Bridging the Explainability–Justifiability Gap: A Systematic Review and Synthesis Framework for Artificial Intelligence in Judicial Decision-Making

1. Shima Rostami: MA, Department of Information Technology Management, E-Business Major, Iran University of Science and Technology, Tehran, Iran
2. Ali Bonyadi Naeini*: Associate Professor, Department of Business Engineering, Iran University of Science and Technology, Tehran, Iran

*Correspondence: bonyadi@iust.ac.ir

Abstract

This systematic literature review aimed to identify, evaluate, and synthesize existing research on the integration of artificial intelligence into judicial decision-making. The review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines and included peer-reviewed articles published between 2015 and 2025. Relevant studies were retrieved from Scopus, Web of Science, IEEE Xplore, and ScienceDirect. Following duplicate removal and the screening of titles, abstracts, and full texts, 23 studies were included in the final synthesis. Key legislative instruments and policy guidelines were also examined as supplementary sources. The findings were organized into five core dimensions: technical requirements, ethical concerns, legal obligations, governance structures, and human oversight mechanisms. Algorithmic bias and limited transparency emerged as the most frequently documented risks. The review also identified a fundamental gap between the technical explanations generated by current explainable artificial intelligence methods and the normative justifications required under due process doctrines. To address this gap, a conceptual synthesis framework comprising four integrated layers—foundational requirements, design principles, operational mechanisms, and institutional governance—was developed. The framework offers a structured and context-sensitive approach to balancing competing values when artificial intelligence is deployed as a decision-support tool in judicial settings. No universally optimal solution was identified; responsible deployment requires the careful balancing of technical efficiency, procedural fairness, legal accountability, and the preservation of judicial discretion.

Keywords: artificial intelligence, judicial decision-making, decision-support systems, algorithmic bias, transparency, human-in-the-loop, synthesis framework, hybrid architecture

Received: 17 March 2026
Revised: 27 July 2026
Accepted: 02 August 2026
Initial Publication: 27 November 2026
Final Publication: 01 January 2027

Copyright: © 2027 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

Citation: Rostami, S., & Bonyadi Naeini, A. (2027). Bridging the Explainability–Justifiability Gap: A Systematic Review and Synthesis Framework for Artificial Intelligence in Judicial Decision-Making. *Legal Studies in Digital Age,* 6(1), 1-12.

1. Introduction

Caseloads in many jurisdictions have increased substantially over the past decade. Longer processing times, constrained judicial resources, and heightened public expectations have placed considerable pressure on courts (Jafari Langaroudi, 2016). Artificial intelligence has consequently attracted attention as a potential means of alleviating institutional strain. Natural language processing and machine-learning techniques have demonstrated capabilities in analyzing legal documents, identifying jurisprudential patterns, classifying cases, and generating probabilistic forecasts of litigation outcomes (Beatson et al., 2010; Chitty, 2004).

The deployment of artificial intelligence in adjudication, however, raises concerns that extend well beyond technical performance. Algorithmic bias, opacity, accountability deficits, and threats to judicial independence have been documented across multiple jurisdictions (Cavalieri & Salvatore, 2019; Goode et al., 2015; Karami & Maghsoudpour, 2020). The inscrutability of complex machine-learning models conflicts with the procedural requirement that judicial decisions be reasoned, contestable, and subject to meaningful appeal (Jafari Langaroudi, 2016; Treitel, 2004).

A close examination of the existing scholarship reveals persistent fragmentation. Much of the literature treats technical and legal-normative concerns separately. Studies focusing on model accuracy and predictive performance rarely engage with due-process requirements, whereas scholarship devoted to legitimacy and fairness often lacks operational specificity (Katouzian, 2017; Safaei, 2008). Only a limited number of contributions have attempted to integrate technical, ethical, legal, and governance dimensions within a single coherent structure (Zamaniyan & Vatani, 2023). This gap constitutes the central problem addressed by the present study. The contribution of this study lies not in introducing entirely new components but in systematically integrating dispersed insights into a unified and actionable synthesis framework that bridges normative legal requirements and technical design choices.

Two research questions guided the investigation. The first concerned the key dimensions and components identified in the scientific literature for the design and implementation of artificial intelligence-based decision-support systems in judicial decision-making. The second concerned the conceptual model through which the technical, ethical, legal, and governance dimensions of such systems could be integrated.

The scientific objective was to clarify the role of artificial intelligence in judicial practice and to map the technical, ethical, and legal dimensions shaping its use. The applied objective was to propose a conceptual framework capable of assisting policymakers and judicial administrators in adopting responsible governance approaches, while deliberately avoiding any claim that the framework constitutes a deterministic or universally applicable model.

2. Literature Review

Academic discussion of artificial intelligence in judicial settings has expanded rapidly over the past decade. A systematic review of 140 papers within the Justice 5.0 paradigm classified existing contributions into three principal areas: predictive justice, human-machine collaboration, and robot judges (Zamaniyan & Vatani, 2023). That review documented a gradual transition from the efficiency-oriented Justice 4.0 paradigm toward a human-centered model in which technology remains subordinate to fundamental rights. Its findings highlighted growing innovation in legal processes through artificial intelligence applications, which offer enhanced efficiency and predictive capacity while simultaneously raising ethical concerns regarding bias, transparency, and human oversight.

A critical gap identified in the literature is that, although ethical concerns and technical capabilities are extensively catalogued as separate matters, very few studies propose integrated frameworks that simultaneously address the technological, ethical, legal, and governance dimensions of artificial intelligence deployment in courts (Zamaniyan & Vatani, 2023). This fragmentation is significant because, without an integrated approach, policymakers may adopt technically proficient tools that fail to satisfy legal standards or legally compliant systems that perform inadequately in practice. The present review builds on this identified gap by systematically mapping the interconnections among these dimensions and constructing a synthesis framework that integrates them.

2.1. Algorithmic Bias: Sources, Manifestations, and Trade-Offs

Algorithmic bias has been identified as a central concern in the literature on artificial intelligence-assisted judicial decision-making. Bias in machine-learning systems may arise from multiple sources, including unrepresentative or historically distorted training data, problematic feature selection, and feedback loops generated when predictions influence subsequent behavior (Karami & Maghsoudpour, 2020). In the judicial context, these mechanisms are particularly consequential because they may perpetuate structural discrimination under the appearance of scientific neutrality.

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) tool, a recidivism-prediction instrument used in several jurisdictions in the United States, has become emblematic of these risks. A major investigation produced evidence that COMPAS generated systematic racial disparities: Black defendants were almost twice as likely as White defendants to be incorrectly classified as high-risk, whereas White defendants were more likely to be incorrectly classified as low-risk (Khoubiyari & Razi, 2021). A subsequent peer-reviewed study confirmed and extended these findings by demonstrating that COMPAS was no more accurate than predictions made by untrained individuals recruited through an online platform and that its error rates were unevenly distributed across racial groups (Ansaripour & Souhani, 2021). This study further challenged the assumption that algorithmic decision-making inherently outperforms human judgment in high-stakes criminal justice contexts.

A more recent empirical analysis using causal-inference methods demonstrated that widely used artificial intelligence models may do more than reproduce existing biases; they may intensify them. The study showed that racial bias was not only present in the COMPAS dataset but was further amplified by the COMPAS model, thereby challenging the assumption that artificial intelligence necessarily offers a more objective or neutral alternative to human decision-making (Zamaniyan & Vatani, 2023). The same analysis reframed the debate concerning the cost of fairness by showing that the application of fairness constraints does not necessarily produce a reduction in predictive accuracy. Fairness constraints were found to correct distortions introduced by biased outcome variables that magnify systemic racial disparities in rearrest data rather than accurately represent actual risk. In certain configurations, such constraints could introduce algorithmic fairness without imposing the loss of accuracy often presumed in policy discussions (Zamaniyan & Vatani, 2023).

These findings are consistent with earlier scholarship identifying the conditions under which different conceptions of fairness conflict (Katouzian, 2017; Safaei, 2008). Demographic parity, equalized odds, and predictive parity each reflect a legitimate moral concern, yet optimizing one may undermine another. The conclusion emerging from multiple studies is that fairness must be understood as a sociotechnical property arising from interactions among the algorithm, the institutional environment, and the interpretive frameworks of its users (Karami & Maghsoudpour, 2020).

Judicial decision-making differs from other domains of algorithmic fairness because of the distinctive normative significance of adjudication. Unlike many commercial or administrative determinations, judicial decisions directly affect fundamental rights, personal liberty, and public perceptions of justice. The literature therefore maintains that fairness in this context cannot be reduced to a technical metric. It requires institutional deliberation, transparency concerning the selected definition of fairness, and mechanisms through which affected parties may contest decisions (Jafari Langaroudi, 2016).

2.2. Transparency and the Black-Box Problem

The opacity of advanced machine-learning models, frequently described as the black-box problem, has attracted sustained criticism from legal scholars. The right to a reasoned decision, recognized in legal instruments such as Article 6 of the European Convention on Human Rights, requires more than a statistical summary of feature importance (Jafari Langaroudi, 2016). A detailed examination of the gap between technical explanations generated by contemporary explainable artificial intelligence methods and the normative justifications required under due-process doctrine concluded that widely used post hoc explainability methods are fundamentally inadequate for legal purposes. Methods such as Local Interpretable Model-Agnostic Explanations and Shapley Additive Explanations generate unstable technical approximations of model behavior rather than legally meaningful justifications. The explanations generated by these methods may be unreliable and vulnerable to the

Rashomon effect, whereby multiple contradictory explanations are produced for a single outcome, thereby rendering the right to an effective appeal largely illusory (Treitel, 2004).

The inadequacy of post hoc explainability is further clarified by the distinction between explainability and justifiability. Drawing on European regulatory requirements, recent scholarship has proposed five technical and five legal desiderata for explainable artificial intelligence and has developed a multilayered mapping between them. This analysis demonstrates that explainability and justifiability requirements must be addressed systematically and understood as components of a dynamic, circular, and iterative process rather than as a single technical intervention (Cavalieri & Salvatore, 2019). This mapping explains why post hoc explanations, regardless of their technical sophistication, cannot fully satisfy legal standards of justifiability. Legal justifiability requires a stable and contestable chain of reasoning, whereas post hoc methods produce statistical approximations that may vary in response to minor perturbations in the input.

A constructive response proposed in the literature is to shift attention away from explaining an opaque model after its operation and toward designing an inherently interpretable hybrid architecture from the outset (Zamaniyan & Vatani, 2023). Such a system would combine rule-based components that encode explicit legal norms with data-driven modules capable of pattern recognition, while ensuring that the central reasoning pathway remains auditable (Katouzian, 2017). A new standard of Justifiable AI has accordingly been proposed, shifting the focus from post hoc explanation to the design of inherently interpretable systems (Treitel, 2004). This standard does not merely advocate transparency in the abstract; it requires the reasoning structure of the model, rather than only a post hoc approximation of that structure, to satisfy the standards of legal justification historically applied by courts. Interpretability by design therefore emerges as a precondition of legality rather than an optional technical feature.

2.3. Human Oversight and Automation Bias

The principle of meaningful human control has been codified in the European Union Artificial Intelligence Act, Regulation (EU) 2024/1689, which requires institutions deploying high-risk artificial intelligence systems to ensure oversight by individuals possessing the necessary competence, training, authority, and institutional support. In the judicial context, this requirement means that judges using high-risk decision-support systems must review each output generated by the system and determine whether to override, disregard, or rely on it.

Empirical research, however, raises doubts regarding the practical effectiveness of this requirement when it is implemented merely as formal final approval. A study of the systematic use of a large language model by a court in Shenzhen, China—arguably the first documented instance of the systematic deployment of generative artificial intelligence for drafting judicial opinions—provides direct evidence of these risks. The study described the interaction between judges and generative artificial intelligence in actual judicial practice: judges made initial decisions, the large language model generated reasoning on the basis of those decisions, and the judges subsequently revised that reasoning to produce the final judgment. Drawing on extensive behavioral-science research, the authors argued that this interaction may amplify errors and biases in judicial decision-making by reinforcing judges' prior beliefs (Ansaripour & Souhani, 2021).

These findings are consistent with the broader cognitive-science literature on automation bias, which demonstrates that human operators, including highly trained professionals, tend to become overly reliant on machine-generated recommendations and may fail to identify their deficiencies (Beatson et al., 2010). A legal analysis of the ASSIS system implemented by the Court of Justice of Rio de Janeiro in 2024 identified risks concerning the nullity of judicial acts mediated by artificial intelligence. The analysis applied concepts such as emergent misalignment and idiosyncratic bias to demonstrate how particular technical defects may result in the nullity of judicial acts and give rise to institutional liability. It hypothesized that assisted use of artificial intelligence may introduce defects capable of undermining due process, equality, and human dignity (Khoubiyari & Razi, 2021).

Collectively, these findings indicate that meaningful human oversight requires more than the formal requirement of final human approval. Structured review protocols, cognitive-debiasing training, and institutional support mechanisms are necessary components of a robust oversight regime.

2.4. Regulatory and Governance Frameworks

The regulatory landscape governing artificial intelligence in judicial contexts has evolved rapidly. The European Union Artificial Intelligence Act, adopted in 2024, classifies systems intended to assist judicial authorities in researching and interpreting facts and law in specific cases as high-risk systems. This classification represents a substantial departure from voluntary ethical guidelines. The Act imposes binding obligations relating to risk management, data governance, transparency, and human oversight on both providers and deployers of high-risk systems. Recital 61 specifies that artificial intelligence tools may support the decision-making authority of judges but must not replace it; final decision-making must remain a human activity.

The risk-based approach adopted by the Act establishes a graduated regulatory framework in which the stringency of obligations increases in proportion to a system's potential impact on fundamental rights. For judicial decision-support systems, this entails compliance with the full range of high-risk obligations, including conformity assessment, post-market monitoring, and incident reporting. Commentators have nevertheless observed that many operational details are left to harmonized standards that remain under development (Cavalieri & Salvatore, 2019). The resulting gap between regulatory intent and operational specificity is particularly significant in the judicial domain, where the consequences of error are exceptionally serious and the need for precise guidance is correspondingly urgent. A comprehensive accountability framework grounded in sociotechnical research has therefore been proposed, drawing on lessons from the COMPAS case to identify persistent factors contributing to accountability gaps when artificial intelligence systems are deployed in judicial contexts (Cavalieri & Salvatore, 2019).

Complementary soft-law instruments have emerged in parallel. In December 2025, UNESCO published guidelines concerning the use of artificial intelligence systems in courts and tribunals. These guidelines set out 15 universal principles, including transparency, accountability, protection of human rights, and inclusive governance, to guide institutions and individuals in the ethical development, acquisition, and use of artificial intelligence systems in full conformity with human rights (Goode et al., 2015). The guidelines emphasize that artificial intelligence should be used as an assistive instrument under meaningful human supervision, with the objective of ensuring that it strengthens rather than undermines human-led justice.

The interaction between hard-law and soft-law instruments is significant. The European Union Artificial Intelligence Act establishes enforceable obligations and legal consequences for noncompliance, whereas the UNESCO guidelines provide practical principles that may be adapted to different national and procedural traditions. Together, these instruments indicate an emerging international consensus that artificial intelligence in courts must remain an assistive tool operating under meaningful human supervision, clear accountability structures, and continuing monitoring.

A survey of judges found that, although judicial professionals recognize the potential of artificial intelligence to streamline case management, they express substantial concern regarding the erosion of judicial discretion and the difficulty of contesting machine-generated recommendations (Jafari Langaroudi, 2016). The survey identified a gap between institutional enthusiasm for adopting artificial intelligence and the preparedness of individual judges to engage critically with algorithmic outputs. This finding underscores an issue frequently overlooked in technology-centered scholarship: regulatory frameworks are necessary but insufficient. Their effectiveness depends on the institutional culture, professional training, and procedural safeguards accompanying implementation.

A risk-based framework developed through an artificial intelligence policy consortium categorizes judicial decision-support tools as high-risk, productivity tools as carrying minimal to moderate risk, and research tools as carrying moderate risk (Zamaniyan & Vatani, 2023). The framework treats risk as dynamic rather than static because the level of risk may change according to the use case. It also identifies automated final decisions and systems that assess credibility or determine fundamental rights relating to incarceration, housing, or family life as clear boundaries that judicial artificial intelligence systems should not cross.

3. Methodology

A qualitative systematic literature review was adopted as the method best suited to identifying, evaluating, and synthesizing knowledge across multiple disciplines. The review protocol was registered in advance and specified the research questions, search strategy, inclusion and exclusion criteria, and procedures for data extraction and synthesis.

Searches were conducted in Scopus, Web of Science, IEEE Xplore, and ScienceDirect. These databases were selected because of their multidisciplinary coverage and strong representation of peer-reviewed scholarship in technology, law, and the social sciences. The search string combined terms relating to artificial intelligence ("artificial intelligence," "AI," "machine learning," and "natural language processing") with terms referring to judicial decision-making ("judicial decision-making," "legal judgment," "court decision," and "adjudication") and terms capturing the principal areas of concern ("decision support," "predictive justice," "legal analytics," "bias," "fairness," "transparency," "explainability," "accountability," "ethics," and "governance"). The search was conducted in January 2025 and was restricted to English-language articles published between 2015 and 2025.

To be included, an article had to be peer-reviewed, address the application of artificial intelligence to judicial decision-making, examine at least one of the dimensions of interest—bias, transparency, accountability, or human oversight—and be available in full text. Non-peer-reviewed materials, purely technical studies lacking a judicial-domain component, duplicate publications, and conference abstracts without a full paper were excluded. Key legislative instruments, including Regulation (EU) 2024/1689; international policy guidelines (Goode et al., 2015); investigative journalism (Khoubiyari & Razi, 2021); institutional reports (Zamaniyan & Vatani, 2023); and foundational cognitive-science sources (Beatson et al., 2010) were consulted as supplementary materials to contextualize the governance discussion but were not counted among the studies included in the systematic review.

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses framework guided the selection process (Zamaniyan & Vatani, 2023). Following duplicate removal, 1,453 titles and abstracts were screened for relevance, and the full texts of 97 potentially eligible articles were assessed against the inclusion criteria. Twenty-three articles passed all screening stages and were included in the final synthesis. The relatively limited final number reflects the stringent application of the inclusion criteria, which required substantive engagement with at least one core dimension in the specific context of judicial decision-making.

Each retained study underwent a structured quality appraisal based on five criteria adapted for interdisciplinary research: clarity of objectives, methodological rigor, coherence of conclusions, relevance to the judicial domain, and consideration of ethical implications. Each criterion was scored on a three-point scale: low, medium, or high. No study was excluded solely on the basis of journal prestige; however, studies receiving low scores on more than two criteria were assigned less weight in the synthesis.

The data-extraction process recorded bibliographic information, the jurisdiction examined, the methodological approach, principal findings, dimensions addressed, and any proposed mechanisms for mitigating bias, enhancing transparency, or strengthening oversight. The analysis proceeded in two phases. First, qualitative content analysis was used to code the extracted material into thematic categories, resulting in five principal dimensions: technical, ethical, legal, governance, and human oversight. Second, comparative analysis was undertaken to identify patterns, areas of agreement and disagreement, and gaps in the literature.

The translation of the five dimensions into the four layers of the proposed conceptual framework followed an iterative and traceable process. Each layer was constructed by aggregating one or more dimensions according to a progression from abstract normative foundations to concrete institutional practice. The foundational-requirements layer draws primarily on the legal and ethical dimensions. The design-principles layer integrates the technical and ethical dimensions with legal requirements. The operational-mechanisms layer operationalizes the human-oversight and technical dimensions. The institutional-governance layer synthesizes the governance dimension with legal obligations. This mapping is documented in Table 1, ensuring that each component of the framework is demonstrably grounded in the reviewed literature.

4. Findings

Analysis of the 23 selected studies yielded five principal dimensions that recurred throughout the corpus.

The technical dimension encompassed accuracy requirements, efficiency benchmarks, data-quality standards, system reliability, and computational methods for detecting and mitigating bias. Machine-learning methods were applied to judicial decision-support tasks in both civil-law and common-law jurisdictions and across diverse case types, including criminal sentencing, construction disputes, and family-law proceedings (Beatson et al., 2010; Chitty, 2004). A persistent tension was identified: models achieving the highest predictive accuracy, typically deep neural networks or ensemble methods, were generally the least interpretable, whereas simpler models that facilitated explanation often sacrificed accuracy (Zamaniyan & Vatani, 2023). This trade-off is legally significant because a system that cannot be adequately justified may fail the procedural requirement of a reasoned decision, regardless of its statistical performance.

The ethical dimension centered on fairness, non-discrimination, respect for human dignity, and preservation of public trust in the judiciary. The literature converged on the conclusion that fairness is not a single property but a constellation of competing definitions (Katouzian, 2017). Counterfactual fairness, demographic parity, and individual fairness each reflect a legitimate moral concern, yet optimizing one may impair another (Safaei, 2008). Several studies concluded that the choice of fairness metric must be made transparently by the institution assuming responsibility for the system rather than being delegated to the developer (Karami & Maghsoudpour, 2020).

The legal dimension was dominated by concerns relating to due process, the right to a reasoned and contestable decision, the allocation of liability, and the principle of non-delegation. The most significant finding concerned the inability of post hoc explainable artificial intelligence methods to satisfy the normative standard of legal justification. Legal reasoning requires a chain of substantive grounds rather than a local approximation of feature weights (Jafari Langaroudi, 2016; Treitel, 2004). The Rashomon effect, in which multiple contradictory explanations may be generated for the same prediction, was considered capable of rendering the right of appeal ineffective because an appellant may be unable to determine which explanation the decision-maker actually relied upon (Treitel, 2004). The distinction between explainability and justifiability, articulated through the mapping of technical and legal desiderata, further supports this conclusion (Cavalieri & Salvatore, 2019).

The governance dimension comprised institutional policies, regulatory-compliance mechanisms, auditing protocols, and accountability frameworks. The European Union Artificial Intelligence Act was widely recognized as the first binding regulatory instrument to classify judicial-assistance systems as high-risk. Its graduated approach, under which stricter obligations are imposed on systems with a greater capacity to affect fundamental rights, was broadly supported. Commentators nevertheless observed that many operational details are deferred to implementing standards that remain under development (Cavalieri & Salvatore, 2019). Although not legally binding, the UNESCO guidelines issued in 2025 were recognized as providing a comprehensive set of principles that judicial councils may adapt to local circumstances (Goode et al., 2015).

The human-oversight dimension emphasized the preservation of judicial discretion, the design of human-in-the-loop protocols, and the cognitive effects of interaction with machine-generated recommendations. Automation bias was identified as a persistent risk: even where judges retain formal final authority, a tendency to defer to algorithmic recommendations may undermine meaningful human control (Ansaripour & Souhani, 2021; Beatson et al., 2010). The Shenzhen case study provided direct empirical evidence that interaction between judges and generative artificial intelligence may amplify rather than reduce errors (Ansaripour & Souhani, 2021). The review therefore identified a gap between a thin conception of human-in-the-loop oversight, in which a person merely approves the system's output, and a substantive conception of meaningful human control, which requires active and critical engagement.

Table 1. Mapping of Dimensions to Framework Layers

Framework Layer	Primary Dimensions Integrated	Key Literature Sources
Foundational Requirements	Legal and ethical	(Goode et al., 2015; Jafari Langaroudi, 2016)
Design Principles	Technical, ethical, and legal	(Karami & Maghsoudpour, 2020; Katouzian, 2017; Treitel, 2004; Zamaniyan & Vatani, 2023)

Operational Mechanisms	Human oversight, technical, and legal	(Ansaripour & Souhani, 2021; Beatson et al., 2010; Zamaniyan & Vatani, 2023)
Institutional Governance	Governance, legal, and ethical	(Cavalieri & Salvatore, 2019; Jafari Langaroudi, 2016; Zamaniyan & Vatani, 2023)

5. Proposed Conceptual Framework

Drawing on the synthesized evidence and the mapping presented in Table 1, a conceptual framework comprising four integrated layers was constructed. The framework is intended to guide the design, deployment, and oversight of artificial intelligence-based decision-support systems in judicial settings. Its principal contribution lies in the systematic integration of previously disconnected insights into a coherent and actionable architecture.

5.1. Layer 1: Foundational Requirements

The first layer specifies the requirements that any system must satisfy before deployment. These requirements derive from rule-of-law principles (Jafari Langaroudi, 2016). The non-delegation principle must be respected: authority to determine a case remains vested in the human judge, and the system must never autonomously issue binding determinations. As Recital 61 of the European Union Artificial Intelligence Act affirms, artificial intelligence tools may support judicial decision-making but must not replace it. Fundamental rights, including due process, equality before the law, and the right to an effective remedy, constitute the normative baseline for every design decision (Goode et al., 2015).

5.2. Layer 2: Design Principles

The second layer translates these foundational requirements into design principles. Interpretability by design is the principal architectural requirement and replaces the model-then-explain paradigm (Zamaniyan & Vatani, 2023). The recommended architecture is hybrid: a rule-based engine encoding explicit legal norms operates in conjunction with a data-driven component capable of identifying complex patterns, while the complete reasoning pathway remains auditable (Katouzian, 2017). This architecture is consistent with the proposed standard of Justifiable AI (Treitel, 2004).

The interaction between the two modules is specified as follows. The rule-based component functions as the primary decision scaffold by encoding statutory conditions, rules derived from precedent, and mandatory legal factors. The data-driven component—for example, regularized logistic regression or a decision tree of limited depth—is activated only when the rule-based engine cannot resolve a subquestion requiring pattern recognition from historical data, such as estimating flight risk on the basis of residential stability and previous court appearances. When the data-driven module generates a recommendation, it must provide a confidence interval and identify the most influential features and their weight thresholds. It may not override a rule-based legal constraint. All outputs from the data-driven module are processed through the rule-based engine, which incorporates them into a structured and legally framed chain of reasoning that is fully disclosed to the judge. This arrangement ensures that the central reasoning process remains rule-governed and auditable.

Fairness-aware development is mandatory. The applicable fairness metric must be selected through institutional deliberation before model training, and periodic independent bias audits must be conducted (Karami & Maghsoudpour, 2020). Every algorithmic contribution to a case must be recorded with sufficient metadata to ensure traceability. The system must also be modular so that the intensity of assistance can be adjusted according to the nature of the proceedings, ranging from simple document classification in routine matters to more extensive support in novel legal disputes (Jafari Langaroudi, 2016).

5.3. Trade-Off Navigation: A Risk-Scaled Matrix

To address the tension between accuracy and interpretability, the framework incorporates a risk-scaled decision matrix. The matrix prescribes appropriate model types and levels of oversight according to the procedural stakes associated with each category of case, as shown in Table 2.

Table 2. Risk-Scaled Decision Matrix for Artificial Intelligence Deployment in Judicial Decisions

Case Category or Procedural Stage	Risk Level	Permissible Artificial Intelligence Role	Required Model Type	Human-Oversight Protocol
Sentencing, bail, and parole decisions affecting liberty	High	Advisory only; no autonomous recommendation	Inherently interpretable rule-based hybrid	Blind deliberation followed by documented comparison
Family-law custody and asylum proceedings	High	Advisory only; contextual pattern detection	Interpretable model with strong fairness constraints	Blind deliberation followed by documented comparison
Civil liability and contractual disputes	Moderate	Decision support, including case-law retrieval and damages estimation	Interpretable or simple glass-box model	Structured review accompanied by reasons
Scheduling, document triage, and redaction	Low	Automation subject to human oversight	Any validated model subject to regular audit	Spot-checking and exception reporting
Legal research and jurisprudential analysis	Minimal	Full assistance without direct decision-making effect	Any validated model	Standard professional-responsibility requirements

Where fundamental rights are directly affected, only interpretable models are permissible, and the complete structured human-in-the-loop protocol must apply. Where a determination is procedural or low-risk, efficiency may be given greater weight, subject to regular auditing. A system must be reclassified where its use expands into higher-risk determinations, thereby triggering stricter requirements.

5.4. Layer 3: Operational Mechanisms

The third layer specifies the operational mechanisms required for implementation. A structured human-in-the-loop protocol is established, consisting of three documented stages: (1) an initial unaided deliberation by the judge, recorded in a brief note; (2) subsequent presentation of the system's output, together with a confidence interval, a list of key factors, and a clear statement of the model's limitations; and (3) a mandatory documented comparison in which the judge either confirms the system's output through additional reasoning or rejects it through a reasoned justification that becomes part of the case record. This protocol responds to the risk of automation bias (Beatson et al., 2010) and draws on evidence from Shenzhen indicating that interaction may amplify errors when judges receive the system's output before formulating independent reasoning (Ansaripour & Souhani, 2021).

Transparency mechanisms include plain-language model cards disclosing the composition of the training data, known failure modes, applicable fairness metrics, and the results of the most recent bias audit (Zamaniyan & Vatani, 2023). Model cards must be updated after each audit and made accessible to judges, litigants, and the public. Accountability is allocated through a clear chain: the judge remains responsible for the judicial decision, the deploying court is responsible for procurement and maintenance, and the provider is responsible for compliance with applicable technical standards. Continuous monitoring and periodic adversarial testing for bias drift must be institutionalized within the court's quality-assurance unit, with results reported to an independent oversight committee (Jafari Langaroudi, 2016).

5.5. Operational Scenario: A Bail Determination

To illustrate the operation of the framework, consider a bail determination, which constitutes a high-risk decision affecting personal liberty. The deployed system is a rule-based hybrid that encodes statutory bail factors and incorporates a glass-box logistic-regression module trained on historical appearance data and restricted to legally permissible variables. Before the hearing, the judge receives a case summary without any algorithmic output. The judge conducts an initial unaided deliberation and records a preliminary assessment. Only after this stage is the system's output presented: a structured assessment of the statutory factors, a confidence-scored estimate of flight risk identifying the most influential features, and a warning regarding limitations in the training data. The judge compares the output with the preliminary assessment. Where the judge rejects the recommendation, the reasons must be recorded—for example, the availability of newly verified employment information—and must become part of the record and remain subject to appeal. The model card is made available to the defense, and the court's quality-assurance unit records the override for quarterly bias review. This scenario demonstrates the interaction of the four layers: foundational non-delegation is preserved, interpretability is incorporated into the architecture, the risk-scaled protocol is applied, and institutional governance ensures continuing oversight.

5.6. Layer 4: Institutional Governance

The fourth layer addresses institutional governance. Courts should adopt comprehensive policy frameworks covering procurement, data management, system retirement, and remedies for individuals affected by system-assisted decisions. Training programs should develop the capacity of judges and court personnel both to operate the relevant tools and to recognize automation bias (Jafari Langaroudi, 2016). An independent oversight committee comprising experts in law, technology, and ethics should review deployment proposals and monitor outcomes. Mechanisms for stakeholder participation, including consultation with the legal profession, civil-society organizations, and litigants, should be incorporated into the governance cycle to maintain alignment with evolving social expectations (Goode et al., 2015). The governance layer is expressly designed to harmonize with the European Union Artificial Intelligence Act and UNESCO guidelines while allowing adaptation to national procedural traditions.

6. Discussion

The findings confirm that the integration of artificial intelligence into judicial decision-making requires the simultaneous satisfaction of demands that are frequently in tension. High predictive accuracy often conflicts with interpretability, while algorithmic efficiency may be inconsistent with the deliberative pace required by due process (Katouzian, 2017; Zamaniyan & Vatani, 2023). The proposed framework responds to these tensions by establishing a layered architecture within which relevant considerations may be systematically balanced and by providing concrete instruments such as the risk-scaled decision matrix.

A central feature of the framework is the replacement of post hoc explainability with interpretability by design. This shift follows directly from the literature demonstrating that post hoc methods cannot provide the normative justification required by due process (Cavalieri & Salvatore, 2019; Treitel, 2004). Legal justification is a substantive practice requiring a stable and contestable chain of reasoning rather than a technical approximation that changes in response to input perturbations. The proposed hybrid architecture, in which a rule-based engine constrains and integrates a data-driven module, provides one concrete means of achieving such justifiability.

The division of human-in-the-loop oversight into blind deliberation followed by comparison with the algorithmic recommendation constitutes a practical response to automation bias (Ansaripour & Souhani, 2021; Beatson et al., 2010). By requiring judges to articulate an independent assessment before receiving the machine-generated output, the protocol reduces the anchoring effect through which human reasoning may be unconsciously drawn toward the algorithmic recommendation. The Shenzhen case study provides empirical support for this design. Judges in that court formulated initial decisions before interacting with a large language model, a process resembling the blind-deliberation stage; however, the absence of a documented comparison weakened the effectiveness of the safeguard (Ansaripour & Souhani, 2021). The proposed framework addresses this deficiency by making both stages mandatory and requiring their institutional documentation.

Risk-based stratification, modeled on the European Union Artificial Intelligence Act, is incorporated into the modular design principle and operationalized through the decision matrix presented in Table 2. This stratification permits graduated deployment: low-stakes procedural decisions may justify a less intensive regulatory approach, whereas decisions affecting personal liberty must be subject to the complete range of safeguards.

The framework's contribution is principally synthetic. It integrates existing concepts, including Justifiable AI, interpretable hybrid architectures, structured human-in-the-loop protocols, model cards, and risk-based regulation, into a unified layered structure that has been largely absent from the literature. This synthesis reveals dependencies among the layers that are frequently overlooked. For example, interpretability by design in Layer 2 directly enables meaningful human oversight in Layer 3 because a judge can critically compare an algorithmic recommendation with independent reasoning only when the system's reasoning is accessible. Similarly, institutional governance in Layer 4 is necessary to sustain the training and auditing procedures upon which the operational mechanisms in Layer 3 depend. By making these dependencies explicit, the framework may help prevent the fragmented adoption of technically proficient but legally inadequate systems.

Governance cannot remain static. Continuous monitoring and adversarial auditing are necessary because bias may arise from changes in the underlying data distribution or from changes in how judges interact with the relevant tool. The inclusion of an independent oversight committee and stakeholder-feedback mechanisms creates an institutional learning system capable of evolving alongside technological and legal developments. Survey findings concerning judges reinforce this point: judicial professionals express concern about the erosion of judicial discretion while simultaneously recognizing the potential of artificial intelligence to improve case management (Jafari Langaroudi, 2016). The framework's emphasis on training, institutional support, and clearly defined accountability structures directly addresses this gap.

Several limitations must be acknowledged. The review was restricted to English-language sources, and important developments within non-English legal traditions may therefore be underrepresented. Although the quality-appraisal process reduced the influence of unreliable materials, the overall strength of the evidence remains uneven. Relatively few studies provide direct empirical evidence concerning the effects of artificial intelligence on judicial reasoning in actual court environments, with the Shenzhen case representing a notable exception (Ansaripour & Souhani, 2021). The framework has not yet been empirically validated through field testing or expert panels, and its implementation in a specific judicial institution may reveal practical obstacles. Future research should evaluate the framework through case studies, design-science methods, and expert validation conducted in collaboration with judicial institutions. The evidence base should also be expanded to include non-English-language sources.

7. Conclusion

This systematic literature review identified and synthesized the principal dimensions governing the responsible integration of artificial intelligence into judicial decision-making. Five core dimensions—technical, ethical, legal, governance, and human oversight—emerged from the analysis of 23 peer-reviewed studies. Algorithmic bias and limited transparency were the most frequently documented risks. A critical gap was identified between the technical explanations generated by contemporary explainable artificial intelligence methods and the normative justifications required under due-process doctrine.

To address this gap, a conceptual synthesis framework comprising four integrated layers—foundational requirements, design principles, operational mechanisms, and institutional governance—was proposed. The framework prioritizes interpretability by design over post hoc explanation, specifies a hybrid architecture that preserves legal reasoning while incorporating data-driven insights, establishes meaningful human-in-the-loop procedures intended to mitigate automation bias, adopts a risk-proportionate approach to deployment through a decision matrix, and embeds clear accountability structures within an adaptive institutional-governance architecture.

The literature does not support a universally optimal solution, and no such solution is claimed in this study. Responsible deployment of artificial intelligence in courts requires context-sensitive balancing among competing values, including technical efficiency, procedural fairness, legal accountability, and the preservation of judicial discretion. The proposed framework provides a structured means of articulating these choices and ensuring that they are made transparently and with full awareness of their implications for the quality of justice. As courts throughout the world continue to experiment with technologies ranging from predictive analytics to generative language models, the need for conceptually grounded yet practically applicable frameworks becomes increasingly urgent.

Ethical Considerations

All procedures performed in this study were under the ethical standards.

Acknowledgments

Authors thank all who helped us through this study.

Conflict of Interest

The authors report no conflict of interest.

Funding/Financial Support

According to the authors, this article has no financial support.

References

Ansaripour, M. A., & Souhani, M. (2021). The concept of voidable contracts in English, French, Iranian law and Islamic jurisprudence. *Comparative Research on Islamic and Western Law*, *8*(2), 41-66.

Beatson, J., Burrows, A., & Cartwright, J. (2010). *Anson's Law of Contract* (29th ed.). Oxford University Press.

Cavalieri, R., & Salvatore, V. (2019). *An Introduction to International Contract Law*. G. Giappichelli Editore.

Chitty, J. (2004). *Chitty on Contracts* (29th ed.). Sweet & Maxwell.

Goode, R., Kronke, H., & McKendrick, E. (2015). *Transnational Commercial Law* (2nd ed.). Oxford University Press.

Jafari Langaroudi, M. J. (2016). *Law of Obligations* (4th ed.). Ganj-e Danesh.

Karami, Z., & Maghsoudpour, R. (2020). Comparative examination of international commercial contracts in common law and civil law legal systems. *Journal of Jurisprudence and History of Civilization*, *6*(1), 10-32.

Katouzian, N. (2017). *General Rules of Contracts* (1st ed., Vol. 2). Ganj-e Danesh.

Khoubiyari, H., & Razi, P. (2021). Critique of the status of contract voidability in Iranian law and its comparison with common law. *Comparative Research on Islamic and Western Law*, *8*(2), 99-128.

Safaei, S. H. (2008). *Civil Law: General Rules of Contracts* (6th ed., Vol. 2). Mizan.

Treitel, G. H. (2004). *An Outline of the Law of Contract* (6th ed.). Oxford University Press.

Zamaniyan, M., & Vatani, Z. (2023). Economic analysis of applying supplementary rules to prevent breach of incomplete contracts in common law, French and Iranian law, and the UNIDROIT Principles of International Commercial Contracts. *Commercial Research Quarterly*, *27*(106), 199-239.