friction is any deliberately preserved point at which a human must actively interpret, decide, compare, justify, or challenge rather than passively accept an artificial output. Examples may include requiring the user to formulate an initial issue statement before receiving an AI-generated one; presenting competing legal hypotheses rather than a single recommendation; requiring verification of cited authorities before a draft can proceed; forcing

³³ Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), art. 14, 2024 O.J. (L) 2024/1689.

explicit confirmation of high-impact factual assumptions; or separating research and drafting agents so that the latter cannot silently transform uncertain research into categorical language.

Such mechanisms resemble the “cognitive forcing functions” studied in human-AI interaction, but the judicial context expands their purpose.³⁴ The objective is not only to reduce immediate overreliance. It is also to preserve institutional authorship and professional competence. A properly calibrated friction can require the responsible human to remain intellectually present at the points where the legal meaning of the case is constructed.

This produces an important normative consequence. Friction should not be maximized. Excessive friction can make systems unusable, undermine efficiency, and lead users to circumvent official tools through unsanctioned “shadow AI.” The relevant task is calibration. Low-risk and reversible tasks may justify minimal friction. High-risk tasks, or tasks that structure later stages of reasoning, may require stronger forms of human engagement.

The same principle applies to organizational design. An institution might permit automated document classification while requiring human validation of contested factual chronologies. It might allow an agent to propose case-law clusters while preventing it from selecting the legally controlling line without explicit confirmation. It might permit draft generation only after the judge records the dispositive direction and core reasons. These are not universal prescriptions. They illustrate a broader proposition: the architecture of delegation should reflect the cognitive significance of the task, not merely its technical automatability.

E. A task-based map of the reorganized cognitive space

The reorganization of judicial cognitive space can be made more precise by distinguishing four categories of institutional activity.

The first includes automation-dominant tasks: high-volume, low-discretion operations such as formatting, document conversion, metadata extraction, duplication detection, scheduling support, and other functions whose legal significance is limited and whose outputs are easily reversible.

The second includes AI-dominant but human-verified tasks: synthesis of large records, identification of candidate precedents, extraction of quotations, comparison of documents, or generation of alternative organizational structures. Here artificial systems may perform most of the operational work,

³⁴ Zana Buçinca, Maja Barbara 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, no. CSCW1, art. 188, at 1–21 (2021), <https://doi.org/10.1145/3449287>; Luis Felipe Salomão & Dierle Nunes, O Jurista que Parou de Tentar: IA, Dependência Cognitiva e Como o Design Importa, ConJur (May 8, 2026), <https://www.conjur.com.br/2026-mai-08/o-jurista-que-parou-de-tentar-ia-dependencia-cognitiva-e-como-o-design-importa-2/>.

but institutional reliability depends on verification protocols and source access.

The third includes genuinely hybrid tasks: issue framing, assessment of competing arguments, construction of legal hypotheses, evaluation of evidentiary significance, and iterative drafting of reasons. These tasks may benefit from AI contributions, but their value lies in interaction rather than unilateral delegation. They are precisely the zone in which cognitive friction and institutional curation become most important.

The fourth includes human-reserved functions. The category should not be defined by claims that machines are metaphysically incapable of performing them. Rather, it should be defined institutionally. In adjudication, the determination of legitimate purposes, the assumption of public responsibility, the exercise of constitutional authority, the final acceptance or rejection of coercive consequences, and the production of reasons that can be attributed to the court remain human functions because democratic legal orders require human institutions to answer for them.

This four-part map differs from the familiar automation-versus-augmentation distinction. It asks not merely who performs a task, but what institutional meaning follows from relocating it. The same technical capability may be acceptable in one part of a workflow and unacceptable in another. A model may draft a neutral chronology with limited concern, yet the same model's silent selection of which facts deserve inclusion may alter the substantive frame of the dispute. A search agent may identify authorities efficiently, yet its ranking criteria may invisibly determine which doctrinal pathway receives attention.

The central object of governance is therefore the boundary between cognitive functions. That boundary should remain revisable because the technology, professional practices, and evidence about human-AI interaction will change. What must remain stable is the institutional principle: delegation is legitimate only when the institution preserves sufficient competence, traceability, and authority to understand what has been delegated and to reclaim the function when necessary.

The analysis so far has focused on the redistribution of cognition between humans and individual AI systems. Agentic AI introduces a further transformation. Once multiple artificial agents can divide cognitive labor among themselves, cognition is redistributed not only between human and machine, but within artificial architectures. That development requires a distinct analytical vocabulary.

III. FROM INDIVIDUAL AI TO ARTIFICIAL COLLECTIVE INTELLIGENCE

A. Agents, multi-agent systems, and the artificial division of cognitive labor

The contemporary development of agentic AI marks a shift from models that generate content to systems that organize action. Multi-agent systems extend that transformation by distributing action and cognition across multiple artificial components. In technical terms, the field encompasses a wide range of designs: centralized or decentralized coordination, fixed or dynamic roles, cooperative or competitive interaction, shared or separated memory, hierarchical orchestration, and iterative critique.³⁵ What unifies these architectures is that the relevant unit is no longer a single model producing a single answer. The relevant unit is a system of interacting agents whose collective process generates the result.

The analogy with organizations is difficult to ignore. Human institutions increase capacity by dividing labor, assigning roles, creating channels of communication, and building mechanisms of review. Courts themselves are structured in this way. Judicial production is rarely the product of a solitary mind. Clerks, advisers, administrative staff, expert evidence, precedent databases, appellate structures, and collegial deliberation all distribute the production and correction of legal knowledge.

Multi-agent AI brings a computational analogue of this organizational logic. In CAMEL, for example, agents are assigned roles and use structured communication to cooperate on tasks.³⁶ ChatDev organizes specialized agents into role-based production processes inspired by software-development organizations.³⁷ Generative Agents demonstrated how memory, reflection, planning, and interaction could produce collective behavior not reducible to a single static instruction.³⁸ These projects do not establish that artificial collectives replicate human institutions in any strong sociological sense. They do, however, demonstrate that capability can emerge from architecture as well as from the strength of the individual model.

That point has major implications for law. If performance increasingly depends on coordination among specialized agents, then the legal object of

³⁵ Taicheng Guo et al., *supra* note 23.

³⁶ Guohao Li et al., CAMEL: Communicative Agents for “Mind” Exploration of Large Language Model Society, in *Advances in Neural Information Processing Systems* 36 (2023), <https://proceedings.neurips.cc/paper/2023/hash/a3621ee907def47c1b952ade25c67698-Abstract-Conference.html>.

³⁷ Chen Qian et al., *ChatDev: Communicative Agents for Software Development*, in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* 15174, 15174–86 (2024), <https://doi.org/10.18653/v1/2024.acl-long.810>.

³⁸ Joon Sung Park et al., *Generative Agents: Interactive Simulacra of Human Behavior*, in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* art. 2 (2023), <https://doi.org/10.1145/3586183.3606763>.

analysis cannot remain confined to the properties of a single model. Questions about bias, reliability, explainability, and safety must be extended to interaction rules, role assignment, communication protocols, memory structures, authority hierarchies, and conflict-resolution mechanisms.

A judicial multi-agent system could, for example, allocate separate functions to a factual-analysis agent, a statutory-research agent, a precedent-retrieval agent, a counterargument agent, a citation-verification agent, and a drafting agent. An orchestrator could sequence their work and determine when the task is complete. The final output might be superior to the result of a single undifferentiated prompt because specialization allows different components to focus on different failure modes.

But the same decomposition makes responsibility harder to reconstruct. If the final draft omits a decisive authority, was the failure caused by retrieval, ranking, orchestration, communication, memory, or drafting? If an intermediate agent incorrectly labels a factual issue as uncontested, downstream agents may rationally build on that premise. Error becomes architectural.

This is the first major lesson of artificial collective intelligence: the quality of an AI-assisted judicial outcome increasingly depends on the design of the deliberative system that produces it.

B. Architecture as a source of epistemic gain

Multi-agent systems are attractive in part because they can introduce forms of diversity and internal challenge that a single model lacks. A single model often produces one coherent path through a problem. Multiple agents can generate competing paths and compare them.

Research on multi-agent debate illustrates the potential. Du and co-authors showed that multiple language-model instances proposing and debating answers across several rounds could improve factuality and reasoning in a range of tasks.³⁹ The mechanism is important: the gain does not come only from adding more computation. It comes from structured disagreement. One agent's initial answer becomes an object of criticism by another, and the system is given opportunities to revise before settling on a final response.

Other work reaches a related conclusion through reflection rather than inter-agent debate. Reflexion uses verbal feedback and episodic memory to allow agents to learn from prior failed attempts without retraining model

³⁹ Yilun Du et al., *Improving Factuality and Reasoning in Language Models Through Multiagent Debate*, 235 *Proceedings of Machine Learning Research* 11733, 11733–63 (2024). <https://proceedings.mlr.press/v235/du24e.html>.

weights.⁴⁰ Although self-reflection and multi-agent disagreement are technically different, they share a governance-relevant insight: intermediate reasoning processes can be designed to expose outputs to deliberate correction rather than treating the first response as final.

For law, this is immediately recognizable. Legal procedure has long institutionalized the intuition that first impressions should be challenged. Parties contest each other's factual and legal claims; appellate review tests lower-court decisions; collegial bodies expose individual views to disagreement; dissents preserve alternative interpretations; procedural burdens force claims to be substantiated. These mechanisms are not infallible, but they treat contestation as epistemically productive.

The emergence of multi-agent systems therefore invites a provocative question: can artificial systems incorporate an internal analogue of legal contestation?

C. Artificial adversariality

This article uses the term artificial adversariality to describe an architecture in which specialized artificial agents are deliberately assigned different epistemic functions so that intermediate conclusions are exposed to structured challenge before a final output is produced.

Artificial adversariality is not a digital version of constitutional due process. It does not satisfy the right to be heard, does not represent the parties, does not create procedural equality, and cannot replace adversarial participation by persons whose rights are affected. The concept is narrower and epistemic. It refers to internal system design.

A minimal form of artificial adversariality may involve a generator and a critic. A stronger form may include several agents with distinct roles: one reconstructs the factual record; another searches for counterevidence; one develops the strongest argument for a proposed legal characterization; another is instructed to identify the best doctrinal objection; a separate verifier checks quotations and sources; and an orchestrator records unresolved disagreement rather than suppressing it.

The concept is useful for three reasons. First, it recognizes that disagreement can be engineered. AI governance is often framed in terms of preventing error through better models. Artificial adversariality adds a different strategy: assume that any agent may err and design the system so

⁴⁰ Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R. Narasimhan & Shunyu Yao, *Reflexion: Language Agents with Verbal Reinforcement Learning*, in *Advances in Neural Information Processing Systems* 36 (2023), https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

that relevant errors are more likely to encounter resistance before reaching the final output.

Second, it makes architecture legally visible. A court evaluating an AI system should not ask only which model it uses. It should also ask whether the system contains mechanisms that challenge its own intermediate conclusions. A more capable model operating in a single-path workflow may, in some contexts, be less institutionally reliable than a somewhat weaker model embedded in a robust architecture of criticism and verification.

Third, artificial adversariality provides a bridge between technical design and legal theory. Law does not merely produce outcomes; it develops procedures for producing outcomes under disagreement. The analogy should not be exaggerated, but it suggests that institutional values can sometimes be translated into design requirements. A system intended to assist adjudication should not be optimized only for fluent convergence. It may need to preserve visible dissent, uncertainty, and competing hypotheses.

This point is especially important because large language models have a strong capacity to produce coherent narratives. Coherence is useful, but it can create a false appearance of settledness. In difficult cases, the institution may be better served by a system that surfaces unresolved conflict than by one that synthesizes disagreement into a smooth answer. Artificial adversariality should therefore be designed not only to reach agreement but also to preserve legally material disagreement when consensus would be epistemically premature.

D. The limits of artificial disagreement

Artificial adversariality should not be romanticized. Multi-agent debate can improve performance in some tasks, but it does not guarantee independence of perspectives. Agents based on the same underlying model may share blind spots, training-data biases, or systematic errors. If all agents search the same database, rely on the same embedding model, or inherit the same faulty factual representation, apparent disagreement may conceal deep homogeneity.

Cemri and co-authors' analysis of multi-agent traces offers a concrete taxonomy for this caution. Their 2025 preprint identifies fourteen failure modes grouped around system design, inter-agent misalignment, and verification. Its tasks and frameworks are not judicial deployments, so the taxonomy is a diagnostic resource rather than a measured judicial failure rate. It nevertheless suggests a testable requirement for artificial adversariality: courts should examine whether agents preserve task constraints, transmit

necessary information, and verify completion, rather than infer reliability merely from the presence of several specialized roles.⁴¹

There is also a risk of manufactured confidence. A system may report that several agents independently reached the same result even though their outputs are highly correlated because they share architecture, training data, prompts, or memory. In human institutions, independent review has value partly because reviewers can possess genuinely distinct information, incentives, and experiences. Artificial systems can simulate role diversity without necessarily achieving epistemic independence.

The architecture can also create escalation problems. Agents may recursively criticize one another, consume additional resources, and still fail to resolve disagreement. An orchestrator must decide when to stop. That stopping rule is normatively relevant. If the system terminates debate because a numerical confidence threshold has been reached, the threshold itself becomes part of the decision architecture. If an orchestrator suppresses minority-agent objections to produce a concise output, the system may erase precisely the information a judge most needs to see.

Furthermore, multi-agent systems can produce emergent behavior. Chan and co-authors argue that increasingly agentic systems should be understood through characteristics such as underspecification, directness of impact, goal-directedness, and long-term planning, and they emphasize that increasing agency can produce harms not fully captured by traditional model-level analysis.⁴² In multi-agent environments, those risks are compounded by interaction. The behavior of the whole may not be predictable from the specification of each component considered separately.

This does not imply that emergent behavior is necessarily uncontrolled or dangerous. It means that validation must occur at the system level. Testing each agent individually is insufficient if the material behavior arises from their coordination.

A suitable evaluation should therefore compare the proposed multi-agent workflow with a simpler single-agent workflow under comparable model, retrieval, time, and resource conditions. Otherwise, any apparent advantage may reflect more computation or better source access rather than the value of disagreement itself. The normative proposal is conditional: artificial adversariality merits deployment only where its additional complexity produces a demonstrable benefit for the intended task without weakening responsibility or contestability.

⁴¹ Mert Cemri et al., *Why Do Multi-Agent LLM Systems Fail?* (arXiv preprint, version 3, Oct. 26, 2025), <https://doi.org/10.48550/arXiv.2503.13657>.

⁴² Alan Chan et al., *Harms from Increasingly Agentic Algorithmic Systems*, in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* 651, 651–66 (2023), <https://doi.org/10.1145/3593013.3594033>.

E. Distributed cognition and distributed accountability

The legal difficulty becomes sharper when cognitive production is distributed but legal responsibility is not. Public institutions cannot simply mirror technical decomposition in their accountability structure. A court cannot answer a litigant's challenge by explaining that an error was an emergent product of several agents whose individual contributions cannot be isolated.

This creates a fundamental asymmetry. Cognition may be distributed; accountability cannot be allowed to dissolve.

The principle has two implications. The first is architectural. Multi-agent systems intended for judicial use must generate records that allow reconstruction of material interactions. It may not be necessary (or even useful) to preserve every token generated in every internal exchange indefinitely. But the institution should be able to identify which agents performed material functions, which sources were accessed, which intermediate conclusions affected later stages, what conflicts arose, how they were resolved, and where a human intervened.

Reconstruction here means an evidentiary record of operations: source identifiers, retrieved passages, tool calls, permissions, material handoffs, versions, and human interventions. It does not mean treating a model-generated explanation of its own reasoning as a faithful causal account. Nor should auditability require indiscriminate publication of confidential case material. A governance protocol should distinguish operational records, materials available for appropriately controlled review, and the public legal grounds of the judgment. These are different records serving different purposes.

The second is organizational. Responsibility must be assigned before deployment rather than improvised after failure. Judicial institutions should identify who is responsible for approving use cases, configuring permissions, validating databases, monitoring system behavior, reviewing incidents, authorizing updates, and ultimately accepting the use of AI-generated work in the exercise of judicial authority.

This distinction between technical causation and institutional responsibility is essential. Complex systems frequently involve many causal contributors. Law nevertheless assigns responsibility through institutional roles. Agentic AI should not be treated differently. If anything, greater technical complexity strengthens the case for ex ante responsibility design.

The transition from individual AI to artificial collective intelligence therefore changes the object of governance once again. The central issue is no longer only whether a model is accurate, nor only whether a human reviews the final output. Governance must reach the architecture of artificial deliberation itself.

That requirement leads to the next question. What kind of institutional mechanisms can preserve direction, contestability, and accountability when AI systems are able not only to generate content but to coordinate sequences of action and artificial deliberation? The answer proposed here is an architecture of governance built around cognitive friction, institutional reins, and meaningful human initiative.

IV. GOVERNING AGENTIC JUDICIAL SYSTEMS: FROM COGNITIVE FRICTION TO INSTITUTIONAL REINS

A. Autonomy as a governance problem

The governance challenge created by agentic AI is not autonomy in the abstract. Courts have long relied on actors with delegated discretion. Clerks organize files, staff prepare research, experts contribute specialized knowledge, and administrative units make operational choices without constant judicial instruction. The institutional problem arises when autonomy is combined with opacity, scale, speed, and weakly defined responsibility.

Agentic systems intensify that combination. An agent may be given a broad objective and then determine intermediate steps that were not individually specified by the user. A multi-agent system may distribute those steps among specialized components. The resulting workflow can be efficient precisely because continuous human authorization is not required. Yet that efficiency creates a tension: the more a system can proceed without intervention, the less meaningful a nominal human checkpoint may become unless the surrounding governance architecture is strong.

The risks are not exhausted by incorrect outputs. Increasing agency expands the range of ways in which systems can produce legally significant effects. Chan and co-authors emphasize that agency should not be treated as a binary property and identify characteristics such as underspecification, direct impact, goal-directedness, and long-term planning as dimensions that can increase agentic risk.⁴³ Shavit and co-authors similarly argue that agentic systems require governance practices addressing actors across the system life cycle, including responsibilities, permissions, monitoring, accountability, and the ability to interrupt harmful activity.⁴⁴

In the judicial domain, these concerns acquire a constitutional dimension. A system that merely drafts text can still create serious problems, but its action space is comparatively narrow. A system that can retrieve data, communicate with external services, modify files, trigger procedural actions, coordinate subordinate agents, or persist across time can affect legal

⁴³ Alan Chan et al., *supra* note 42.

⁴⁴ Yonadav Shavit et al., *supra* note 24.

processes before a human sees the final output. Governance must therefore focus not only on the quality of recommendations but on the scope of authorized action.

This suggests a basic principle: the greater the operational autonomy of the system, the more explicit the institution must be about objectives, permissions, escalation rules, logging, and interruption.

B. Why human oversight is not enough

Human oversight remains indispensable, but it is not self-executing. The assumption that a human can correct whatever an AI system does wrong is one of the most fragile foundations of contemporary AI governance.

Green's critique of government-algorithm oversight is important precisely because it separates the normative appeal of human review from its empirical effectiveness. Human overseers can over-rely on recommendations, misunderstand system limitations, lack access to relevant information, or be structurally pressured to approve machine outputs at scale.⁴⁵ Research on automation bias and cognitive forcing reinforces the same point. Oversight must be designed, resourced, and institutionally supported.

The EU AI Act reflects this more demanding approach. For high-risk systems, Article 14 requires that oversight measures enable natural persons to understand relevant capabilities and limitations, monitor operation, remain aware of automation bias, interpret outputs, disregard or reverse them, and intervene or stop the system.⁴⁶ These requirements are significant because they treat human oversight as a set of capacities rather than the mere existence of a person somewhere in the decision chain.

The concept of meaningful human control developed in the ethics of autonomous systems offers a related contribution. Santoni de Sio and van den Hoven propose "tracking" and "tracing" conditions: systems should respond to relevant human reasons, and outcomes should remain traceable to human agents capable of bearing responsibility.⁴⁷ This framework is highly relevant to judicial AI because it links technical design to moral responsibility.

Yet the judicial context requires an additional step. A court can satisfy a formal tracing condition—there is always a judge whose name appears on the decision—while still allowing the cognitive architecture to become substantively machine-directed. The judge may remain legally responsible

⁴⁵ Ben Green, *supra* note 32.

⁴⁶ Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), art. 14, 2024 O.J. (L) 2024/1689.

⁴⁷ Filippo Santoni de Sio & Jeroen van den Hoven, *Meaningful Human Control over Autonomous Systems: A Philosophical Account*, 5 *Frontiers Robotics & AI* 15 (2018), <https://doi.org/10.3389/frobt.2018.00015>.

but exercise little meaningful initiative if objectives, issue framing, information selection, and argumentative structure are substantially generated upstream by artificial agents.

For this reason, the present article distinguishes human oversight from meaningful human initiative. Oversight is often implemented reactively, through review, correction, or interruption after a system has acted. However, established oversight and meaningful-control frameworks also include anticipatory design. The distinction proposed here is therefore an emphasis within governance, not a claim that prior scholarship overlooks *ex ante* control.

Initiative emphasizes the constitutive choices that judicial institutions must retain: why a system is used, which questions it may pursue, what may be delegated, which constraints govern it, and what legal grounds the court ultimately adopts. These choices should remain revisable throughout the workflow.

This distinction does not render oversight obsolete. Meaningful human initiative includes oversight but goes further. It seeks to preserve human normative authorship throughout the architecture rather than at a single review point.

C. The institutional rein

To operationalize this idea, an institutional rein is the integrated set of normative, technological, and organizational mechanisms through which a judicial institution directs, constrains, monitors, audits, interrupts, and, when necessary, reclaims functions delegated to agentic AI systems.

The metaphor is deliberate. A rein does not immobilize. Its purpose is to permit movement while preserving direction and the possibility of correction. Applied to AI governance, the concept rejects two equally inadequate extremes: the assumption that innovation requires unrestricted autonomy, and the assumption that risk can be managed only by prohibiting meaningful delegation.

The institutional rein is composed of at least seven elements.

First, purpose specification. Agentic systems should operate within institutionally defined purposes rather than open-ended optimization. In judicial settings, “assist in deciding cases faster” is not a sufficient purpose. The institution must distinguish legitimate objectives—such as improving retrieval, organizing records, or supporting consistency—from outcomes that cannot become machine-defined objectives, such as maximizing conviction rates, minimizing appeals, or optimizing throughput regardless of procedural quality.

Second, permission boundaries. An agent should have only the tools, databases, communication channels, and action rights necessary for the

specific use case. A research agent that needs access to a jurisprudential database does not necessarily need the capacity to modify case files. A drafting agent does not need authority to file an order. Least-privilege design is therefore not merely a cybersecurity principle; it is a constitutional governance principle when technical permissions determine the practical reach of delegated public power.

Third, cognitive checkpoints. Institutions should identify points at which human judgment must occur before the workflow can continue. These checkpoints can be calibrated to risk and cognitive significance. A system may proceed autonomously through formatting and retrieval but pause before resolving contested factual classifications or before converting research into a dispositive recommendation.

Fourth, traceability. Material steps should be reconstructable. Logs should identify relevant sources, tools, agents, intermediate outputs, revisions, and human interventions. Traceability should be designed for institutional use, not merely for technical debugging. A record so voluminous that no responsible actor can understand it does not produce practical accountability.

Fifth, interruptibility and escalation. Systems should contain mechanisms for safe interruption, and users should know when they are required to escalate uncertainty. The ability to stop an agentic workflow is meaningless if the human does not know that a dangerous deviation has occurred. Monitoring and escalation thresholds are therefore part of the rein.

Sixth, institutional learning. Governance must incorporate incident reporting, periodic reassessment, model or prompt changes, newly discovered vulnerabilities, and shifts in actual use. Agentic systems evolve not only because models are updated but because users develop new practices around them. A governance regime that evaluates a system only at procurement is structurally obsolete.

Seventh, competence preservation. The institution must preserve sufficient human expertise to supervise, challenge, and, when necessary, perform critical functions without AI. Training should not be limited to teaching users how to operate tools. It should also teach failure modes, verification practices, and the boundaries of appropriate reliance.

These seven elements convert the institutional rein from metaphor into governance architecture. They also show why governance cannot be reduced to a single principle such as transparency. A transparent system can still have excessive permissions. A well-logged system can still create deskilling. A human-review requirement can still be meaningless under severe time pressure. Governance is systemic.

Prompt injection makes the relationship between permissions and authority particularly concrete. An agent can encounter instructions

embedded in a retrieved page or case-file attachment that attempt to redirect its conduct. In the judicial setting, the design implication is to treat retrieved material as evidence to analyze, not as authority to change permissions. Separation of trusted instructions from case content, restricted tool access, and authorization checks outside the model should accompany source verification.⁴⁸

Recent security evidence also cautions against importing benchmark scores into procurement decisions. Zhuowen Liu's October 2026 preprint finds poor transfer of prompt-injection detector rankings across agent benchmarks. This is preliminary evidence, not a universal estimate of detector failure. It supports evaluating controls on the actual kinds of tool outputs and documents a court expects to process, including benign material that may trigger false alarms. A detector can be part of a defense, but should not be treated as proof that an agent is safe to exercise broader permissions.⁴⁹

D. Resolution No. 615/2025 as a Brazilian governance paradigm

Brazilian National Council of Justice Resolution No. 615/2025 is an unusually important reference for this discussion because it establishes a judicial-specific governance framework for AI. The Resolution explicitly addresses the development, use, and governance of AI solutions in the judiciary and replaced the earlier Resolution No. 332/2020.⁵⁰ It was amended in 2026 by Resolution No. 674, which modified the composition of the National Committee on Artificial Intelligence in the Judiciary but left the broader regulatory architecture intact.⁵¹

The Resolution's significance is not limited to individual prohibitions. Its broader contribution lies in institutionalizing a risk-based governance logic. It recognizes the need for transparency, auditability, human supervision and review, security, accountability, and institutional structures capable of monitoring AI use. High-risk systems are subjected to stronger controls, while certain uses are prohibited. The framework also requires public cataloging of AI applications and creates a national governance committee.⁵²

⁴⁸ Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer & Florian Tramèr, *AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents* (arXiv, version 3, Nov. 24, 2024), <https://doi.org/10.48550/arXiv.2406.13352>.

⁴⁹ Zhuowen Liu, *Passing the Test You Trained On: Re-evaluating Prompt-Injection Detectors for LLM Agents* (arXiv preprint, Oct. 2, 2026), <https://doi.org/10.48550/arXiv.2610.03448>.

⁵⁰ Conselho Nacional de Justiça [CNJ], Resolução No. 615, de 11 de março de 2025, DJe/CNJ No. 54/2025, at 2–17 (Mar. 14, 2025), <https://atos.cnj.jus.br/atos/detalhar/6001>.

⁵¹ Conselho Nacional de Justiça [CNJ], Resolução No. 674, de 25 de março de 2026, DJe/CNJ No. 71/2026, at 2 (Mar. 26, 2026), <https://atos.cnj.jus.br/atos/detalhar/6836>.

⁵² Conselho Nacional de Justiça [CNJ], Resolução No. 615, de 11 de março de 2025, as amended by Resolução No. 674, de 25 de março de 2026, compiled text, <https://atos.cnj.jus.br/files/compilado1243232026032669c529ebd7bbb.pdf>; Conselho Nacional

Tamer's doctrinal analysis situates the Resolution within international principles of human supervision, proportionality, risk assessment, and data protection. That reading supports treating the instrument as a substantial governance framework rather than a merely aspirational declaration. The argument advanced here is more specific: agentic composition requires implementation choices about the unit of assessment, changing permissions, intermediate records, and the preservation of contestation. Those choices should be presented as applications or proposed elaborations of the framework, rather than attributed to provisions that do not expressly prescribe them.⁵³

This architecture is particularly relevant in light of the speed of AI adoption in Brazilian courts. The National Council of Justice's 2024 survey of magistrates and court staff already documented extensive interest in, and use of, generative AI.⁵⁴ The subsequent mapping of projects showed a rapid institutional expansion, with 157 initiatives across forty-three courts and a strong concentration of projects launched in 2024 and 2025.⁵⁵ This is precisely the environment in which governance rules must move from abstract principle to organizational practice.

Resolution No. 615/2025 provides several building blocks for the institutional rein. Risk classification helps calibrate controls. Requirements of human supervision and review support cognitive checkpoints. Audit and impact mechanisms reinforce traceability. Institutional committees create organizational responsibility. Public catalogs can improve external accountability. Restrictions on systems that do not permit human review acknowledge that technical design can make legal oversight impossible.

The doctrinal anchors should be explicit. Article 19, § 3, II, preserves judicial orientation, interpretation, verification, review, and responsibility when generative AI is used; Article 32 protects internal users' autonomy. The Annex's low-risk category BR4 concerns supporting texts produced under judicial supervision and instructions, including instructions on preliminary and merits issues. It does not justify treating an independently framed dispositive recommendation as low risk merely because a judge later signs it.

de Justiça [CNJ], Comitê Nacional de Inteligência Artificial do Judiciário (CNIAJ), <https://www.cnj.jus.br/sistemas/plataforma-sinapses/comite-nacional-de-inteligencia-artificial-do-judiciario-cnaij/>.

⁵³ Maurício Antonio Tamer, *Regulação da Inteligência Artificial no Judiciário Brasileiro: Análise do Objeto e Definições da Resolução CNJ n. 615/2025*, 36 Rev. TRF3, no. 162, 17–36 (2025), <https://doi.org/10.65674/rev-trf3.v36i162.742>.

⁵⁴ Conselho Nacional de Justiça [CNJ], *O Uso da Inteligência Artificial Generativa no Poder Judiciário Brasileiro: Relatório de Pesquisa* (Brasília: CNJ 2024), ISBN 978-65-5972-158-0, <https://bibliotecadigital.cnj.jus.br/jspui/handle/123456789/976>.

⁵⁵ Conselho Nacional de Justiça [CNJ], *Mapeamento de IA no Poder Judiciário 2* (2025), <https://www.cnj.jus.br/wp-content/uploads/2025/12/relatorio-mapeamento-projetos-iniciativas-inteligencia-artificial.pdf>.

The practical question is whether the actual workflow satisfies the conditions attached to its classification.⁵⁶

At the same time, agentic and multi-agent systems test the boundaries of a framework largely written at the level of “AI solutions,” “models,” and “systems.” The challenge is not that the Resolution is defective. The challenge is that a system capable of orchestrating multiple agents introduces new questions that conventional risk categories may not answer cleanly.

Consider, hypothetically, a judicial application with separate components for extracting facts, retrieving law, drafting, criticism, and citation checking. Their labels do not determine their regulatory status. If the integrated workflow evaluates disputed evidence or silently frames the dispositive grounds, classification must follow its actual function and applicable Annex criteria. The problem is therefore not an exemption for “six low-risk agents,” but the risk of evaluating components in isolation while overlooking what their combination is authorized to do.

A similar problem arises with dynamic tool use. An agent may be low risk when confined to an internal database but acquire a different risk profile when authorized to access external sources, send messages, modify records, or invoke other agents. Risk is partly a function of permissions and context, not only model identity.

Shared memory creates another grey area. A multi-agent system may store intermediate representations or summaries that later agents treat as authoritative. If the shared memory contains a mistaken characterization, the error can propagate through the architecture even when each downstream agent performs correctly relative to its inputs. Governance must therefore ask how shared memory is created, validated, revised, and retained.

Finally, agentic systems complicate the concept of “human review.” What exactly is the human expected to review? The final text? The sources? The intermediate plan? The actions taken? The interaction among agents? The answer cannot be “everything” in complex systems, because exhaustive review would eliminate the efficiency that motivates agentic delegation. The governance task is to identify which points are legally material enough to require human attention.

These examples support a broader conclusion: Resolution No. 615/2025 should be understood not as a finished answer to AI governance but as an institutional platform capable of supporting more specific practices for agentic systems. Its enduring value will depend on adaptive interpretation, technical standards, governance guidance, and organizational learning.

⁵⁶ Conselho Nacional de Justiça [CNJ], Resolução No. 615, de 11 de março de 2025, as amended by Resolução No. 674, de 25 de março de 2026, art. 19, § 3, II; art. 32; Annex, BR4, <https://atos.cnj.jus.br/atos/detalhar/6001>.

Transparency also requires a distinction between the existing rule and the reform proposed here. Article 33, § 3, makes mention of AI use in the text of a judicial decision discretionary, subject to its cross-references; Article 19, § 6, addresses internal recording. A general duty to publish every prompt cannot be inferred from those provisions. This article instead argues for proportionate access to the information necessary to contest a materially relevant AI-mediated premise. The appropriate disclosure depends on procedural relevance, confidentiality, and the applicable legal basis.⁵⁷

The constitutional basis is independent of the technology. Article 5, XXXV, LIV, and LV, of the Brazilian Constitution protects access to judicial protection, due process, and adversarial participation; Article 93, IX, requires reasoned judicial decisions. In civil procedure, Articles 9 and 10 protect participation and prohibit decisions based on grounds on which the parties had no opportunity to comment, subject to the statutory framework. Article 489, § 1, specifies deficiencies of judicial reasoning. Accordingly, an internal agent debate cannot substitute for participation by affected persons. If an AI-generated ground becomes material to the decision, the procedural question concerns the parties' opportunity to address that ground, not whether another agent has criticized it.⁵⁸

Technical traceability and legal justification should therefore remain distinct. A complete execution log may show how a draft was generated without demonstrating why its conclusion follows from the applicable law and the record. Conversely, a persuasive judgment does not cure an unauthorized disclosure of confidential data or a procedurally impermissible basis for decision. The framework demands both operational accountability and legally sufficient grounds, with neither used as a substitute for the other.

E. Regulating is not governing

The distinction between regulation and governance is central. Regulation establishes *ex ante* normative boundaries. It tells institutions what is permitted, required, restricted, or prohibited. Governance is the continuing organizational activity through which those boundaries are translated into real systems and revised in light of experience.

An institution can comply formally with a regulation and still govern poorly. It may classify a system once and never reassess it after major updates. It may maintain a nominal human-review requirement while

⁵⁷ Conselho Nacional de Justiça [CNJ], Resolução No. 615, de 11 de março de 2025, as amended by Resolução No. 674, de 25 de março de 2026, art. 19, § 6; art. 33, § 3, <https://atos.cnj.jus.br/atos/detalhar/6001>.

⁵⁸ Constituição Federal [C.F.] [Constitution], arts. 5, XXXV, LIV–LV, and 93, IX (Braz.), https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Lei No. 13.105, de 16 de março de 2015 [Código de Processo Civil], arts. 9–11 and 489, § 1 (Braz.), https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/113105.htm.

allocating only seconds per output. It may keep logs no one knows how to interpret. It may provide training focused on interface use rather than critical verification. It may prohibit unsanctioned tools but fail to provide functional alternatives, encouraging shadow AI.

The distinction matters because agentic AI is dynamic. Models change, tools are added, prompts evolve, permissions expand, users improvise, and vendors update services. A regulatory snapshot is therefore insufficient. Governance must be iterative.

The NIST AI Risk Management Framework captures this organizational orientation through its functions of Govern, Map, Measure, and Manage.⁵⁹ UNESCO's Recommendation on the Ethics of Artificial Intelligence similarly emphasizes that governance must integrate human rights, oversight, accountability, capacity building, and ongoing institutional processes rather than rely on isolated technical fixes.⁶⁰ These international frameworks are not judicial rules, but they reinforce a core proposition: trustworthy AI depends on institutions capable of continuous risk management.

The 2026 International AI Safety Report reinforces the need to manage uncertainty about capability and risk without assuming that a single intervention resolves it. In judicial governance, this means separating evidence that a control works under specified conditions from an institution's decision about acceptable residual risk. The report is a scientific synthesis, not a source of binding duties for Brazilian courts; its relevance is to the evidential discipline of ongoing assessment.⁶¹

For courts, continuous governance is especially important because institutional legitimacy can be damaged even when the underlying technical failure is minor. A fabricated citation, undisclosed use of AI in drafting, an unexplained data source, or an automated action taken without authority can undermine confidence in the process. Judicial governance must therefore be sensitive not only to statistical risk but to role legitimacy: whether the technology is performing a function compatible with the legal meaning of judicial authority.

F. Meaningful human initiative

The concept of meaningful human initiative is proposed as the normative center of the governance architecture. It's the institutionally preserved capacity of human actors to define purposes, set limits, frame the legally

⁵⁹ National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1 (Jan. 2023), <https://doi.org/10.6028/NIST.AI.100-1>.

⁶⁰ UNESCO, Recommendation on the Ethics of Artificial Intelligence (Nov. 23, 2021), <https://unesdoc.unesco.org/ark:/48223/pf0000381137>.

⁶¹ Yoshua Bengio et al., International AI Safety Report 2026, DSIT 2026/001 (2026), <https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026>.

relevant problem, require or suspend artificial assistance, challenge intermediate and final outputs, interrupt workflows, and assume public responsibility for the resulting exercise of institutional authority.

The concept contains five dimensions. The first is teleological initiative: humans define the legitimate ends. AI systems may optimize means, but they should not silently redefine the institutional objective. In courts, efficiency is subordinate to legality, due process, reason-giving, equality, and rights protection.

The second is epistemic initiative: humans retain the capacity to decide what questions require attention, what uncertainty matters, and what evidence or authority must be examined. This prevents AI systems from becoming invisible framers of the dispute.

The third is interruptive initiative: humans possess practical authority and technical means to pause, redirect, disregard, or terminate the workflow. This dimension connects meaningful human initiative to meaningful human control but emphasizes the surrounding institutional conditions that make interruption real.

The fourth is responsibility initiative: the institution identifies humans who can answer for the use of the system. Responsibility should not be reduced to liability after harm; it includes the *ex ante* duty to justify why delegation was appropriate.

The fifth is legitimating initiative: the final exercise of judicial authority must remain attributable to a human institution capable of publicly giving reasons. The court cannot outsource the normative act by which reasons become the court's reasons.

Meaningful human initiative is therefore stronger than the idea that a judge has the "final say." Final approval can be empty if the system has already framed the issue, selected the evidence, ranked the authorities, and drafted the result in a way that the human lacks time or competence to reconstruct. Initiative asks whether humans remain authors of the institutional direction of the process.

This conception also explains why cognitive friction, the institutional rein, and artificial adversariality are complementary rather than competing concepts. Artificial adversariality can improve the internal quality of artificial deliberation. The institutional rein governs the autonomy through which that deliberation occurs. Cognitive friction preserves points of human engagement. Meaningful human initiative provides the normative criterion for evaluating whether the architecture, taken as a whole, preserves human institutional authorship.

The next section turns from governance architecture to professional transformation. If meaningful human initiative becomes more important as artificial systems perform more cognitive execution, what becomes of the

role of judges and judicial teams? The answer proposed here is not replacement, but institutional AI curation.

V. INSTITUTIONAL AI CURATION AND THE FUTURE OF JUDICIAL WORK

A. Beyond the replacement narrative

Public discussion about AI and professions is often organized around a binary question: will machines replace humans? In the judiciary, the question is especially misleading. Courts are not simply collections of tasks. They are institutions authorized to exercise public power under legal constraints. Even if artificial systems become technically capable of performing many operations associated with judging, it does not follow that the institution can delegate the normative meaning of adjudication in the same way that it delegates document formatting or information retrieval.

A more useful framework begins with the reorganization of tasks. Ethan Mollick’s account of human-AI collaboration distinguishes different modes in which people may work alongside AI, including “centaur” and “cyborg” patterns that alternate or intertwine human and machine contribution.⁶² Those metaphors are valuable because they move beyond simple substitution. Yet judicial institutions require an additional layer: the human professional must not only collaborate with AI but also curate the institutional conditions under which artificial contributions become legally usable.

On the proposed account, the judge’s role is not exhausted by personally performing every preparatory task or by approving a completed draft. Judicial judgment must extend to defining the permitted purpose of assistance, identifying which premises require independent scrutiny, and critically adopting the grounds for decision. This is a normative allocation of responsibility, not a prediction that all courts will converge on an orchestration model.

B. Human capacities that remain institutionally non-delegable

The category of “non-delegable” functions requires care. It should not be based on predictions about permanent technological incapacity. Capabilities that appear uniquely human today may be simulated or approximated by future systems. The argument here is institutional rather than anthropological. Certain functions should remain human because their attribution to human officeholders is constitutive of democratic and legal legitimacy.

The first is the definition of institutional purposes. AI systems can optimize objectives, but the choice of what ought to be optimized is normative. A court exists to adjudicate according to law, protect rights,

⁶² Ethan Mollick, *Co-Intelligence: Living and Working with AI* (Portfolio 2024).

provide reasons, and resolve disputes through legitimate procedure. Productivity is relevant, but it cannot become the master objective from which all other values are derived.

The second is public responsibility. Judicial authority carries an obligation to answer for the use of coercive power. A person whose liberty, property, family relations, or legal status is affected is entitled to confront an institution that can explain and defend its action. Artificial systems can contribute reasons, but responsibility for adopting those reasons cannot be displaced into the technical architecture.

The third is moral and contextual judgment. Adjudication often involves the interpretation of open-textured standards, competing constitutional values, vulnerability, proportionality, credibility, and consequences that cannot be reduced to a stable optimization function. AI may illuminate relevant considerations, but the institution must retain human responsibility for deciding which considerations are normatively controlling.

The fourth is legitimation through reason-giving. Judicial reasons do not merely communicate a result. They connect public power to law by explaining why the outcome should be accepted as legally justified. A generated text becomes a judicial reason only when the court has critically adopted it as its own.

The fifth is the power to interrupt. Legitimate institutions must be able to refuse technological momentum. A judge must remain capable of saying that a system should not be used in a case, that an output is inadequate, that a workflow must be repeated, or that a type of delegation has become inconsistent with legal obligations.

These functions correspond to the concept of meaningful human initiative developed above. They also explain why professional transformation should be evaluated not by asking how much work remains human, but by asking which forms of human agency remain institutionally decisive.

C. Institutional AI curation

The concept of institutional AI curation captures the emerging competence through which human actors organize artificial contributions without surrendering institutional authorship.

Curation is not merely selection. In museums, archives, scholarship, and information systems, curation involves deciding what enters a collection, how it is classified, how relationships are constructed, what is excluded, and how the resulting whole is presented and interpreted. The analogy is useful for AI-assisted adjudication because the central human task increasingly concerns the management of cognitive inputs rather than their exclusive production.

Institutional AI curation includes at least six functions: (i) Selection: Choosing which systems, models, agents, databases, and tools are appropriate for a given institutional purpose; (ii) Composition: Deciding how different artificial and human capabilities are combined in a workflow. A research agent, citation verifier, and drafting agent should not simply be connected because they are available; their interaction should reflect the legal structure of the task; (iii) Validation: Determining which outputs require independent verification and what counts as sufficient verification. This includes distinguishing factual checking from legal evaluation; (iv) Rejection: Preserving the practical and cultural capacity to disregard artificial outputs. A curation regime fails if AI recommendations become default answers that humans rarely feel authorized to reject; (v) Contextualization: Integrating machine-generated content into the particular procedural, factual, and normative context of the case. An authority retrieved by an agent may be technically relevant yet institutionally inappropriate because of hierarchy, temporal context, procedural posture, or factual distinctions; (vi) Responsibility: Converting selected and validated artificial contributions into institutional work for which identifiable human actors remain answerable.

Institutional AI curation is therefore not a job title. It is a distributed competence that may be exercised by judges, advisers, court staff, technology teams, governance committees, and judicial schools at different levels. Its distribution should correspond to institutional role. Technology units may curate model access and permissions; governance committees may curate approved use cases; judicial schools may curate competence formation; judges and teams may curate the cognitive contributions used in individual cases.

The concept also clarifies why training is a governance issue. If judicial education focuses only on “how to prompt,” it risks producing more efficient users without producing more competent curators. Training should include system limitations, source verification, automation bias, privacy and security, risk classification, agentic permissions, incident reporting, and the legal significance of delegating particular cognitive functions.

D. From executor to orchestrator — without becoming a rubber stamp

The language of orchestration is useful but must be handled carefully. An “orchestrator judge” should not be imagined as someone who delegates all substantive thinking to artificial subordinates and merely chooses among outputs. Orchestration itself must remain cognitively substantive.

A competent orchestrator must know enough about the underlying task to recognize when the artificial process has gone wrong. This returns us to the supervision paradox. The future of judicial work cannot consist in eliminating all lower-level cognitive practice and expecting higher-level supervision to

remain magically intact. Expertise is cumulative. Institutional design must therefore preserve training environments in which new professionals develop independent legal reasoning before, and alongside, reliance on powerful AI systems.

This may require differentiated use according to professional experience. A senior judge and a newly appointed clerk need not have identical AI workflows. An experienced professional may safely delegate tasks that a novice still needs to perform personally for training purposes. Governance can therefore incorporate developmental friction: constraints designed not because the AI is unsafe, but because the human user needs to build or preserve competence.

The same principle applies to judicial teams. If AI absorbs first-draft research and writing, the institution should deliberately redesign team roles rather than simply reduce them. Staff may shift toward source verification, comparative legal research, identification of outlier cases, monitoring of agent behavior, and quality assurance. Judicial schools may need to teach not only legal doctrine and digital literacy but also how to supervise artificial divisions of cognitive labor.

Institutional AI curation thus provides a professional counterpart to the institutional rein. The rein governs the system; curation describes the human competence required to use that system responsibly.

E. A model of hybrid judicial intelligence

The concepts developed can be integrated into a model of hybrid judicial intelligence. At the operational layer, artificial systems perform tasks such as retrieval, synthesis, classification, comparison, drafting, and verification. At the deliberative layer, artificial adversariality exposes intermediate conclusions to structured challenge. At the governance layer, institutional reins constrain permissions, preserve traceability, require cognitive checkpoints, and enable interruption. At the human layer, meaningful human initiative determines purposes, frames problems, adopts or rejects artificial contributions, and assumes responsibility. Across all layers, institutional AI curation coordinates how these capacities are selected and combined.

The value of the model is that it does not require choosing between technological optimism and technological resistance. It accepts that artificial systems may become substantially more capable and that courts have legitimate reasons to use them. At the same time, it rejects the assumption that capability determines authority.

This distinction is fundamental. Artificial capability is a technical property; institutional authority is a legal construction. The judiciary may use systems that read faster than humans, search more comprehensively, remember more information, or generate multiple argumentative structures

in seconds. None of those capabilities answers the question of who may exercise judicial authority or how the exercise of that authority must be justified.

The most productive future is therefore neither a “human-only” judiciary nor a machine-directed one. It is a judiciary capable of converting artificial capability into institutional intelligence through governed distribution of cognition.

F. Evaluating the proposed architecture

The framework needs observable tests if it is to guide institutional practice. The following protocol is a design proposal derived from the preceding analysis, not an evaluated intervention or a claim of legal sufficiency. It treats the complete workflow as the unit of analysis and separates technical performance, professional competence, and procedural legitimacy. These dimensions should be reported separately: an aggregate score could conceal a serious rights-related failure behind gains in speed.

First, specify a bounded task and a defensible comparator. A pilot could evaluate preparation of a research memorandum from a closed record using a human-only process, a human-assisted single-agent process, and the proposed multi-agent process. Participants, source access, case complexity, model versions, and time allowances should be documented. Legally qualified reviewers should assess outputs without knowing which process produced them where feasible. Such a pilot would test the proposed contribution of orchestration; it would not establish the acceptability of autonomous adjudication.

Second, evaluate fidelity rather than fluency alone. Magesh and co-authors’ preregistered evaluation found that retrieval-based legal research tools could still generate false or inadequately grounded answers. The result concerns particular systems and a historical testing period, not all current products. For the proposed pilot, every material factual claim should be checked against the record and every legal proposition against the cited authority, including its jurisdiction, procedural context, and continuing relevance. Evaluators should record omissions of material contrary authority as well as fabricated citations. Source retrieval is a control to test, not a guarantee of correctness.⁶³

Third, test the exercise of human initiative. Relevant observations include whether reviewers detect deliberately seeded errors, whether they can recover the underlying source, whether a dissenting agent’s objection remains visible, and whether users can stop or redirect the workflow before an external action

⁶³ Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning & Daniel E. Ho, *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*, 22 J. Empirical Legal Stud. 216–242 (2025), <https://doi.org/10.1111/jels.12413>.

occurs. Review time, workload, and the consequences of refusing a recommendation should also be recorded. A low override rate has no fixed meaning: it may reflect reliable assistance, ineffective review, or organizational pressure. Evaluation must distinguish these explanations rather than reward agreement with the system.

Fourth, examine abstention, security, and changes over time. A system should expose when it lacks a source or cannot resolve a conflict, rather than convert uncertainty into a categorical recommendation. If it presents numerical confidence, that confidence should be evaluated against actual task performance rather than accepted as self-verifying. Regression tests should follow changes to models, prompts, retrieval, memory, or permissions. The NIST Generative AI Profile offers a cross-sectoral resource for organizing evaluation and risk treatment; specific judicial acceptance criteria still require institutional definition.⁶⁴

Consider an illustrative civil dispute in which an agent prepares a chronology and identifies possible precedents. The research component may retrieve authorities but cannot modify the record or communicate externally. A verifier links each material proposition to its source; a critic identifies counterarguments and missing facts; the orchestrator preserves unresolved issues. The judicial reviewer then checks the contested premises and determines the legal grounds to be pursued. A missing source or unresolved material conflict triggers escalation rather than automatic drafting. Any newly material ground must enter the applicable adversarial procedure. The example illustrates the proposed allocation of tasks and authority; it is not evidence of an existing deployment.

Finally, feasibility and proportionality constrain the proposal. Smaller courts may lack the staff to maintain complex agent architectures, and a simpler workflow may be preferable. Training, security, audit access, vendor dependence, and the ability to continue essential work during an outage belong in the institutional cost assessment. Competence preservation should be evaluated over time through appropriately designed unassisted tasks and qualitative assessment of professional learning, not inferred from immediate productivity. The framework is useful only if it helps an institution identify, test, and refuse delegations it cannot responsibly govern.

CONCLUSION

The history of judicial technology is often narrated as a sequence of increasingly sophisticated tools: digitized documents, electronic proceedings,

⁶⁴ National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1 (July 2024), <https://doi.org/10.6028/NIST.AI.600-1>.

automated routines, predictive systems, generative AI, and now autonomous agents. This article has argued that such a technological chronology, although useful, does not capture the most important transformation currently underway.

Generative and agentic AI systems do not merely accelerate pre-existing judicial work. They enter the cognitive infrastructure through which judicial work is produced. Reading, synthesis, research, comparison, drafting, verification, planning, and coordination can increasingly be distributed among human professionals and artificial systems. Multi-agent architectures extend the redistribution further by allowing cognitive functions to be divided among artificial components themselves.

Once cognition becomes distributed in this way, the legal problem changes. Accuracy remains necessary but is no longer sufficient. Governance must address where cognitive work occurs, how artificial contributions interact, which functions humans continue to perform, how errors can be reconstructed, who can interrupt the process, and who remains responsible for the exercise of public authority.

The article developed five analytical categories to structure that problem.

The first, reorganization of institutional cognitive space, describes the phenomenon itself: the redistribution of cognitive activities among judges, judicial teams, informational infrastructures, and artificial systems.

The second, artificial adversariality, identifies a possible deliberative architecture in which specialized artificial agents challenge, revise, and validate intermediate conclusions. The concept draws attention to the epistemic value of designed disagreement while insisting that artificial internal review cannot substitute for constitutional adversarial procedure or the parties' rights of participation.

The third, the institutional rein, describes the governance mechanisms necessary when systems possess operational autonomy. Purpose specification, permission boundaries, cognitive checkpoints, traceability, interruptibility, institutional learning, and competence preservation transform the abstract demand for "human oversight" into a more concrete governance architecture.

The fourth, meaningful human initiative, provides the normative principle that organizes the framework. Human presence at the end of an automated process is not enough. Judicial institutions must preserve the human capacity to define purposes, frame legally relevant problems, establish limits, interrupt workflows, challenge outputs, and publicly assume responsibility for the resulting exercise of authority.

The fifth, institutional AI curation, describes the corresponding transformation of professional work. Judges and judicial teams increasingly need to select, compose, validate, contextualize, reject, and assume

responsibility for artificial cognitive contributions. Curation is not passive approval. It is a higher-order form of institutional judgment.

Taken together, these concepts support the article's central thesis: AI does not eliminate legal thinking; it reorganizes the institutional space in which legal thinking occurs. The relevant transformation is therefore not adequately described as replacement. It is a redistribution of cognitive labor and a relocation of human responsibility.

This conclusion has direct implications for the Brazilian judiciary. National Council of Justice Resolution No. 615/2025, as amended by Resolution No. 674/2026, already provides an important governance infrastructure based on risk, supervision, transparency, auditability, and institutional responsibility. Its next challenge will be interpretive and organizational: applying those principles to systems whose risks arise not only from individual models but from orchestration, shared memory, dynamic tool use, emergent interaction, and distributed artificial decision steps.

The distinction between regulation and governance is decisive in that process. Rules can identify obligations, but only institutions can build the competences, routines, technical controls, training structures, incident mechanisms, and cultures of challenge that make those obligations real. Agentic AI will therefore test not only the quality of judicial regulation but the adaptive capacity of judicial institutions themselves.

Several questions remain open for empirical research. How do different forms of AI assistance affect the development of expertise among judges and court staff over time? Which cognitive frictions most effectively reduce overreliance without making official systems unusable? When do multi-agent architectures genuinely improve epistemic diversity, and when do they merely reproduce correlated model errors? Which logs and intermediate records are necessary for meaningful legal audit without producing unmanageable volumes of data? How should risk classification change when individually low-risk agents are composed into a system that materially influences adjudication? And what training practices best preserve the capacity of professionals to reclaim delegated tasks when systems fail?

These questions suggest a broader research agenda on the institutional ecology of AI-assisted adjudication. The next generation of judicial AI research should not study models only as tools used by judges. It should study the hybrid systems formed by models, agents, people, procedures, interfaces, incentives, databases, and governance structures.

In that sense, the transition from artificial intelligence to institutional intelligence is not a claim that institutions themselves become intelligent in a human sense. It is a demand that they become more deliberate about how intelligence is organized (human and artificial).

Courts have always distributed cognition. What changes now is the arrival of artificial actors capable of performing an expanding share of the cognitive work that precedes institutional decisions. The constitutional challenge is to ensure that, as this distribution changes, responsibility and legitimacy do not disappear into the architecture.

REFERENCES

- Bednar, Nicholas, David R. Cleveland, Allan Erbsen & Daniel Schwarcz, *Artificial Intelligence and Human Legal Reasoning*, 121 Nw. U. L. Rev. (forthcoming 2027). SSRN No. 6525800, <https://doi.org/10.2139/ssrn.6525800>.
- Bengio, Yoshua et al., *International AI Safety Report 2026*, DSIT 2026/001 (2026), <https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026>.
- Brynjolfsson, Erik, Danielle Li & Lindsey R. Raymond, *Generative AI at Work*, 140 Q.J. Econ. 889–942 (2025), <https://doi.org/10.1093/qje/qjae044>.
- Buçinca, Zana, Maja Barbara 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, no. CSCW1, art. 188, 1–21 (2021), <https://doi.org/10.1145/3449287>.
- Cemri, Mert et al., *Why Do Multi-Agent LLM Systems Fail?* (arXiv preprint, version 3, Oct. 26, 2025), <https://doi.org/10.48550/arXiv.2503.13657>.
- Chan, Alan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger & Tegan Maharaj. *Harms from Increasingly Agentic Algorithmic Systems*, In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 651–66 (2023), <https://doi.org/10.1145/3593013.3594033>.
- Conselho Nacional de Justiça [CNJ], *Comitê Nacional de Inteligência Artificial do Judiciário (CNIAJ)*, <https://www.cnj.jus.br/sistemas/plataforma-sinapses/comite-nacional-de-inteligencia-artificial-do-judiciario-cnij/> (last visited Oct. 21, 2026).
- Conselho Nacional de Justiça [CNJ], *Mapeamento de IA no Poder Judiciário*. 2025, <https://www.cnj.jus.br/wp-content/uploads/2025/12/relatorio-mapeamento-projetos-iniciativas-inteligencia-artificial.pdf>.
- Conselho Nacional de Justiça [CNJ], *O Uso da Inteligência Artificial Generativa no Poder Judiciário Brasileiro: Relatório de Pesquisa*. Brasília: CNJ, 2024, <https://bibliotecadigital.cnj.jus.br/jspui/handle/123456789/976>.

- Conselho Nacional de Justiça [CNJ], Resolução No. 615, de 11 de março de 2025, DJe/CNJ No. 54/2025, 2–17 (Mar. 14, 2025), as amended by Resolução No. 674, de 25 de março de 2026, <https://atos.cnj.jus.br/atos/detalhar/6001>.
- Conselho Nacional de Justiça [CNJ], Resolução No. 674, de 25 de março de 2026, DJe/CNJ No. 71/2026, 2 (Mar. 26, 2026), <https://atos.cnj.jus.br/atos/detalhar/6836>.
- Constituição Federal [C.F.] [Constitution], arts. 5, XXXV, LIV–LV, and 93, IX (Braz.), https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm.
- Debenedetti, Edoardo, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer & Florian Tramèr, AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents (arXiv, version 3, Nov. 24, 2024), <https://doi.org/10.48550/arXiv.2406.13352>.
- Du, Yilun, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum & Igor Mordatch, Improving Factuality and Reasoning in Language Models Through Multiagent Debate, 235 *Proceedings of Machine Learning Research* 11733–63 (2024), <https://proceedings.mlr.press/v235/du24e.html>.
- Faraj, Samer, Stella Pachidi & Karla Sayegh, Working and Organizing in the Age of the Learning Algorithm, 28 *Info. & Org.* 62–70 (2018), <https://doi.org/10.1016/j.infoandorg.2018.02.005>.
- Follesdal, Andreas & Malcolm Langford, Trustworthy AI in Courts: A Legitimacy Analysis, 2 *Cambridge Forum on AI: Law and Governance* e20 (2026), <https://doi.org/10.1017/cfl.2026.10061>.
- Green, Ben, The Flaws of Policies Requiring Human Oversight of Government Algorithms, 45 *Comput. L. & Sec. Rev.* 105681 (2022), <https://doi.org/10.1016/j.clsr.2022.105681>.
- Guo, Taicheng, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest & Xiangliang Zhang, Large Language Model Based Multi-Agents: A Survey of Progress and Challenges, In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 8048–57 (2024), <https://doi.org/10.24963/ijcai.2024/890>.
- Hutchins, Edwin, *Cognition in the Wild* (MIT Press 1995).
- Lee, Hao-Ping, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks & Nicholas Wilson, The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers, in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, art. 1121, 1–22 (2025), <https://doi.org/10.1145/3706598.3713778>.
- Lei No. 13.105, de 16 de março de 2015 [Código de Processo Civil], arts. 9–11 and 489, § 1 (Braz.), https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/113105.htm.

- Li, Guohao et al., CAMEL: Communicative Agents for “Mind” Exploration of Large Language Model Society. In *Advances in Neural Information Processing Systems* 36 (2023), <https://proceedings.neurips.cc/paper/2023/hash/a3621ee907def47c1b952ade25c67698-Abstract-Conference.html>.
- Li, Mahei Manhai, Philipp Reinhard & Jan Marco Leimeister, *Agentic AI: From Autonomous Agents to Enterprise Agentic AI Systems*, 68 *Bus. & Info. Sys. Eng.* 1191–1204 (2026), <https://doi.org/10.1007/s12599-026-01009-w>.
- Liu, Grace, Brian Christian, Tsvetomira Dumbalska, Michiel A. Bakker & Rachit Dubey, *AI Assistance Reduces Persistence and Hurts Independent Performance*. arXiv preprint, version 6, Oct. 13, 2026, <https://doi.org/10.48550/arXiv.2604.04721>.
- Liu, Zhuowen, *Passing the Test You Trained On: Re-evaluating Prompt-Injection Detectors for LLM Agents* (arXiv preprint, Oct. 2, 2026), <https://doi.org/10.48550/arXiv.2610.03448>.
- Magesh, Varun, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning & Daniel E. Ho, *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*, 22 *J. Empirical Legal Stud.* 216–242 (2025), <https://doi.org/10.1111/jels.12413>.
- March, James G. & Johan P. Olsen, *Rediscovering Institutions: The Organizational Basis of Politics* (Free Press 1989).
- Mollick, Ethan, *Co-Intelligence: Living and Working with AI* (Portfolio 2024).
- National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, NIST AI 600-1 (July 2024), <https://doi.org/10.6028/NIST.AI.600-1>.
- National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1, Jan. 2023, <https://doi.org/10.6028/NIST.AI.100-1>.
- Park, Joon Sung et al., *Generative Agents: Interactive Simulacra of Human Behavior*, In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, art. 2 (2023), <https://doi.org/10.1145/3586183.3606763>.
- Qian, Chen et al., *ChatDev: Communicative Agents for Software Development*, In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 15174–86 (2024), <https://doi.org/10.18653/v1/2024.acl-long.810>.
- Raisch, Sebastian & Sebastian Krakowski, *Artificial Intelligence and Management: The Automation–Augmentation Paradox*, 46 *Acad. Mgmt. Rev.* 192–210 (2021), <https://doi.org/10.5465/amr.2018.0072>.
- Re, Richard M. & Alicia Solow-Niederman, *Developing Artificially Intelligent Justice*, 22 *Stan. Tech. L. Rev.* 242 (2019).

- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), 2024 O.J. (L) 2024/1689.
- Salomão, Luis Felipe & Dierle Nunes, O Jurista que Parou de Tentar: IA, Dependência Cognitiva e Como o Design Importa. *ConJur*, May 8, 2026, <https://www.conjur.com.br/2026-mai-08/o-jurista-que-parou-de-tentar-ia-dependencia-cognitiva-e-como-o-design-importa-2/>.
- Santoni de Sio, Filippo & Jeroen van den Hoven, Meaningful Human Control over Autonomous Systems: A Philosophical Account, 5 *Frontiers Robotics & AI* 15 (2018), <https://doi.org/10.3389/frobt.2018.00015>.
- Sapkota, Ranjan, Konstantinos I. Roumeliotis & Manoj Karkee, AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges. 2(3) *SuperIntelligence – Robotics – Safety & Alignment* (2025), <https://doi.org/10.70777/si.v2i3.15161>.
- Shavit, Yonadav et al., Practices for Governing Agentic AI Systems. 2023, <https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf>.
- Shinn, Noah, Federico Cassano, Ashwin Gopinath, Karthik R. Narasimhan & Shunyu Yao, Reflexion: Language Agents with Verbal Reinforcement Learning. In *Advances in Neural Information Processing Systems* 36 (2023), https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.
- Shrestha, Yash Raj, Shiko M. Ben-Menahem & Georg von Krogh, Organizational Decision-Making Structures in the Age of Artificial Intelligence, 61 *Cal. Mgmt. Rev.* 66–83 (2019), <https://doi.org/10.1177/0008125619862257>.
- Simon, Herbert A. *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations*, 4th ed (Free Press 1997).
- Socol de la Osa, David Uriel & Nydia Remolina, Artificial Intelligence at the Bench: Legal and Ethical Challenges of Informing—or Misinforming—Judicial Decision-Making Through Generative AI, 6 *Data & Policy* e59 (2024), <https://doi.org/10.1017/dap.2024.53>.
- Susskind, Richard, *Online Courts and the Future of Justice* (Oxford University Press 2019).
- Tamer, Maurício Antonio, Regulação da Inteligência Artificial no Judiciário Brasileiro: Análise do Objeto e Definições da Resolução CNJ n. 615/2025, 36 *Rev. TRF3*, no. 162, 17–36 (2025), <https://doi.org/10.65674/rev-trf3.v36i162.742>.
- UNESCO, Recommendation on the Ethics of Artificial Intelligence. Nov. 23, 2021, <https://unesdoc.unesco.org/ark:/48223/pf0000381137>.

Vaccaro, Michelle, Abdullah Almaatouq & Thomas Malone, When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis, 8 Nature Hum. Behav. 2293–2303 (2024), <https://doi.org/10.1038/s41562-024-02024-1>.

Wang, Lei, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei & Jirong Wen, A Survey on Large Language Model Based Autonomous Agents, 18 Frontiers Comput. Sci. 186345 (2024), <https://doi.org/10.1007/s11704-024-40231-1>.

João Sergio dos Santos Soares Pereira

Ph.D. in Law, State, and Constitution from the University of Brasília (UnB), with a focus on Procedural Law and Technology, and LL.M. in Constitutional Law from the Brazilian Institute of Education, Development, and Research (IDP). He has undertaken postdoctoral research fellowships at the University of Coimbra, Rio de Janeiro State University (UERJ), Fluminense Federal University (UFF), and the University of Brasília (UnB). He holds several postgraduate specializations and MBA degrees in Procedural Law, Digital Law, Artificial Intelligence, Data Governance, and Neuroscience. He is a professor at FGV Rio and at judicial and public prosecution training institutions, as well as a researcher involved in projects affiliated with UnB, the University of São Paulo (USP), the National Council of Justice (CNJ), and the United Nations Development Programme (UNDP), focusing on the application of artificial intelligence to the justice system. He has contributed to technological innovation projects at the Federal Regional Court of the 1st Region (TRF1), the CNJ, and the Rio de Janeiro State Court of Justice (TJRJ). Since 2002, he has been a career civil servant at the Rio de Janeiro State Court of Justice, where he serves as a judicial advisor and Chief of Staff to an Appellate Judge.

Email: josh.sergio@uol.com.br

ORCID iD: <https://orcid.org/0000-0001-6271-4242>