Transformative Society (2026), 2(2): 1-30.
https://doi.org/10.63336/TransSoc.53

Judicial Regulation of Generative AI "Machine Opinions": Scrutiny Dilemmas, Layered Blocking Mechanisms, and Institutional Safeguards

Yilin Chen[1*]

[1] The University of Edinburgh, Edinburgh, United Kingdom

Corresponding author: Yilin Chen (Cadenchen@outlook.com)

Submitted: 2026-03-30 / Accepted: 2026-04-28 / Published: 2026-04-30

Abstract: Generative artificial intelligence increasingly produces outputs that participate in judicial fact-finding and adjudication — both as litigant-submitted materials and as components of court-internal assistance systems — yet existing evidentiary classifications and procedural review mechanisms address neither the distinctive features of these outputs nor the organizational conditions of their deployment. This article develops “machine opinion” as a functional analytical category covering both forms, distinguishing explicit from implicit machine opinions to support a structurally specific regulatory analysis. The article then constructs a three-dimensional blocking framework in which procedural blocking and managerial blocking function as specialized responses to risks specific to each entry form, while cognitive blocking operates at a regulatory layer distinct from both, addressing the common downstream risk of preemptive directional influence on judicial conviction-forming. As litigation-internal blocking faces inherent limits where upstream technological conditions are not adequately shaped, external institutional support providing minimum upstream conditions is necessary as complement. The analysis is normative rather than empirical, primarily situated in Chinese procedural law with comparative materials drawn upon as institutional reference points; its central concept of cognitive sovereignty designates the normatively protected institutional capacity that the regulatory architecture is designed to support.

Keywords: generative artificial intelligence, machine opinion, cognitive sovereignty, three-dimensional blocking, layered regulation

1. Introduction

Generative artificial intelligence is reshaping judicial systems worldwide. In the European Union, the Artificial Intelligence Act (EU AI Act, 2024) has classified judicial AI applications as "high-risk" systems subject to mandatory transparency, human oversight, and lifecycle documentation obligations. In the United States, the Advisory Committee on Evidence Rules (2025) has proposed a new Rule 707 for machine-generated evidence, signaling a paradigm shift toward reliability-based admissibility standards for AI outputs. These developments reflect a recognition across multiple jurisdictions that generative AI outputs differ in important respects from conventional evidence and require dedicated regulatory responses.

In China, generative AI is being rapidly integrated into judicial practice. From AI-assisted document review and intelligent sentencing recommendations in criminal proceedings to automated case classification, similar-case retrieval, and predictive analytics in civil litigation, both the Supreme People's Court and grassroots courts are deploying generative AI tools at unprecedented scale (Supreme People's Court, 2022). Yet existing Chinese procedural law was designed around presuppositions — human witnesses, deterministic electronic records, identifiable expert authors — that generative AI's probabilistic, generative, and non-factual outputs systematically violate.

This combination of rapid deployment and conceptual mismatch generates a regulatory gap that this article addresses. The existing scholarship on generative AI in judicial settings spans substantial ground, but it can be organized, for the purposes of the present article's intervention, along two lines of substantive inquiry and one body of normative and comparative resources drawn upon for evaluative reference.

The first line of substantive inquiry treats AI as an evidentiary object, examining the doctrinal status, reliability, and explainability of AI outputs at the level of admissibility and weight. Domestic Chinese scholarship along this line has analyzed the legal nature and regulatory pathways for generative AI evidence (Xiong, 2025; Han, 2025; Zhang D., 2022). International work has identified how machine-generated outputs reveal structural gaps in existing evidence law (Roth, 2023) and has mapped the specific phenomenon of legal hallucination in large language models (Dahl et al., 2024).

The second line of substantive inquiry treats AI as adjudicative infrastructure, examining how Smart Court systems and digital-prosecution mechanisms reshape adjudicative workflows, human-machine division of labor, and the organizational conditions within which judges decide. Work along this line has analyzed the auxiliary positioning and procedural boundaries of AI in judicial practice from a human–machine collaboration perspective (Yu P., 2023; Gong, 2023).

Beyond these two substantive lines, the article draws upon a body of normative and comparative resources — the EU AI Act's lifecycle documentation and human-oversight regime; the proposed U.S. Federal Rule of Evidence 707; Gless's comparative analysis of machine evidence in criminal trials (Gless, 2020); and theoretical work on the responsibility gaps raised by autonomous AI systems (Santoni de Sio & Mecacci, 2021). These are not the primary objects of this article's integrative work; they function as evaluative benchmarks and institutional reference points for the regulatory architecture developed below.

Two underlying disagreements organize this field. The first concerns whether generative AI outputs should be assimilated into existing evidentiary categories (such as electronic data or expert opinions) or addressed through a new category. The second concerns whether the principal regulatory response should occur at the procedural level — through reform of evidentiary rules and admissibility standards — or at the organizational level, through the governance architecture of Smart Court systems and judicial-administration design. The present article does not enter the first disagreement directly: by treating "machine opinion" as a functional analytical category rather than as a proposal for a new statutory evidence type, the article avoids taking a position on assimilation-versus-new-category. On the second disagreement, the article argues that neither end can effectively address the problem in isolation: the procedural level and the organizational level must operate as complementary components of a single regulatory architecture, coordinated rather than substitutive.

The specific gap the present article addresses, within this structured field, is that the two substantive lines of inquiry have developed in parallel without being brought into integrated analysis. The dual role of generative AI in adjudication — as an explicit evidentiary object submitted to courts and, simultaneously, as an implicit component of the adjudicative infrastructure within which judges decide — has not been examined under a unified framework capable of tracing how procedural-level and organizational-level risks interact. Existing procedural mechanisms remain insufficiently equipped to address the cognitive and organizational risks that algorithmic outputs may engender (Gong, 2023).

This article advances three interrelated propositions. First, it proposes and develops the concept of "machine opinion" as a functional analytical category under which the distinctive features of generative AI outputs in adjudication can be coherently examined. The category distinguishes between explicit machine opinions (materials submitted by litigants to courts) and implicit machine opinions (embedded within court-internal AI assistance systems). The contribution is not the introduction of a new statutory evidence type, but the organization of a regulatory analysis that existing doctrinal classifications cannot organize coherently. Second, it constructs a three-dimensional blocking framework — procedural blocking, cognitive blocking, and managerial blocking — in which procedural and managerial blocking function as specialized responses to specialized risks originating in distinct entry forms, while cognitive blocking, operating at a regulatory layer distinct from both, addresses the common downstream risk of preemptive directional influence on judicial conviction-forming that both forms generate. Third, it argues that litigation-internal safeguards alone are unlikely to absorb upstream technological risks generated before judicial outputs enter the courtroom or the judicial workflow, and that a layered governance structure integrating external institutional conditions with internal blocking mechanisms is therefore required to protect judges' cognitive sovereignty as a normatively anchored institutional capacity.

The analytical context of this article is Chinese procedural law. Comparative instruments — including the EU AI Act, the proposed U.S. Federal Rule of Evidence 707, the Council of Europe's 2024 Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (Council of Europe, 2024), and related materials — are drawn upon as normative benchmarks and institutional reference points rather than as foundations for a cross-jurisdictional general theory. Where these instruments are referenced in what follows, their role in the argument is specified explicitly — as a normative benchmark, an institutional contrast, or evidence of problem-commonality — rather than as generic indicators of an "international trend."

The remainder of this article proceeds as follows. Section 2 establishes the conceptual foundation by arguing that generative AI outputs cannot be adequately classified under existing evidentiary categories and proposes "machine opinion" as the governing analytical concept. Section 3 examines the structural failures of traditional review mechanisms when confronted with machine opinions, covering both the cross-examination breakdown triggered by explicit machine opinions and the cognitive erosion caused by implicit machine opinions coupled with administrative performance discipline. Section 4 constructs the three-dimensional blocking framework. Section 5 addresses the inherent limits of litigation-internal safeguards and articulates the external institutional conditions necessary to sustain the blocking mechanisms. Section 6 concludes.

2. Conceptual Foundation: Machine Opinion as a Functional Analytical Category

This section establishes the evidentiary character of generative AI outputs and argues that they require a new conceptual category. The analysis proceeds in three steps: first, it demystifies the technical nature of generative AI to reveal its essentially "non-factual" quality; second, it demonstrates the doctrinal mismatch between such outputs and existing evidentiary categories, establishing "machine opinion" as the governing concept; and third, it distinguishes two operational forms — explicit and implicit machine opinions — that structure the regulatory analysis throughout this article.

2.1 Technical Demystification: The Non-Factual Nature of Generative AI Outputs

Traditional evidence law presupposes that electronic data constitutes "objective records," emphasizing a relatively stable correspondence between the storage medium, source, and content, with a core focus on maintaining the integrity and tamper-resistance of original records. However, generative AI powered by large language models disrupts this epistemological premise. The outputs of generative AI are not mechanical reproductions or mirror images of pre-existing facts; rather, they are generative expressions formed through the model's processing of massive corpora, probabilistic associations, and contextual predictions. This generative mechanism determines that the output content is "probabilistic" rather than "deterministic" (Xiong, 2025), and is essentially "non-factual" — it does not correspond to any objectively existing original fact but is instead content generated according to data and algorithmic structures (Han, 2025).

This generative mechanism inevitably gives rise to two inherent technical deficiencies. The first is "hallucination" risk: model outputs may contain fabricated facts, misattributed citations, or logically spliced content. Even when the output formally appears complete, coherent, and highly realistic, such appearance cannot serve as a basis for presuming the truthfulness of its content (Dahl et al., 2024). The second is the inexplicability produced by the "black box." The decision pathways of deep learning models are highly complex and opaque; there is no intuitive logical connection between input and output (Wei, 2024), making it difficult to reconstruct the reasoning in terms comprehensible to human understanding. The questions raised thereby no longer concern merely whether "this content genuinely exists on a particular storage medium," but extend to the generative causes of the output, the reliability of the inferential pathway, and whether any verifiable correspondence exists between the conclusions and the facts of the case. This exceeds the scope of traditional electronic data review.

2.2 Evidentiary Breakthrough: Establishing the Concept of "Machine Opinion"

Given the foregoing characteristics of generative AI outputs, classifying them simply as "electronic data" or "expert opinions" presents doctrinal mismatches. This section accordingly develops "machine opinion" as a functional analytical category rather than as a proposal for a new statutory evidence type, organizing the subsequent procedural analysis around those distinctive features that resist absorption into either existing classification.

Forcibly classifying them as "electronic data" entirely ignores their generative subjectivity, judgmentality, and predictiveness, treating evaluative outputs from algorithmic models as if they were neutral physical records. Treating them as "expert opinions" likewise produces a mismatch: machine opinions lack a juridically responsible human subject capable of bearing perjury liability, the methodological transparency essential to forensic expert analysis, and the procedural mechanisms (oath, cross-examination, demeanor observation) traditionally used to test reliability.

To resolve this doctrinal misalignment, this article adopts "machine opinion" in a broadened sense as a functional analytical category — an organizing term under which the distinctive features of generative AI outputs can be examined without forcing them into existing statutory classifications. As Gless (2020) has shown, AI-driven systems that observe and evaluate human behavior can produce outputs that go beyond mere measurement, conveying judgmental and evaluative content of their own — what Gless analyzes as "machine evidence," building on Roth's discussion of "machine testimony." Such outputs exhibit subjective evaluative characteristics that approximate, in functional terms, the judgmental dimension of human opinion. This characterization captures the dual nature of generative AI outputs: they possess the technical dependence of electronic data while simultaneously exhibiting the subjective inferential and predictive qualities akin to expert testimony or character evidence. What matters for the analysis that follows is not whether this category acquires independent standing in positive evidence law, but whether — as a functional analytical category — it enables a coherent institutional response to risks that the existing classifications cannot adequately address.

The concept of "machine opinion" as deployed in this article draws on two principal theoretical resources, to which a third — a Chinese-language doctrinal treatment — contributes the determinate terminological positioning the concept carries in the present context.

The first resource is Roth's analysis of machine testimony (Roth, 2017), which identifies the need for a standalone evidentiary category for machine outputs that is independent of the human-witness presuppositions built into traditional hearsay doctrine and cross-examination procedures. Machine outputs, Roth argues, function in the judicial process as a source of information distinct from party statements and documentary records, but the doctrinal instruments designed to test the reliability of human testimony — including cross-examination, oath, and demeanor observation — cannot be applied to machine outputs without substantial modification. A standalone analytical category is therefore required. The present article adopts this core insight: machine outputs resist coherent absorption into evidentiary categories premised on human subjects, and an independent analytical frame is needed.

The second resource is Gless's comparative analysis of machine evidence in criminal trials (Gless, 2020), which develops the Rothian insight in a comparative-law register. Gless adopts "machine evidence" as her organizing concept while explicitly engaging Roth's notion of "machine testimony," and her broader contribution lies in articulating a normative framework concerned with how technology reshapes the distribution of adjudicative authority through evidentiary channels. Although developed in the context of criminal machine evidence, that framework provides an important basis for the present article's extension to court-internal AI assistance systems. Gless does not put forward "machine opinion" as a doctrinal category in her own work; rather, her comparative analysis offers conceptual and normative material that the present article reorganizes under the heading of machine opinion in the Chinese-law context.

A structural question arises when transposing this Rothian and Glessian analytical frame into the Chinese-law context. In common-law systems, "testimony" as a doctrinal category encompasses both lay witness testimony and expert testimony, so that Rothian machine testimony spans both functional locations. The Chinese law of evidence, however, draws a categorical distinction between witness testimony (zhèngrén zhèngyán, 证人证言) — the factual account given by a witness based on direct perception — and expert or forensic opinion (jiàndìng yìjiàn, 鉴定意见) — the judgmental conclusion rendered by a qualified person on a specialized question (Chen R., 2021). Witnesses are not permitted to render judgmental conclusions on specialized questions; forensic opinions cannot be produced by persons other than qualified experts. Machine outputs, which produce judgmental conclusions through specialized methods rather than reporting direct perceptions, are functionally aligned with the expert-opinion location rather than with the witness-testimony location. The cross-legal-system repositioning is therefore not a matter of translation but of identifying the functionally correct doctrinal address: Rothian testimony maps, in the Chinese-law context, to opinion, not to testimony.

The third resource makes this repositioning determinate at the level of Chinese doctrinal discussion. Yu P. (2024), in a dedicated doctrinal treatment of AI evidence in criminal procedure, defines AI evidence at the outset as a "machine opinion" (jīqì yìjiàn, 机器意见) that supports fact-finding, and characterizes it as a new form of opinion evidence. Yu distinguishes machine opinion from ordinary-witness opinion, from expert or forensic opinion, and from electronic evidence, emphasizing throughout that the core feature of AI evidence lies in its being an inference or judgment derived from machine-mediated analysis rather than a direct account of perceived facts. The reliability of such outputs, Yu argues, depends on the internal standardization of the underlying system rather than on human perceptual and executive capacities, so that the procedural mechanisms traditionally used to secure the reliability of witness testimony — oath, demeanor observation, and cross-examination — cannot be directly applied. The present article adopts this terminological positioning: "machine opinion" is the doctrinally appropriate designation in the Chinese-law context, and the term is used here with that doctrinal orientation in mind.

Building on these three resources, this article undertakes three structurally distinct extensions that together constitute its analytical contribution. The first is a normative-channel extension: Gless's normative framework examines how technology reshapes the distribution of adjudicative authority through evidentiary channels, limited in her treatment to criminal-evidence cases. The present article retains that normative concern but extends the channel structure from a single channel (evidence) to a coordinated set of channels that includes the organizational embedding of AI assistance systems within courts. Technology-mediated reshaping of adjudicative authority, on this extended view, can occur both through materials entering the courtroom and through infrastructure operating on the judicial workflow, and an adequate regulatory analysis must address both. The second is an internal structuring: the explicit/implicit differentiation developed in §§2.3–2.4 supplies the category with an internal analytical architecture that enables differentiated institutional responses. The third is an elevation to an overarching analytical category: under this architecture, the category carries the defining feature and two independent analytical functions specified in the three-clarifications block below, and supports the differentiated regulatory anchoring formally specified at the close of §2.4. The article's contribution, so situated, is a structural repositioning and extension of existing conceptual resources — not a claim to have originated the term or the underlying insight — oriented toward a specific procedural-law regulatory task that the existing resources alone do not support.

Building on the foregoing genealogy and the three extensions, three clarifications on the analytical role of "machine opinion" are in order before turning to its typological development in §§2.3–2.4.

First, the category operates at a functional rather than doctrinal level. Its purpose is not to propose an additional statutory evidence type alongside electronic data and expert opinions, but to gather, under a single analytical frame, those judicial outputs of generative AI whose distinctive features — probabilistic generation, non-factual character, and the absence of a human subject capable of bearing perjury liability — produce institutional consequences that cut across existing classifications.

Second, the category carries one defining feature and two independent analytical functions. The defining feature is integrated comparability: by placing externally submitted AI-generated materials and internally embedded AI assistance systems within a single analytical frame, the category allows their respective effects on adjudicative formation to be assessed along common dimensions — reliability, cognitive influence, and transparency — that segmented frameworks do not accommodate. Beyond this defining feature, the category performs two independent analytical functions. The first is boundary demarcation: the category clarifies regulatory scope by delimiting which forms of AI involvement in adjudication fall within the article's analytical purview, distinguishing them from general judicial digitalization — such as case management, automated scheduling, and AI-assisted legal research used by litigants outside court — whose outputs do not directly enter fact-finding or conviction formation. The second is differentiated regulatory anchoring: under this category, the explicit/implicit differentiation triggers structurally distinguishable institutional consequences, allowing the regulatory architecture to allocate specialized blocking mechanisms (procedural blocking to explicit machine opinions, managerial blocking to implicit machine opinions) while retaining a shared blocking mechanism (cognitive blocking) for the common downstream risk. This second function is developed at the close of §2.4, after the explicit/implicit differentiation has been substantively introduced.

Third, the category's reach is limited by the regulatory concern that motivates it. "Machine opinion" in this article extends to those AI outputs that substantively participate in fact-finding or in the formation of adjudicative conviction — whether by being submitted to the court as litigation materials or by being embedded within court-internal assistance systems that operate on adjudicative inputs. Pure infrastructural functions that do not operate on adjudicative reasoning itself — including basic docketing, scheduling, and undifferentiated document retrieval — fall outside its scope.

2.3 Explicit Machine Opinions: Practical Forms and Review Priorities

Having established "machine opinion" as a functional analytical category, it is now necessary to develop its operational forms within the judicial domain — both to prevent conceptual conflation from disrupting regulatory coherence and to lay the structural basis on which differentiated regulatory anchoring will operate in the analysis that follows. In practice, machine opinions manifest in two forms: explicit machine opinions that enter courts from outside as litigation materials, and implicit machine opinions that are embedded within courts as components of adjudicative assistance systems.

Explicit machine opinions, as externally generated materials, primarily appear in four typical forms in current judicial practice.

The first is commercial valuation materials: AI-generated property valuation reports, commercial loss assessments, and intellectual-property infringement damage calculations submitted by parties as evidence. Such materials apply parametric inference and predictive modeling rather than the deterministic application of tested professional standards (Federal Rules of Evidence, Rule 702, as amended 2023; Advisory Committee on Evidence Rules, 2025); their reliability accordingly depends on whether the methodological assumptions, training data, and parameter choices of the underlying model have been transparently disclosed and independently verifiable.

The second is AI-generated content nested within forensic expert reports. As DeepSeek and other generative-AI tools become incorporated into forensic-expert workflows, expert opinions increasingly contain AI-generated analytical components — for instance, document-authenticity verification using AI image-analysis subroutines, or forensic-accounting reports incorporating AI predictive components. Such content occupies a doctrinally ambiguous position: nominally part of the human expert's opinion, yet substantively produced by AI. Han (2025) has argued that the AI-generated portion should be subject to its own reliability review independent of the broader expert opinion.

The third is AI-enhanced original evidence: image-restoration outputs, audio-enhancement outputs, video-stabilization outputs, and similar AI-mediated processing of original evidence. Such enhancement may simultaneously increase the apparent clarity of the evidence and introduce content alterations not present in the original (Norman & Farid, 2024). The procedural review priority lies in distinguishing what is enhanced (form) from what is generated (content).

The fourth is unconfirmed deepfakes: AI-generated synthetic media — typically video, audio, or image content depicting persons or events that did not occur — submitted as evidence either by the producing party (claiming the synthetic media to be authentic) or by an opposing party (offering it as evidence of fabrication or manipulation). NIST (2024a) provides technical-detection frameworks; the procedural review priority is the integration of such technical detection with the courtroom evidentiary process.

Because these four types of explicit machine opinions involve different modes of technological intervention, their litigation review priorities should not be treated uniformly: the first type emphasizes methodological reliability; the second emphasizes the demarcation between human expert judgment and AI-generated content; the third emphasizes the boundary between enhancement and generation; the fourth emphasizes the integration of technical detection with the evidentiary process.

One boundary clarification is warranted to complete the typology. AI-assisted legal information-retrieval outputs — including case-law recommendation tools, statute-matching utilities, and similar-case search systems used by litigants outside court — are not treated as explicit machine opinions within the present framework. The reason is functional: such tools perform information-location work on existing legal materials rather than generating new factual or analytical content, and their outputs are best understood as the outputs of legal-research instruments, whose procedural status may be handled, in principle, under the rules applicable to existing legal-research tools. The four types identified above are accordingly those forms in which AI intervention generates content that directly enters fact-finding or substantive legal analysis, not those forms in which AI facilitates access to existing legal sources. Where the boundary between the two becomes contested

in practice — for example, where a retrieval tool presents synthesized summaries rather than raw retrieval results, thereby introducing generative content into what formally remains a retrieval output — the tool's procedural treatment should follow the functional criterion just stated rather than its nominal designation.

2.4 Implicit Machine Opinions: Deep Embedding in Judicial Workflows

Unlike explicit machine opinions that enter courts from outside as litigation materials, implicit machine opinions are AI components embedded within courts as parts of internal assistance systems — functioning not as evidentiary objects but as infrastructural elements within the adjudicative workflow itself.

A clarification is warranted here. Not every form of digitalization within courts gives rise to implicit machine opinion concerns. Basic case management, automated scheduling, document filing, and similar infrastructural digitalization — even when AI-supported — does not directly operate on adjudicative reasoning and is not the focus of this article. The implicit machine opinions of regulatory concern are those AI assistance systems whose outputs do operate on adjudicative reasoning: similar-case retrieval and recommendation systems whose outputs influence conviction or sentencing; risk-assessment systems whose outputs shape evidentiary decisions; deviation-alert systems whose outputs shape judicial reasoning under accountability pressure; and integrated decision-support platforms whose outputs participate substantively in conviction-forming.

In civil proceedings, this embedding takes the form of similar-case retrieval and recommendation systems integrated with case-management workflows. Through Smart Court infrastructure deployed across Chinese courts, AI-driven systems intervene in civil judicial handling through mechanisms such as similar-case retrieval, predictive case-outcome analyses, deviation alerts, and case-allocation and process-management tools, thereby structuring the informational and procedural environment within which judges work (Gong, 2023). The institutional concern is that, when such recommendations reach judges before independent case analysis is undertaken, the recommendations may anchor subsequent reasoning.

In criminal proceedings, the embedding goes further. AI risk-assessment, sentence-prediction, deviation-alert, and integrated decision-support systems are deployed within courts as components of digitalized adjudicative workflows (Fan, 2024). Shen L. (2025) and Ye (2026) have analyzed how these embeddings reshape the structural conditions of criminal proof and conviction-forming respectively.

Through deep learning on massive corpora, generative AI outputs often present highly structured content, rigorous logical closure, and extremely standardized legal language, thereby acquiring a highly deceptive "appearance of superiority" (Liu Y., 2024). When confronting this appearance of superiority, judges — constrained by a structural review-capacity gap — can neither access the model's internal operational information nor penetrate the black box to investigate the algorithmic weights and logical causation behind the results (Liu Y., 2024).

Zhang B. (2024) warns that when judges over-rely on the efficiency, speed, and apparent accuracy of generative AI outputs, adjudicative decisions risk being reduced, in substance, to algorithmic outcomes. This concern, situated within the broader institutional conditions identified by Fan (2024) — in which performance assessment and adjudicative supervision constitute standardized pressure on judges' trial conduct — grounds the present article's analytical interest in whether and how such pressures may, in turn, shape judges'

engagement with AI outputs. Whether this further step — from pressure on trial conduct to pressure on AI-use patterns — obtains as an empirical regularity in Chinese courts is a question this article treats as an open institutional-risk hypothesis rather than as an established behavioral finding.

The foregoing typological development makes it possible to specify the structural form of the differentiated regulatory anchoring introduced in §2.2. Explicit and implicit machine opinions exhibit partially differentiated and partially shared risk profiles. Explicit forms concentrate risks at the point of admission and cross-examination, where the question is whether externally submitted AI-generated materials can be subjected to meaningful reliability review and adversarial testing before entering the factual record. Implicit forms concentrate risks at the organizational layer, where the question is whether court-internal assistance systems, once coupled with performance evaluation, transform from reference tools into instruments of organizational discipline. These two risk profiles call for specialized regulatory responses operating at different institutional layers: a procedural response addressed to admission and cross-examination for explicit forms, and a managerial response addressed to internal organizational coupling for implicit forms.

At the same time, the two forms converge at a common downstream point: the formation of judicial conviction, where the algorithmic "appearance of superiority" can exert preemptive directional influence regardless of whether the influence enters through externally submitted evidence or through internally embedded assistance. The cognitive layer at which such preemptive directional influence operates is unitary, not partitioned by entry form; the regulatory response addressing that layer must therefore span both forms rather than being assimilated into either specialized response. Cognitive blocking, accordingly, occupies a regulatory layer distinct from both the admissibility layer addressed by procedural blocking and the organizational layer addressed by managerial blocking, and — within that distinct layer — necessarily covers both explicit and implicit forms. The specific operational mechanisms through which cognitive blocking engages explicit versus implicit forms may differ in their concrete modalities, but these operational differences occur within a single regulatory layer directed at a unitary normative object.

The three-dimensional blocking framework developed in §4 is structured accordingly. Procedural blocking and managerial blocking function as specialized responses to specialized risks originating in distinct entry forms; cognitive blocking functions as a shared response to the common downstream risk of preemptive directional influence on judicial conviction-forming. This architecture is neither a uniform response to machine opinion in general, nor three parallel responses proceeding in isolation, but a differentiated allocation calibrated to where and how generative AI intervenes in adjudicative formation. The next section turns to the two distinct types of structural threats that machine opinions, understood in these two forms, pose to judges' institutional capacity for independent conviction-forming.

3. Scrutiny Dilemmas: Two Types of Structural Threats to Judges' Cognitive Sovereignty

Having established that machine opinions cannot be adequately absorbed by traditional evidentiary classification systems, the next question is whether existing procedural mechanisms can effectively handle this novel object of review. This section demonstrates that traditional review mechanisms — whether adversarial or inquisitorial — encounter structural failure when confronting machine opinions.

Before proceeding to the institutional analysis, a core analytical category employed throughout this article requires definition, and its relation to two principal adjacent concepts requires clarification. These two concepts frame the conceptual space within which cognitive sovereignty is introduced: one drawn from the empirical literature on human interaction with automated decision systems, the other drawn from the long-established legal doctrine of judicial independence.

The first is automation bias: the empirically documented tendency of human decision-makers to over-rely on automated outputs, to neglect disconfirming information when automation is present, and to defer to algorithmic recommendations even when these are demonstrably less accurate than unaided judgment (Goddard, Roudsari, & Wyatt, 2012). Automation bias is a descriptive-psychological phenomenon: it characterizes what decision-makers in fact do, and the supporting findings rest on empirical research conducted principally in aviation and clinical-medical settings rather than on judicial decision-makers. While judges and the professional decision-makers studied in this literature differ materially in institutional role, the expert-decision-maker character of both suggests that the basic mechanism carries structural reference value across these domains.

The second is institutional judicial independence: the organization-level absence of overt external interference from administrative organs, appellate courts, or other identifiable sources of influence over adjudicative decision-making. Institutional independence is an institutional-factual concept: it characterizes whether, and from whom, external interference is or is not occurring at the organizational level, and it is historically addressed to episodic, identifiable forms of interference.

These two concepts are analytically valuable in their respective domains, and cognitive sovereignty is not proposed here as a replacement for either of them in those domains. They share, however, a common feature relevant to the present argument: each operates within a different analytical register than the one needed for procedural-law regulation: automation bias is descriptive in character, while institutional independence addresses a different class of interference. Neither, on its own or in combination, performs the specific work required of a regulatory concept in procedural law — the work of identifying a normatively protected institutional capacity on which specific regulatory obligations can rest against covert, continuous, and structurally embedded influence.

The analytical work that cognitive sovereignty performs in this article is precisely this: to serve as a normative anchor that translates the risks descriptively identified by the foregoing concepts into a procedurally protected institutional capacity on which specific regulatory duties can rest. Without such a normative anchor, observations of automation bias in judicial settings remain observations about how judicial decision-making in fact proceeds; with this anchor, they become identifiable threats to a normatively protected institutional capacity, against which procedural law can ground specific regulatory obligations. By 'institutional capacity' the article means the structural conditions enabling judges to form adjudicative conviction independently, not the empirical fact of any particular judge's doing so. The protection at issue is therefore directed at the preservation of those conditions, rather than at the frequency with which independent formation actually occurs. The analytical gain, accordingly, is not the integration of adjacent descriptive concepts — each of which remains usable in its own domain — but the establishment of a regulatory-law platform on which those descriptive findings can acquire normative traction.

The choice of 'sovereignty' as the operative term — rather than weaker adjacent terms such as 'decisional autonomy' (in AI ethics contexts), 'integrity', or 'independence' applied to the cognitive dimension — reflects a specific normative claim. Decisional autonomy in particular characterizes the preservation of human agency under algorithmic influence, but it does so at a register the present analysis is designed to exceed. What the concept aims to protect is not merely a descriptive autonomy — the contingent fact that judges' cognitive processes are, to a greater or lesser extent, self-directed — but a reserved zone of final judgment that remains available to the judge against covert, continuous, and structurally embedded institutional pressures. Covertness, continuity, and structural embedding are the features that distinguish the influence at issue from episodic external interference of the kind addressed by traditional judicial-independence doctrine. Weaker adjacent terms do not carry, with equal force, the normative claim that this reserved zone must remain institutionally secured rather than merely descriptively preserved. The article retains "sovereignty" because this normative load-bearing function has not been replaceable by vocabulary of lesser strength, while being disciplined in its deployment: the term is used where the reserved-zone claim is analytically load-bearing, and not as a general-purpose label.

On this basis, cognitive sovereignty can be defined along two components. The normative-functional component, stated at operative strength: cognitive sovereignty is judges' institutional capacity to autonomously form adjudicative conviction and to bear corresponding reasoning obligations through direct examination and independent evaluation of evidence, with the normative function of preserving a reserved zone of final judgment against covert, continuous, and structurally embedded institutional pressures. This concept differs from the principle of free evaluation of evidence (freie Beweiswürdigung), which concerns the formal authority of judges to evaluate evidence; cognitive sovereignty addresses the preconditions under which that authority can operate substantively — namely, whether judges can independently access evidentiary information, autonomously assess its probative value, and form conviction without preemptive directional influence from external systemic factors (Chen R., 2021). It also differs from the principle of judicial independence, which primarily addresses overt interference from identifiable sources such as administrative organs or appellate courts.

The risk-type component of the definition is stated as a typology of institutional risks against which the foregoing capacity is to be protected, rather than as a description of already-occurring empirical phenomena. The risks relevant to the present analysis are of three types: (i) the risk that the algorithmic appearance of superiority — the structured, logically closed, and linguistically standardized form in which generative AI outputs present — may exert preemptive directional influence on judicial conviction-forming; (ii) the risk that technological information asymmetry between algorithmic systems and non-specialist judicial users may undermine judges' capacity to subject algorithmic outputs to meaningful critical evaluation; and (iii) the risk that the institutionalized embedding of AI assistance systems within judicial workflows, particularly when coupled with performance-evaluation mechanisms, may create organizational conditions under which the reserved zone of final judgment is structurally compressed. The question whether, and to what extent, each of these risks is actually occurring in any particular judicial system is an empirical question for further investigation; what the concept of cognitive sovereignty provides is the normative framework under which such empirically investigated risks can be assessed as threats to a protected institutional capacity (Agudo et al., 2024).

The subsequent analysis demonstrates that this institutional capacity faces two distinct families of structural threat — one originating at the admission and cross-examination of explicit machine opinions (§3.1), the other originating at the internal-organizational coupling of implicit machine opinions with performance discipline (§3.2) — both of which converge at the cognitive layer the present concept identifies as normatively protected.

3.1 The Cross-Examination Breakdown of Explicit Machine Opinions: Structural Failure of Traditional Review Mechanisms

Whether in the adversarial system centered on cross-examination in common law jurisdictions or in the inquisitorial and hybrid structures of civil law jurisdictions and China that emphasize judicial gatekeeping, traditional review mechanisms encounter structural failure when confronting explicit machine opinions. This failure stems not from the inadequacy of particular rules, but from the fundamental transformation of the object of review combined with systemic capacity asymmetries.

U.S. judicial practice has progressively raised reliability standards for AI evidence, attempting to address such risks through case law and rule revisions (Federal Rules of Evidence, Rule 702, as amended 2023). Recognizing this crisis, the Advisory Committee on Evidence Rules has proposed a new Rule 707 for machine-generated evidence, attempting a paradigm shift toward reliability-based admissibility, with the proposal continuing to develop through subsequent committee discussions (Advisory Committee on Evidence Rules, 2025; Advisory Committee on Evidence Rules, 2026).

Yet rule-level amendments cannot conceal the deeper failure of the cross-examination mechanism. When the object of cross-examination shifts from natural persons to generative AI, the cross-examination mechanism encounters the dual obstacles of "absent subject" and "inexplicability." Algorithms possess no consciousness and are not subject to perjury sanctions; the honesty-assurance mechanism in the adversarial system that operates through deterrence thus becomes ineffective (Santoni de Sio & Mecacci, 2021). Simultaneously, because the algorithmic black box is difficult to penetrate, existing explanation methods can in many cases only present the model's external behavioral characteristics or provide post-hoc approximate accounts, without fully revealing why the model selected particular features and assigned them critical weights (Rudin, 2019). Even the introduction of "Explainable AI" (XAI) often remains difficult for non-specialists to interpret meaningfully and may produce a misleading "placebo effect" (Okolo et al., 2024). More importantly, model interpretability should not be conflated with causal or substantive validity. As observed in the broader social science context, sophisticated analytical tools can create a misleading impression of rigor that masks, rather than reveals, the actual mechanisms at work (Li et al., 2025). A deep structural review-capacity gap thus forms between complex AI algorithms and non-specialist users such as judges, jurors, and defense attorneys.

Inquisitorial and hybrid systems face the same structural failure but at a different procedural locus. In China's evidentiary architecture, the threshold question for explicit machine opinions is whether such outputs satisfy the authenticity, relevance, and legality requirements that govern admissibility under the Criminal Procedure Law and the Civil Procedure Law. The doctrinal response within the existing electronic-data framework focuses on hash verification, chain of custody, and source authentication. However, in the face of deepfake synthesis and generative large model technologies, "formal source authenticity" can no longer be equated with

"content truthfulness" (NIST, 2024b). Generative AI outputs may pass formal source authentication while remaining substantively unreliable: a hash-verified video may be a deepfake; a chain-of-custody-tracked audio recording may be AI-synthesized.

The deeper inadequacy lies in the static character of existing review mechanisms. Existing electronic-evidence rules presuppose that authenticity is a property determinable at a single point in time. Generative AI, however, exhibits dynamic credibility: model versions, parameter updates, training-data revisions, and operating-condition changes can all affect output reliability after the original authentication. Given that the reliability of AI models may fluctuate with changes in version, parameters, training data, and operating conditions, prior review conclusions should not be treated as exhausting the compliance status of the system, particularly where the regulatory framework imposes continuing requirements with respect to technical documentation, logging, human oversight, and the maintenance of accuracy, robustness, and cybersecurity (EU AI Act, 2024, arts. 11, 12, 14, 15). The EU AI Act provides the first comprehensive legal-institutional instantiation of this dynamic-credibility concept, requiring high-risk AI systems to maintain sufficient transparency throughout their lifecycle, encompassing automated activity logs, bias-controlled training datasets, technical documentation, and continuous human oversight and risk-mitigation measures (EU AI Act, 2024). The EU AI Act's regulatory adoption of dynamic-credibility review highlights the deep inadequacy of traditional post-hoc courtroom authentication rules as a matter of institutional design.

Deficiencies in training data also exhibit a tree-like propagation: contaminated origins infect downstream products. If training data includes unlawfully obtained private data or embeds severe social discrimination, the judicial analyses or predictions generated therefrom should, as a matter of legal doctrine, be treated as contaminated "fruits of the algorithmic poison tree" (Chen R., 2018)[1]. The technical mechanism is that, due to deep-learning models' architectural memorization, removing illegal training data does not eliminate its influence on the model — even after costly machine-unlearning techniques (Bourtoule et al., 2021). In litigation, this creates a reality where "fruits cannot be removed but the source cannot be cured": the defense cannot prove that a particular output originated from specific unlawful data, and judges likewise face difficulty conducting effective review of data contamination within the model.

On the other hand, a "generative unreliability" inspection dilemma persists. Large language models in legal contexts remain susceptible to user-provided erroneous premises, producing factual hallucinations and even entirely fabricated legal citations (Dahl et al., 2024). Minimal input perturbations can induce AI to generate entirely erroneous conclusions — so-called adversarial examples — rendering the reliability foundations of such evidence extremely fragile (Goodfellow et al., 2015). Simultaneously, due to the "black box" nature of AI systems, achieving a high degree of transparency and inspectability typically requires substantial technical costs and specialized knowledge that exceeds the technical review capacity ordinarily possessed by judges (Hauer et al., 2023).

The structural inadequacy thus identified is not specific to any particular legal tradition. Rather, the review of explicit machine opinions is no longer merely a question of detecting truth or falsehood, but increasingly depends on whether external parties can access the critical technical information necessary for model operation and verification and translate relevant technical doubts into actionable procedural objections (EU AI Act, 2024). Although Liu (2024) does not expressly theorize a systematic three-dimensional gap among model

visibility, interpretive resources, and procedural translation capacity, her discussion of algorithmic opacity, incomplete explainability, and the procedural risks of automated adjudication offers partial support for the analytical reconstruction proposed here: when these three dimensions are systematically asymmetric between prosecution and defense, courts and technology providers, traditional evidence rules, even if formally extant, become difficult to substantively activate.

In sum, following the entry of generative AI into litigation-proof procedures, traditional review mechanisms centered on cross-examination or formal authentication have undergone structural failure in both adversarial and inquisitorial systems. The root cause lies in a structural review-capacity gap: the accessibility and comprehensibility of critical technical information are simultaneously imbalanced, materially undermining the conditions of symmetric review that traditional procedures presuppose.

3.2 The Cognitive Erosion of Implicit Machine Opinions: Administrative Discipline and the Compression of Judicial Independence

Methodological note. The analysis that follows proceeds as a normative inference from institutional conditions identifiable in China's digital-judicial architecture — in particular, the coupling of internal AI assistance systems with performance-evaluation mechanisms, the quantification of adjudicative workflow, and the accountability structures that surround deviations from system recommendations. It is not based on empirical fieldwork examining how individual judges actually use AI assistance systems, nor does it purport to establish behavioral regularities. Where the analysis identifies structural pressures that may shape judicial conduct, those identifications should be read as institutional-risk hypotheses warranting further empirical investigation rather than as demonstrated behavioral findings. This methodological orientation informs the analysis throughout the section: where institutional inference can carry a claim, the claim is stated at the strength such inference supports; where a claim would otherwise require empirical support about individual judicial psychology or behavior, the claim is reframed as a structural pressure on the institutional environment rather than as a finding about what judges actually do. Comparable empirical research on algorithmic-decision-making interactions in adjacent professional domains — for instance, on how human decision-makers engage with risk-assessment algorithms in adjacent decision-making contexts — is also limited in scope and not directly transferable to the Chinese judicial AI context (Green & Chen, 2019). With this methodological orientation in place, the section's analysis proceeds as follows.

The foregoing analysis revealed primarily the risks triggered when explicit machine opinions enter litigation procedures at the levels of admissibility, cross-examination, and reliability review. However, in China's digital justice context, these risks extend beyond the procedural review failure of external explicit materials, reaching into the compression of judges' independent judgment space through the coupling of internal assistance systems with performance governance mechanisms.

In practice, even when nominally positioned as "auxiliary references," Smart Court systems' similar-case retrieval, sentence prediction, and case-handling recommendations are deeply embedded in the digital adjudicative process. File-reading duration, document-revision histories, and the degree of deviation from system recommendations may all enter the internal management field of vision, becoming reference factors for performance evaluation and workflow control (Fan, 2024). While digital justice systems transform judges'

working methods, they may simultaneously compress their space for independent judgment formation in practice.

One of the institutional conditions plausibly generating tendencies toward "defensive adjudication" among judges lies not in technological compromise with external AI evidence alone, but in the institutional bundling between implicit machine opinions within courts and performance management. When algorithmic recommendations become coupled with quantified assessments, deviation alerts, and accountability evaluations, internal assistance systems may transform from reference tools into de facto management instruments (Gong, 2023).

The institutional risk is that, when algorithmic recommendations become bundled with quantified performance evaluation, deviation alerts, and accountability structures, persisting in deviating from the algorithmic recommendation triggers system alerts, requires the judge to expend considerable communication costs justifying the deviation to supervisory authorities, and may invite real-time supervisory intervention or unfavorable assessment scores (Fan, 2024). Confronted with additional explanation obligations, assessment deductions, or even disciplinary accountability risks, the institutional conditions plausibly create incentives toward compliance with system recommendations, fostering defensive adjudicative dispositions oriented toward minimizing organizational risk.

The deeper crisis extends beyond the algorithmic biases and model limitations of technological tools themselves; it lies in the failure of current digitalized judicial management mechanisms to provide sufficient institutional incentives for judges to identify, question, and correct technological biases. When wrongful-conviction accountability, deviation-degree alerts, and efficiency assessments are highly quantified and institutionally bundled with judges' professional evaluations, the technical standards embedded in machine opinions may acquire institutional weight disproportionate to their demonstrated reliability (Cheng, 2021).

The institutional consequence of this technological discipline operates not through algorithmic replacement of judges' adjudicative authority, but through sustained organizational pressures that structurally reduce the institutional space within which judges maintain prudent distance from system outputs, raise reasonable doubts, and make substantive departures (Shen L., 2025). Yu P. (2023) has identified, within the AI-judicial context, how wrongful-conviction accountability pressures may operate through judges' risk-aversion dispositions to influence behavior in algorithmically assisted adjudication, increasing reliance on system outputs as a means of strengthening the defensibility of adjudicative reasoning and thereby raising the risk of preemptive judgment-formation. In line with this analysis, the institutional incentives generated by such accountability pressures may, in turn, generate structural pressures on judges' space for free evaluation of evidence, and judicial adjudication faces the institutional risk of shifting from a judicial logic that seeks substantive justice in individual cases toward a managerial logic that caters to system performance indicators.

This institutional risk derives from the deep coupling between China's digital judicial governance system and judicial responsibility mechanisms. Concrete instantiations of this dynamic can be observed in deployed judicial AI systems in China: Shanghai's "206 System" provides sentencing-prediction references, while Jiangsu's "Smart Adjudication Suzhou Model" performs similar-case matching, autonomously generates adjudicative recommendations, and triggers alerts on inconsistent outcomes across analogous cases (Zhang S.,

2023). Although nominally framed as auxiliary references, such mechanisms may, at the organizational level, gradually translate into external pressures bearing on judges' independent conviction-formation.

The intervention of generative AI in the judicial domain has simultaneously triggered two distinct types of risk. Explicit machine opinions entering litigation procedures cause traditional authenticity, reliability, and static review mechanisms to malfunction. In China's digital justice context, implicit machine opinions form institutional bundles with administrative discipline, performance assessment, and bureaucratic management, further compressing judges' space for independent judgment. With respect to cognitive sovereignty as defined in §3.0, the former affects the first two risks in the §3.0 typology — the preemptive directional influence of algorithmic appearance of superiority and the technological information asymmetry between algorithmic systems and non-specialist judicial users — by undermining the institutional conditions for substantive review and independent evaluation of machine opinions. The latter corresponds to the third risk — institutionalized embedding coupled with performance-evaluation mechanisms — by compressing the institutional space for independent judgment within the organizational environment. Although their mechanisms differ, both point toward the systemic erosion of judges' institutional capacity to autonomously form adjudicative conviction. These failures do not result from isolated technical defects but from structural mismatches between the nature of the objects and the design assumptions of existing institutions. The response pathway must therefore move beyond piecemeal repair toward layered, differentiated institutional reconstruction.

4. Three-Dimensional Blocking: A Differentiated Regulatory Architecture

Machine opinions generate risks at two distinct entry forms. Explicit machine opinions trigger risks at the admissibility and cross-examination layer, where the question is whether externally submitted AI-generated materials can be subjected to meaningful reliability review and adversarial testing before entering the factual record; these risks call for a procedural-blocking response addressed to admission and cross-examination. Implicit machine opinions, by contrast, generate risks at the organizational layer: the concern here is whether court-internal assistance systems, once coupled with performance evaluation, transform from reference tools into instruments of organizational discipline; these risks call for a managerial-blocking response addressed to internal organizational coupling. At a common downstream point, however, the two risk profiles converge — namely, the formation of judicial conviction, at which the algorithmic appearance of superiority can exert preemptive directional influence regardless of which entry form supplies the influence; this convergence calls for a cognitive-blocking response operating at a regulatory layer distinct from both procedural and managerial blocking, and necessarily covering both entry forms at that distinct layer.

The three-dimensional blocking framework developed in this section is structured accordingly. Procedural blocking and managerial blocking function as specialized responses to specialized risks originating in distinct entry forms; cognitive blocking functions as a shared response to the common downstream risk of preemptive directional influence on judicial conviction-forming that both forms generate. This architecture is neither a uniform response to machine opinion in general nor three parallel responses proceeding in isolation, but a differentiated allocation calibrated to where and how generative AI intervenes in adjudicative formation. All three mechanisms serve a common normative purpose: the institutional protection of judges' cognitive sovereignty as a normatively anchored institutional capacity in the sense developed in §3.0.

4.1 Procedural Blocking: Substantive Reliability Gatekeeping for Explicit Machine Opinions

Procedural blocking's regulatory object is limited to externally submitted explicit machine opinions. Through admissibility-stage substantive reliability review, dynamic credibility assessment, objection-trigger procedures, and public algorithmic-expert support, procedural blocking transforms the cross-examination breakdown into a structured gatekeeping process. Four institutional measures constitute the procedural blocking framework.

First, substantive reliability gatekeeping at the admissibility threshold. Drawing on the proposed reliability-review standards for AI evidence (Advisory Committee on Evidence Rules, 2025): for machine outputs that functionally serve as expert opinions, the proponent should provide sufficient explanation regarding the training data foundation, methodological logic, error rates, and applicability to the case at hand. Where the proponent fails to provide sufficient explanation, the output should not be admitted as evidence.

Second, dynamic credibility assessment that replaces the static authentication paradigm. Given that the reliability of AI models may fluctuate with changes in version, parameters, training data, and operating conditions, prior review conclusions should not be treated as final confirmations (EU AI Act, 2024). Procedural mechanisms should accordingly require ongoing documentation of model state at the time of output generation, with such documentation subject to disclosure on objection.

Third, an objection-trigger procedure that activates intensified review. Where a party raises specific technical objections to a machine opinion's reliability — for instance, alleging deepfake synthesis, training-data illegality, or output-instability under perturbation — the court should, in principle, first require the prosecution to provide sufficient explanation regarding the model's compliance, foundational reliability, and applicability to the case, and should exercise caution in treating the output as a basis for substantive adjudication pending review (cf. Daubert v. Merrell Dow Pharmaceuticals, 1993). This is a procedural rather than substantive standard: it requires the proponent to bear the burden of explanation when the opponent raises specific reliability objections, without requiring the proponent to satisfy a particular epistemic threshold. Comparable procedural proposals are under consideration in U.S. federal evidence-rule reform as well (Advisory Committee on Evidence Rules, 2025).

Fourth, a dual-track public algorithmic-expert support mechanism. Recognizing that judges and defense counsel may lack the technical expertise to subject machine opinions to substantive review, courts may, ex officio or upon application, appoint neutral algorithmic experts to assist in identifying technical issues and explaining relevant methodological limitations (Federal Rules of Evidence, Rule 706). Defendants without resources to retain their own technical experts should, drawing on the institutional logic of state-provided necessary expert assistance, receive publicly funded defensive technical expert assistance to help them understand relevant technical materials and prepare targeted cross-examination (cf. Ake v. Oklahoma, 1985). The institutional objective of procedural blocking is not to achieve complete algorithmic transparency, but rather — on the premise that the structural review-capacity gap cannot be entirely eliminated — to prevent insufficiently verified machine opinions from entering the adjudicative chain through responsibility reallocation, strengthened disclosure obligations, and supplemented review capacity.

4.2 Cognitive Blocking: Direct Intervention at the Conviction-Formation Layer

Cognitive blocking is characterized by two intersecting features that distinguish it from procedural and managerial blocking. First, its regulatory layer — the formation of judicial conviction itself — is distinct from both the admissibility layer addressed by procedural blocking and the organizational layer addressed by managerial blocking. Procedural blocking controls what enters the factual record at the admissibility stage; managerial blocking removes structural pressures at the organizational stage; cognitive blocking acts directly on the conviction-formation process where, after admission and within whatever organizational environment obtains, the algorithmic appearance of superiority would otherwise exert preemptive directional influence on judicial judgment. Second, within this distinct regulatory layer, cognitive blocking necessarily spans both entry forms — explicit and implicit — because the cognitive layer at which preemptive directional influence operates is unitary, not partitioned by entry form. The specific operational mechanisms through which cognitive blocking engages explicit versus implicit machine opinions may differ in their concrete modalities, but those operational differences occur within a single regulatory layer directed at a unitary normative object: the protection of judges' institutional capacity for independent conviction-forming.

The following addresses three intervention points at this layer: ex ante detection (counterfactual reasoning); ex post evaluation (tiered probative-weight control); and temporal sequencing (the human-before-machine ordering of judicial exposure to algorithmic recommendations).

First, a counterfactual reasoning mechanism should be introduced as the substantive review tool for cognitive blocking. Counterfactual reasoning examines whether the AI model's output would have changed had specific features been altered — for instance, whether removing references to a defendant's racial background or zip code would have changed a risk-assessment prediction (Wachter, Mittelstadt, & Russell, 2018). When counterfactual analysis reveals that the model's output exhibits sensitivity to features that should be normatively irrelevant, this provides specific grounds for skepticism about the output's reliability as a basis for adjudicative conviction (Santosh et al., 2022). Counterfactual reasoning is particularly valuable in legal contexts because it can be applied without requiring full algorithmic transparency: even when the underlying model remains a black box, counterfactual probing can reveal whether the model's output structure reflects normatively defensible reasoning patterns (Ye, 2026; Shen H. & Yang, 2025). The technical generation of counterfactual outputs and the assessment of their results are not the judge's direct responsibility but rely on the dual-track public algorithmic-expert support mechanism established under procedural blocking (§4.1). Under that mechanism, the responsibility of the judge is the substantive review of the expert-supplied counterfactual analysis, not the engineering-level operation that produces it; this division of labor preserves the institutional realism of the cognitive-blocking design by aligning judicial responsibilities with judicial competence.

Second, a tiered probative-weight control mechanism should be established. AI outputs that fail substantive reliability review should, in principle, not be granted evidentiary qualification. Machine opinions that have been admitted but still exhibit deficiencies such as insufficient explanation, weak reproducibility, or excessive weight given to non-legal feature influence on decisions should be subject to strict scrutiny in probative-weight evaluation, with their probative weight restricted as appropriate based on the severity of deficiencies (cf. Daubert v. Merrell Dow Pharmaceuticals, 1993). Generative AI outputs that create new information, make

independent inferences, or functionally substitute for expert judgment require particular vigilance and should, in principle, serve only as leads, auxiliary means, or materials credible only after corroboration by independent evidence. In criminal proceedings, they should especially not serve as sole grounds for conviction; foreign legislative proposals have similarly limited the independent probative function of such outputs (New York State Assembly Bill A1338, 2025). The basis for this restriction is that such machine opinions, even after passing admissibility review, may still exert influence in fact-finding beyond their actual reliability; Under the EU AI Act, high-risk AI systems are subject to continuing obligations concerning technical documentation, record-keeping and logging, human oversight, and accuracy, robustness, and cybersecurity, such that reliability is better understood not as a status conclusively established once and for all, but as a condition that must be continuously maintained, monitored, and revisited over time (EU AI Act, 2024, arts. 11, 12, 14, 15).

Third, for high-impact adjudicative assistance scenarios, a "human-before-machine" temporal sequencing mechanism should be constructed. This mechanism responds to the problem of excessive judicial reliance on internal implicit machine opinions rendering human intervention merely formal (Zhang S., 2024). Its core lies not in rejecting algorithmic assistance itself, but in controlling the temporal sequence of judges' exposure to algorithmic recommendations, thereby preventing AI from intervening before judges have formed preliminary independent judgments to directionally guide their evidence screening and conclusion trajectories. At the normative level, the EU AI Act and the European Commission for the Efficiency of Justice (CEPEJ) have both established "human oversight" and "user control" as core principles for high-risk AI systems, emphasizing that the ultimate decision must remain a human-led decision (EU AI Act, 2024; CEPEJ, 2023). Empirical research has shown that when decision-makers are exposed to algorithmic recommendations before forming autonomous judgments, even erroneous recommendations may significantly reduce the independence of subsequent review (Agudo et al., 2024). The institutional implementation of this temporal sequencing, however, admits multiple pathways at the level of institutional design type, each carrying distinct trade-offs between enforcement strength and procedural flexibility.

Three pathways for the institutional implementation of human-before-machine sequencing can be identified at the level of institutional design type. The first is **temporal gating**: institutional rules that condition the system's display of recommendations on the judge's progression through specified procedural stages, controlling exposure timing through procedural rather than technical thresholds — for instance, a rule that recommendations are not displayed until after the judge has completed file review and substantive examination of the case materials. Temporal gating is procedurally explicit and norm-grounded but depends on the operational definition of the procedural-stage thresholds.

The second is **procedural gating**: institutional rules that require judges to record their preliminary independent assessment in the case file before the system unlocks recommendations, with the record subject to ex post review. Procedural gating creates documentary evidence of the human-led independent judgment at the moment it is required to occur, supporting both internal supervision and external accountability; its limitation lies in the additional documentary burden it places on judges and the difficulty of distinguishing genuine independent assessment from formal compliance with the recording requirement.

The third is **incentive-based gating**: no architectural enforcement of sequencing, but performance-evaluation criteria that exclude or specifically scrutinize cases in which judges consulted system recommendations before independent preliminary assessment, providing indirect institutional incentive against premature consultation. Incentive-based gating preserves judges' procedural flexibility and avoids architectural rigidity, but its effectiveness depends on the design and consistent application of the corresponding evaluation criteria.

The choice among these pathways depends on the specific assistance scenario, case-type sensitivity, and the institutional capacity for ex post supervision. These pathways are presented as the institutional design types whose trade-offs require open consideration in actual deployment, without prescriptive commitment to technical parameters that lie beyond the scope of legal-institutional analysis. The "human-before-machine" temporal sequencing, in any of these forms, does not negate the principle of human oversight; rather, by controlling exposure timing, it helps reduce the preemptive directional influence of algorithmic recommendations on judges' independent conviction-forming.

4.3 Managerial Blocking: Decoupling Internal Assistance Systems from Performance Discipline

Managerial blocking primarily targets the use of implicit machine opinions in the operation of court-internal adjudicative systems and their coupling with performance management. Its regulatory objective is to prevent the institutional bundling of AI assistance system outputs with judicial evaluation, deviation alerts, and accountability mechanisms from systematically eroding judges' capacity for independent judgment formation.

The necessity of managerial blocking also derives from the principle of adjudicative responsibility: adjudicative conclusions should be independently rendered by judges who have personally experienced the case, formed conviction, and bear reasoning obligations (Supreme People's Court, 2015). If internal system outputs become institutionally bundled with performance evaluation, judges' deviation from algorithmic recommendations may be transformed into additional explanation obligations, communication costs, and even unfavorable assessment consequences (Zhang S., 2024). The transformation of judicial role under managerial pressures is not unprecedented in the comparative literature: Resnik (1982) identified, in a different jurisdictional context, how the institutional pressures of case management movements in judicial administration can reshape the judge's role from neutral adjudicator toward case-flow manager, with consequences for how adjudicative judgment is exercised. The present concern is structurally analogous in form, though distinct in its specific institutional architecture: in the Chinese digital-judicial context, the relevant pressures emerge through the coupling of AI assistance systems with quantified performance evaluation rather than through the case-management-oriented mechanisms Resnik analyzed. In this regard, while the "auxiliary" positioning has already been affirmed at the policy level, the institutional significance of managerial blocking lies precisely in converting this macro-level positioning into operationalizable organizational constraints, preventing the institutional imbalance whereby judicial responsibility remains attributed to judges while their space for independent judgment is preemptively compressed.

Before turning to specific institutional design, the regulatory scope of managerial blocking, as proposed in this article, requires explicit demarcation. The article's proposal is normatively scoped to the design of assessment dimensions — identifying which categories of data should not feed into performance evaluation, preserving a

legally defensible institutional space for departure from system recommendations, and reorienting evaluation criteria toward explanatory rigor. The proposal does not extend to the broader transformation of judicial-administration governance, the political-economy reconfiguration of case-management hierarchies, or the comprehensive reform of judicial evaluation systems as such; these are general questions of judicial reform that exceed the specific procedural-law response this article is proposing.

Two clarifications attend this scope demarcation. First, demarcating the proposal's scope at the level of assessment-dimension design does not deny that informal channels — administrative communication, supervisory review, professional reputation effects — may continue to generate pressures of the kind managerial blocking is designed to address. The proposal's scope reflects a judgment about what procedural-law analysis can responsibly prescribe, not a denial that other institutional arrangements may also bear on the same risks. Comprehensive treatment of the political-economy dimensions of judicial AI governance would require analytical resources beyond those of the present article. Second, the design of any specific managerial-blocking institutional arrangement must rest on an accurate assessment of the actual operational state of AI assistance systems in Chinese courts at the time of design, rather than on the policy-discourse representation of those systems' embeddedness. Local empirical scholarship has identified a substantive gap between policy enthusiasm and the realities of judicial AI deployment in China, with many systems remaining at preparatory or symbolic stages of operation rather than achieving routinized practical use (Zuo, 2019). This local-context observation provides an important calibration: managerial blocking, as proposed here, should be understood as institutionally proactive rather than reactive — designed to address coupling risks at the current relatively limited stage of AI embeddedness, so as to forestall the more difficult corrective response that would be required after deeper embedding has consolidated.

First, institutional norms should explicitly provide that data from any dimension of AI system usage — including degree of deviation from algorithmic recommendations, frequency of system consultation, and concordance between final judgment and system output — should not enter judicial performance evaluation, accountability assessment, or promotion procedures. The objective is to prevent algorithmic outputs from acquiring the institutional weight of management commands through performance bundling, returning them to their nominal status as auxiliary references.

Second, a legally defensible institutional space for departure from system recommendations should be preserved. When judges, based on case-specific judgment, deviate from system recommendations, the institutional response should accommodate such deviation rather than treat it as anomalous. The procedural treatment of departure should require only that the judge's reasoning for the substantive judgment is articulable, not that the deviation per se requires explanation — the latter requirement reverses the institutional positioning of algorithmic outputs from auxiliary to authoritative (Gong, 2023).

Third, the underlying logic of judicial evaluation should be reoriented from concordance with system outputs toward the rigor and articulability of judges' substantive reasoning. This third measure responds to the deeper institutional question raised by the implicit machine opinion problem: when judicial evaluation tracks system concordance, it indirectly elevates system outputs to evaluative standards; when judicial evaluation tracks reasoning rigor, it preserves the institutional space within which judges develop independent conviction even when departing from system recommendations. An independent external supervision mechanism should be

established to prevent internal management actors from being simultaneously rule-maker and adjudicator over conformity to such rules.

Through this combination of measures, the institutional bundling between implicit machine opinions and performance discipline can be loosened, allowing judges' independent capacity for adjudicative conviction-forming to operate within a managerial environment that does not, by its design, structurally compress that capacity. Managerial blocking, so designed, is the institutional precondition for cognitive sovereignty in the implicit-machine-opinion context, just as procedural blocking is the institutional precondition for cognitive sovereignty in the explicit-machine-opinion context.

5. The Inherent Limits of Litigation-Internal Blocking and the External Institutional Conditions

5.1 The Inherent Limits of Three-Dimensional Blocking

The three-dimensional blocking framework developed above operates within litigation procedures and judicial organization. Its operative conditions, however, depend on the upstream technological state of the AI systems being regulated — a state shaped well before any output enters the courtroom or any system is deployed in a court. Procedural blocking presupposes that the technical information necessary for substantive reliability review is, in principle, accessible. Cognitive blocking presupposes that the algorithmic outputs subject to counterfactual probing are amenable to such probing. Managerial blocking presupposes that the AI systems being decoupled from performance evaluation are themselves subject to baseline institutional constraints on their deployment, function, and update. Where these upstream technological conditions are not met, the corresponding internal blocking mechanism faces inherent limits not by virtue of its own design but by virtue of the conditions on which its operation depends.

Procedural blocking faces a parallel constraint: corpus selection during model training, preference alignment, version updates, and scenario deployment have already shaped the content boundaries and judgment tendencies of outputs before those outputs are formed (Bender et al., 2021). When upstream training-data, model-update, and scenario-deployment conditions are not transparent, even procedural blocking's intensified review measures encounter ceilings: the technical materials necessary for substantive review may simply not be producible by the proponent, regardless of the procedural rule requiring their production.

Cognitive blocking faces a comparable constraint at the model-output layer. Counterfactual probing requires that the model's outputs be sensitive to feature alterations in ways that are detectable by external probing; where the model's architecture systematically obscures such sensitivity — for instance, where outputs depend on deeply nested compositional structures whose feature attribution is not externally recoverable — counterfactual probing's substantive review function is constrained (Weidinger et al., 2022).

Managerial blocking faces a parallel constraint: without upstream constraints on the deployment boundaries, functional purposes, and update rules of internal assistance systems, managerial blocking risks operating as post-hoc pressure relief rather than addressing the prior institutional question of pre-deployment conditions, cross-stage usage boundaries, and the nature of system outputs (CAC et al., 2023).

5.2 External Institutional Conditions: Three Correspondences with the Three-Dimensional Architecture

External support manifests in three dimensions, each corresponding to one mechanism within the three-dimensional blocking architecture.

(1) External support for procedural blocking — minimum review footholds. For generative outputs intended to enter judicial processes, the necessary conditions of source records, version identification, operational traces, and generation documentation should be met to the extent possible, providing the minimum material foundation that procedural blocking's objection-trigger, reliability-review, and reproducibility-testing mechanisms require. The training-data compliance review and algorithmic transparency requirements established in the Interim Measures for the Management of Generative Artificial Intelligence Services can serve as an institutional reference for pre-deployment assessment of generative AI products intended for judicial scenarios (CAC et al., 2023).

(2) External support for cognitive blocking — minimum knowability preconditions. For intelligent assistance systems that may influence judges' judgment formation, arrangements regarding usage boundaries, exposure timing, functional documentation, and risk warnings should prevent such systems from entering judges' judgment starting points without explanation. This arrangement constitutes the organizational precondition for the human-before-machine temporal sequencing developed in §4.2 to be practically implemented; without clear functional documentation and usage-boundary labeling, judges, even with a disposition toward prudence, cannot determine when to initiate independent judgment and when to consult system outputs.

(3) External support for managerial blocking — clear deployment boundaries and organizational preconditions. For all judicial-adjudication internal systems carrying machine opinions, their reference nature, scope of use, and the premise that they must not directly serve as assessment bases should be institutionally clarified, ensuring that the performance decoupling, reasonable departure space, and reasoning-oriented evaluation developed in §4.3 are not re-hollowed in practice.

Such external arrangements are not substitutes for litigation-internal blocking; they are upstream preconditions whose absence renders internal blocking partially or wholly inoperable. The relationship is one of complementarity rather than substitution: external arrangements supply the minimum conditions on which internal blocking depends, while internal blocking supplies the procedural-law and adjudicative-organization mechanisms that transform external preconditions into actual protection of judges' cognitive sovereignty.

6. Conclusion

This article has developed three interrelated propositions concerning the judicial regulation of generative AI. First, the distinctive features of generative AI outputs in adjudication — probabilistic generation, non-factual character, and the absence of a human subject capable of bearing perjury liability — produce institutional consequences that existing evidentiary classifications cannot adequately accommodate; the functional analytical category of "machine opinion," as developed in §2, organizes a regulatory analysis around these distinctive features without proposing a new statutory evidence type. Within this category, the differentiation between explicit machine opinions (externally submitted) and implicit machine opinions (court-internally

embedded) supports a structurally specific allocation of regulatory responses. Second, the three-dimensional blocking framework developed in §4 implements this allocation: procedural and managerial blocking function as specialized responses to specialized risks originating in distinct entry forms; cognitive blocking, operating at a regulatory layer distinct from both, addresses the common downstream risk of preemptive directional influence on judicial conviction-forming. Third, as developed in §5, litigation-internal blocking faces inherent limits where the upstream technological conditions are not adequately shaped; external institutional support, providing minimum upstream conditions corresponding to each blocking mechanism, is therefore necessary as a complement rather than a substitute. Across these three propositions, judges' cognitive sovereignty — as defined in §3.0 along normative-anchoring grounds — provides the protected institutional capacity that the entire regulatory architecture is designed to support.

Several limitations of this article warrant explicit acknowledgment, and they correspond to specific points at which further research is most needed. The article's analytical context is Chinese procedural law; while the conceptual framework of machine opinion and the three-dimensional architecture address structural challenges any legal system integrating generative AI into adjudication will confront, the institutional specifics — including the precise design of admissibility rules, the operational thresholds of human-before-machine sequencing, and the regulatory scope of managerial blocking — depend on jurisdictional features that any extension would need to revisit. Beyond this jurisdictional limit, three substantive limitations remain. First, the article does not include independent treatment of the distinctive procedural features of civil litigation — including the principle of party disposition, evidence contracts, and the embedding of AI in online litigation; these features warrant focused doctrinal investigation in their own terms. Second, the institutional analysis of §3.2 and §4.3 develops normative inferences from identifiable institutional conditions in China's digital-judicial architecture; whether and how the resulting structural pressures translate into observable behavioral patterns among Chinese judges remains an empirical question, as the article's methodological note in §3.2 acknowledges. Third, the technical operability of counterfactual reasoning and other cognitive-blocking mechanisms in actual judicial settings — including the conditions under which the dual-track public algorithmic-expert support mechanism in §4.1 can supply counterfactual analyses at the scale that systematic adoption would require — remains globally in a nascent state and requires sustained interdisciplinary research.

Three lines of further inquiry follow directly from the limitations enumerated above and constitute, in our view, the most pressing extensions of the present analysis. First, focused doctrinal-empirical research is needed on how Chinese judges actually interact with AI assistance systems under existing performance-evaluation regimes, with particular attention to whether the institutional-risk hypotheses developed in §3.2 obtain as observable behavioral patterns. Such research would not only test the empirical foundations of the present argument but also calibrate which of the three implementation pathways for human-before-machine sequencing developed in §4.2 is best suited to specific assistance scenarios. Second, comparative work is needed on how different legal systems integrating generative AI into adjudication respond to the structural challenges this article has identified — in particular, on how regulatory frameworks such as the EU AI Act's lifecycle documentation regime, the proposed U.S. Federal Rule of Evidence 707, China's Smart Court initiative, and the Council of Europe's 2024 Framework Convention shape, in different institutional environments, the implementation and effectiveness of analogous blocking mechanisms. Third, sustained

interdisciplinary research is needed on the technical-feasibility conditions of cognitive-blocking mechanisms — particularly counterfactual reasoning, tiered probative-weight control, and the operational design of public algorithmic-expert support — at the scale that systematic adoption in judicial practice would require. These three lines of inquiry are mutually reinforcing: empirical findings on judicial behavior would calibrate institutional design; comparative analysis would identify cross-system transferable lessons; technical-feasibility research would determine the institutional-realism boundary within which legal-institutional analysis can responsibly prescribe operational mechanisms.

Generative AI's intervention in judicial adjudication is structurally novel in ways that existing procedural and evidentiary frameworks were not designed to accommodate. The judicial regulation of this intervention requires neither rejection of technological assistance nor uncritical adoption of it, but the construction of layered institutional arrangements that maintain judges' institutional capacity for independent conviction-forming as a normatively protected matter — a capacity that, this article has argued, is best understood through the concept of cognitive sovereignty and best protected through the differentiated allocation of procedural, cognitive, and managerial blocking mechanisms, supported by the upstream institutional conditions external support is meant to provide.

Author Contributions

Conceptualization, Methodology, Writing – original draft, Writing – review & editing: Y. Chen.

Competing Interests

The author declares no competing interests.

Data Availability

This article is a normative legal analysis; no datasets were generated or analyzed.

References

[1] Advisory Committee on Evidence Rules. (2025). Report of the Advisory Committee on Evidence Rules. United States Courts. https://www.uscourts.gov/sites/default/files/document/2025-12-01_evidence_rules_committee_report.pdf

[2] Advisory Committee on Evidence Rules. (2026). Evidence Rules Hearing Schedule and Testimony Packet. United States Courts. https://www.uscourts.gov/sites/default/files/document/jan-29-hearing-schedule-and-testimony-packet.pdf

[3] Agudo, U., Liberal, K. G., Arrese, M., et al. (2024). The impact of AI errors in a human-in-the-loop process. *Cognitive Research: Principles and Implications, 9*, Article 1.

[4] *Ake v. Oklahoma*, 470 U.S. 68 (1985).

[5] Bender, E. M., Gebru, T., McMillan-Major, A., et al. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM.

[6] Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., et al. (2021). Machine unlearning. In *2021 IEEE Symposium on Security and Privacy* (pp. 141–159). IEEE.

[7] CEPEJ. (2023). *Assessment Tool for the Operationalisation of the European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment*. Council of Europe. https://rm.coe.int/cepej-2023-16final-operationalisation-ai-ethical-charter-en/1680adcc9c

[8] Chen, R. (2018). Theoretical development of the illegal evidence exclusion procedure [非法证据排除程序的理论展开]. *Comparative Law Studies* [比较法研究], (1), 1–10. (In Chinese)

[9] Chen, R. (2021). *Criminal Evidence Law* [刑事证据法学] (4th ed.). Peking University Press. (In Chinese)

[10] Cheng, L. (2021). Problems and pathways of AI-assisted sentencing [人工智能辅助量刑的问题与出路]. *Journal of Northwest University (Philosophy and Social Sciences Edition)* [西北大学学报(哲学社会科学版)], 51(6), 163–174. (In Chinese)

[11] Council of Europe. (2024). *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*, CETS No. 225. Council of Europe. https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

[12] Cyberspace Administration of China (CAC) et al. (2023). *Interim Measures for the Management of Generative Artificial Intelligence Services* [生成式人工智能服务管理暂行办法]. (In Chinese)

[13] Dahl, M., Magesh, V., Suzgun, M., et al. (2024). Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis, 16*(1), 64–93.

[14] *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993).

[15] EU AI Act. (2024). *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)*. Official Journal of the European Union, L 2024/1689.

[16] Fan, C. (2024). The surveilled judge: The impact of digital justice on the operation of adjudicative power [被敞视的法官：数字司法对审判权运行的影响]. *Law and Social Development* [法制与社会发展], (3), 137–153. (In Chinese)

[17] Federal Rules of Evidence, Rule 702 (as amended effective Dec. 1, 2023). United States Courts.

[18] Federal Rules of Evidence, Rule 706. United States Courts.

[19] Gless, S. (2020). AI in the courtroom: A comparative analysis of machine evidence in criminal trials. *Georgetown Journal of International Law, 51*(2), 195–253.

[20] Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121–127.

[21] Gong, S. (2023). Practical reflections and improvement of AI judicial application [人工智能司法应用的实践审思与完善]. *Journal of National Prosecutors College* [国家检察官学院学报], (5), 95–108. (In Chinese)

[22] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations*.

[23] Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 1–24.

[24] Han, X. (2025). Research on the evidential capacity of generative AI information represented by DeepSeek [以DeepSeek为代表的生成式人工智能信息证据能力问题研究]. *Journal of Hunan University (Social Sciences)* [湖南大学学报(社会科学版)], 39(4), 131–139. (In Chinese)

[25] Hauer, M. P., Krafft, T. D., & Zweig, K. (2023). Overview of transparency and inspectability mechanisms to achieve accountability of artificial intelligence systems. *Data & Policy, 5*, e36.

[26] Li, D. M., Zhang, H. L., & Ju, Q. R. (2025). Statistical significance, narrative, and the scholastic fallacy: How ritualized statistics exaggerate social science theories. *Transformative Society, 1*(2), 39–62. https://doi.org/10.63336/TransSoc.28

[27] Lin, P., Abney, K., & Bekey, G. A. (Eds.). (2011). *Robot Ethics: The Ethical and Social Implications of Robotics*. MIT Press.

[28] Liu, Y. (2024). Algorithm-centered governance model for AI judicial security risks [人工智能司法安全风险的算法中心治理模式]. *Oriental Law* [东方法学], (4), 83–94. (In Chinese)

[29] *New York State Assembly Bill A1338 (2025–2026)*, proposed CPL § 60.80 and CPLR § 4552.

[30] NIST. (2024a). *Reducing Risks Posed by Synthetic Content (NIST AI 100-4)*. National Institute of Standards and Technology. https://www.nist.gov/publications/reducing-risks-posed-synthetic-content-overview-technical-approaches-digital-content

[31] NIST. (2024b). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1)*. National Institute of Standards and Technology.

[32] Norman, J., & Farid, H. (2024). An investigation into the impact of AI-powered image enhancement on forensic facial recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 4306–4314).

[33] Okolo, C. T., Agarwal, D., Dell, N., et al. (2024). "If it is easy to understand, then it will have value": Examining perceptions of explainable AI with community health workers in rural India. *Proceedings of the ACM on Human-Computer Interaction, 8*(CSCW1), Article 71.

[34] Resnik, J. (1982). Managerial judges. *Harvard Law Review, 96*(2), 374–448.

[35] Roth, A. (2017). Machine testimony. *Yale Law Journal, 126*(7), 1972–2053.

[36] Roth, A. (2023). How machines reveal the gaps in evidence law. *Vanderbilt Law Review, 76*(6), 1631–1652.

[37] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.

[38] Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy and Technology, 34*(4), 1057–1084.

[39] Santosh, T. Y. S. S., Xu, S., Ichim, O., & Grabmair, M. (2022). Deconfounding legal judgment prediction for European Court of Human Rights cases towards better alignment with experts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 1120–1138).

[40] Shen, L. (2025). The impact of artificial intelligence on criminal proof structures and institutional responses [人工智能对刑事证明结构的影响及其制度因应]. *Chinese Journal of Law* [法学研究], (6), 204–222. (In Chinese)

[41] Shen, H., & Yang, Y. (2025). On the legality review rules for AI evidence in criminal proceedings [论刑事诉讼中人工智能证据的合法性审查规则]. *Journal of UESTC (Social Sciences Edition)* [电子科技大学学报(社科版)]. https://doi.org/10.14071/j.1008-8105(2025)-3046 (In Chinese)

[42] Supreme People's Court. (2015). *Several Opinions on Improving the Judicial Responsibility System of People's Courts* [关于完善人民法院司法责任制的若干意见]. Fa Fa No. 13. (In Chinese)

[43] Supreme People's Court. (2022). *Opinions on Regulating and Strengthening the Judicial Application of Artificial Intelligence* [关于规范和加强人工智能司法应用的意见]. Fa Fa No. 33. (In Chinese)

[44] Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology, 31*(2), 841–887.

[45] Wei, B. (2024). Legal argumentation analysis of the explainability difficulty of judicial artificial intelligence [司法人工智能可解释性难题的法律论证分析]. *Law and Social Development* [法制与社会发展], 30(4), 76–92. (In Chinese)

[46] Weidinger, L., Mellor, J., Rauh, M., et al. (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214–229). ACM.

[47] Xiong, X. (2025). Dilemmas and regulatory pathways for generative AI evidence determination [生成式人工智能证据认定的困境与规范进路]. *Legal Science* [法律科学(西北政法大学学报)], 43(1), 72–93. (In Chinese)

[48] Ye, X. (2026). Predictive evidence in criminal proceedings: Attributes, risks, and application rules [刑事诉讼中的预测性证据：属性、风险与适用规则]. *Journal of USTB (Social Sciences Edition)* [北京科技大学学报(社会科学版)], 42(1), 85–96. (In Chinese)

[49] Yu, P. (2023). Phenomenon, principle, and regulation: The path toward integrating AI justice with criminal procedural justice [现象、原理和规制：人工智能司法与刑事程序正义的融合之路]. *Tianfu New Idea* [天府新论], (1), 108–123. (In Chinese)

[50] Yu, P. (2024). Legal nature and rules of application of artificial intelligence evidence in criminal procedure [刑事诉讼中人工智能证据的法律性质和运用规则]. *Chinese Journal of Criminal Law* [中国刑事法杂志], (5), 36–54. (In Chinese)

[51] Zhang, B. (2024). Adaptation, risks, and optimization of embedding generative AI in Smart Court construction [生成式人工智能嵌入智慧法院建设的适配、风险与优化]. *Journal of Southeast University (Philosophy and Social Sciences)* [东南大学学报(哲学社会科学版)], 26(Suppl. 2), 95–99. (In Chinese)

[52] Zhang, D. (2022). Application of AI in criminal proof: Precise positioning, conceptual reflection, and path optimization [刑事证明中人工智能的应用：精准定位、理念反思与路径优化]. *Journal of HUST (Social Sciences)* [华中科技大学学报(社会科学版)], 36(4), 64–73. (In Chinese)

[53] Zhang, S. (2023). Legal regulation of criminal justice AI: Constructing substantive human intervention mechanisms [刑事司法人工智能的法律规制——以构建实质化人工干预机制为视角]. *Journal of BIT (Social Sciences)* [北京理工大学学报(社会科学版)]. https://doi.org/10.15918/j.jbitss1009-3370.2023.1019 (In Chinese)

[54] Zhang, S. (2024). On substantive human intervention in criminal justice AI [论刑事司法人工智能的实质化人工干预]. *Shandong Social Sciences* [山东社会科学], (3), 183–192. (In Chinese)

[55] Zheng, X. (2023). Application of generative AI in the judiciary: Prospects, risks, and regulation [生成式人工智能在司法中的运用：前景、风险与规制]. *China Applied Jurisprudence* [中国应用法学], (4), 81–93. (In Chinese)

[56] Zuo, W. (2019). Heat and cold: Rethinking legal artificial intelligence in China [热与冷：中国法律人工智能的再思考]. *Global Law Review* [环球法律评论], (2). (In Chinese)

Footnotes

1. The "algorithmic poison tree" is the present article's own analogical extension of the poisonous-tree doctrine in illegal evidence exclusion theory, as developed in the Chinese procedural law context by Chen R. (2018). It is used here to illuminate the technical mechanism through which training-data illegality propagates to model outputs, rather than to advocate direct application of the full doctrinal apparatus of the illegal evidence exclusion rule to AI-generated evidence. ↩