AI-ASSISTED JUDGING: LEGAL, ETHICAL, AND COGNITIVE IMPLICATIONS OF AI NUDGES IN THE JUDICIARY

By

Aysegul Sivri

Submitted to Central European University

Department of Legal Studies

In partial fulfilment of the requirements for the degree of Master of Global Business Law and Regulation

Supervisor: Asst. Prof. Tommaso Soave

Vienna, Austria

2025

Abstract

Role of AI has been increasing recently in our judicial systems and its role in shaping legal reasoning requires immediate examination. AI-powered decision-support systems that cannot be considered as neural tools can influence judicial cognition through mechanisms such as anchoring bias, automation bias, and framing effects. This thesis investigates the extent to which these AI-generated "nudges" impact judges' reasoning and the resulting legal, ethical, and procedural implications.

Through a multidisciplinary lens that combines legal analysis, cognitive psychology, and comparative study, the research addresses how AI tools structure legal information, interact with human cognitive heuristics, and potentially alter the dynamics of judicial discretion. Special attention is paid to AI's deployment in sentencing and parole decisions, particularly risk-assessment tools like COMPAS, and to the broader normative challenges posed by opaque, non-explainable AI systems.

The thesis further examines how different legal systems, including those of the U.S., European Union, and China, regulate AI in the judiciary, and proposes best practices to ensure that AI serves to augment, rather than erode, the human-centered foundations of justice. In the end, this study argues that AI systems should be designed and regulated not merely for efficiency, but with careful attention to transparency, explainability, and judicial autonomy. The findings aim to contribute to the development of policy frameworks that preserve the integrity and legitimacy of judicial decision-making in the age of algorithmic governance.

AUTHOR'S DECLARATION

I, the undersigned, Aysegul Sivri, candidate for the LLM degree in Global Business Law and Regulation declare herewith that the present thesis titled "AI-Assisted Judging: Legal, Ethical, and Cognitive Implications of AI Nudges in the Judiciary" is exclusively my own work, based on my research and only such external information as properly credited in notes and bibliography.

I declare that no unidentified and illegitimate use was made of the work of others, and no part of the thesis infringes on any person's or institution's copyright.

I also declare that no part of the thesis has been submitted in this form to any other institution of higher education for an academic degree.

Vienna, 15.06.2025

Aysegul Sivri

Author's Note on the Use of AI Tools

In line with academic integrity principles and transparency in research practices, I would like to disclose that AI (AI)-based tools, specifically language models such as OpenAI's ChatGPT, were used to assist in the drafting, structuring, and revision of this thesis.

Such tools were employed as a support mechanism for tasks including language refinement, improving clarity, generating alternative phrasings, and summarising complex arguments — always under my full supervision and critical evaluation. The intellectual content, analysis, argumentation, and academic judgement reflected in this thesis remain entirely my own. All cited sources and references have been personally selected, verified, and incorporated by me.

The use of AI tools in this thesis was strictly limited to enhancing writing efficiency and clarity, without compromising academic rigour or the originality of the substantive contributions.

INTRODUCTION

The growing and pervasive influence of AI had already extended into judicial decision-making processes long before the conception of this thesis. AI-powered decision-support tools—ranging from legal research assistants to risk-assessment systems used in sentencing and bail decisions—are now embedded in various stages of adjudication.[1] As widely discussed in legal scholarship, it is evident that, in addition to their undeniable advantages, such as efficiency, consistency, and speed, they may also raise serious concerns based on legal reasoning, judicial decision-making, and due process.[2] This thesis begins from the premise that AI is not only an instrument of automation, but a cognitive and institutional actor that may *nudge* judicial reasoning in subtle yet normatively significant ways.[3]

Courts around the world are increasingly experimenting with AI tools. Chinese "smart courts" have been deployed to streamline case management and resolve minor disputes. In Colombia, judges have admitted to consulting large language models (LLMs), such as ChatGPT, in the decision-making process.[4] At first glance, these examples might suggest harmless efficiency gains. However, beneath this surface lies a more complex epistemic and ethical dilemma. As AI systems become more sophisticated in structuring legal data and producing recommendations, they can frame options, rank interpretations, and even suggest legal

[1] John Morison and Tomás McInerney, 'When should a computer decide? Judicial decision-making in the age of automation, algorithms and generative AI' - S Turenne and M Moussa (eds) Research Handbook on Judging and the Judiciary (2024), p. 1-6

[2] John Morison and Tomás McInerney, 'When should a computer decide? Judicial decision-making in the age of automation, algorithms and generative AI' - S Turenne and M Moussa (eds) Research Handbook on Judging and the Judiciary (2024), p. 1-5

[3] Christian Schmauder and others, 'Algorithmic Nudging: The Need for an Interdisciplinary Oversight' (2023) 42 Topoi p. 799-803.

[4] Luke Taylor, 'Colombian Judge Says He Used ChatGPT in Ruling' *The Guardian* (3 February 2023) https://www.theguardian.com/technology/2023/feb/03/colombia-judge-chatgpt-ruling accessed 13 April 2025.

outcomes, which might cause an invisible or even visible influence on judicial decision-making.[5] The very act of presenting information in a specific format or order can activate cognitive heuristics such as anchoring, automation bias, or framing effects.[6] In this light, AI is not a neutral advisor but a shaper of judgment.

The influence of AI on judicial reasoning is not merely technical but deeply legal and ethical. Delegating even part of this process to machines risks undermining the performative and deliberative aspects of judging. These concerns are raised by Morison and McInerney, who caution against reducing legal reasoning to a form of "algorithmic deduction," warning that law's social embeddedness and normative fluidity cannot be sufficiently captured by data-driven models.[7] From this perspective, the risk is not that AI will make extremely wrong decisions, but that it may subtly reshape how judges arrive at their conclusions, blurring the boundary between assistance and influence.

Three major challenges emerge from the literature on AI-assisted decision-making. First is the problem of *complementarity*: under what conditions can human-AI cooperation yield better decisions than either actor alone?[8] While some studies suggest that well-calibrated human-AI teams outperform isolated agents, others warn that humans often overtrust or underuse AI recommendations, depending on the context.[9] This leads to the second challenge: mental models. Judges, like other decision-makers, develop beliefs about AI's competence, limitations, and reliability. Inaccurate mental models may lead to mis-calibrated reliance, which can distort

[5] Caterina Fregosi and Federico Cabitza, 'A Frictional Design Approach: Towards Judicial AI and Its Possible Applications'. *HHAI-WS 2024: Workshops at the Third International Conference on Hybrid Human-AI (HHAI), June 10—14, 2024, Malm., Sweden p. 1-2*

[6] Adebola Olaborede and Lirieka Meintjes-van Der Walt, 'Cognitive Bias Affecting Decision-Making in the Legal Proce Olaborede, Adebola & Meintjes-Van Der Walt, Lirieka. (2021). Cognitive Bias Affecting Decision-Making in the Legal Process. Obiter. 41. 806-830. 10.17159/obiter.v41i4.10489. p. 2-3, 14, 27

[7] Morison and McInerney (n 1) pp. 1-2, 18-20, 24-25, 28-31.

[8] Steyvers, M., & Kumar, A. (2023). Three Challenges for AI-Assisted Decision-Making. *Perspectives on Psychological Science*, *19*(5), 722-734. https://doi.org/10.1177/17456916231181102 (Original work published 2024)

[9] Mark Steyvers and Aakriti Kumar, 'Three Challenges for AI-Assisted Decision-Making'.

judicial outcomes.[10] Third is the design problem: how should AI tools be structured to present information without cognitively overloading the judge or implicitly nudging their decisions?[11] These challenges, as Steyvers and Kumar note, require interdisciplinary responses that integrate cognitive psychology, legal theory, and human-computer interaction.[12]

In addition to these cognitive and procedural concerns, ethical questions loom large. AI tools often operate as "black boxes," making it difficult for judges to understand how outputs are generated.[13] In systems where judicial accountability and reason-giving are cornerstones of legitimacy, such opacity is especially troubling. If a judge adopts an AI recommendation without understanding its rationale, the capacity to offer public justification—a core element of legal legitimacy—may be compromised.

This thesis seeks to answer a central question: To what extent do AI-powered decision-support tools influence judges' reasoning, and what are the legal, ethical, and procedural implications of these AI nudges? In exploring this question, the thesis will examine:

- how AI systems structure legal information and frame decision options;
- the cognitive biases potentially reinforced by algorithmic design;
- the comparative regulatory responses across jurisdictions (e.g. EU, US, China);
- and the normative boundaries of legitimate judicial assistance.

Methodologically, the study adopts a multidisciplinary approach. Legal analysis will focus on case law, statutory instruments, and AI governance frameworks. Cognitive psychology will offer tools for analyzing how information framing, automation bias, and decision fatigue affect

[10] ibid 726–728.
[11] ibid 730–733.
[12] ibid.
[13] Morison and McInerney (n 1).

human judgment. Comparative insights will be drawn from common law and civil law systems, paying attention to how different legal traditions conceptualize judicial discretion.

In the end, this introduction frames AI not simply as a tool but as a potential epistemic actor in judicial decision-making—capable of shaping how judges think, not just what they decide. In doing so, it raises urgent questions about how courts can harness AI's benefits while safeguarding the human-centered foundations of justice.

The urgency of these questions is amplified by the accelerating pace of AI adoption in courts worldwide. Jurisdictions that adopt these tools without critically examining their cognitive, legal, and ethical impacts risk undermining the very legitimacy they seek to enhance. This thesis therefore aims not only to analyze current challenges but to contribute to the emerging scholarly and policy conversation on responsible AI integration in the judiciary. By foregrounding the epistemic, normative, and procedural stakes of AI-assisted judging, the study calls for a recalibration of how legal systems think about automation—not as a shortcut to justice, but as a tool whose legitimacy must be earned.

CHAPTER 1 – THE COGNITIVE AND PSYCHOLOGICAL IMPACT OF AI ON JUDGES

AI is increasingly embedded in the judicial process not just as an aid, but as a potential influencer—one that shapes how legal questions are framed and answered. This chapter investigates how AI-powered decision-support tools may nudge judicial reasoning by interacting with—and at times reinforcing—cognitive biases and mental shortcuts commonly used in human decision-making. Drawing on recent findings from cognitive psychology, behavioral law, and empirical studies of human-AI interaction, it explores how seemingly neutral tools can affect outcomes in subtle yet normatively significant ways.

1.1 The Rise of AI-Enhanced Judicial Tools

From various corners of the globe, we receive almost daily reports of AI being integrated into legal processes. While initially confined to research purposes, judges and other legal professionals have gradually succumbed to AI's undeniable appeal in streamlining tasks. AI in law has not only revolutionized how legal information is accessed—a departure from the era of getting lost in voluminous law books—but also transformed how this information is processed and utilized. Contemporary AI applications in the judiciary span a wide spectrum—from foundational legal research tools to advanced systems offering predictive analytics, risk assessment, and even automated drafting capabilities. These tools go far beyond simple data retrieval; they structure, prioritize, and contextualize information to support judges in navigating increasingly complex caseloads and procedural requirements. However, this assistance is not without its cognitive and procedural implications. The way AI platforms organize content—through visual cues, rankings, confidence scores, and default options—can subtly shape judicial reasoning by influencing which paths appear more salient or preferable. Moreover, as these tools become seamlessly integrated into daily judicial workflows, their

framing effects risk becoming embedded in routine decision-making, raising important questions about transparency, autonomy, and the preservation of judicial discretion. To better understand how these cognitive and procedural dynamics have evolved, it is helpful to trace the historical trajectory of AI applications in the judiciary—from early legal databases to today's more advanced predictive and decision-support systems.

The integration of AI into judicial practice has not occurred overnight; it reflects an incremental evolution shaped by both technological progress and increasing pressures on judicial systems worldwide. Initially, AI's role was largely confined to enhancing legal research through databases such as Westlaw and Lexis.[14] These tools revolutionized access to statutes, case law, and secondary materials, streamlining the research process and allowing judges to navigate vast bodies of legal information more efficiently. However, they operated as neutral aids, leaving the core reasoning and decision-making processes firmly in human hands.

A more consequential shift began with the introduction of predictive analytics and risk assessment tools, which moved AI from the periphery of judicial work into its very heart. Systems such as COMPAS started to generate probabilistic evaluations about defendants' likelihood of recidivism or flight risk, thereby directly informing high-stakes decisions on bail, sentencing, and parole. This phase involved integrating algorithmic outputs into judicial workflows, raising significant concerns about transparency, fairness, and the potential for cognitive biases to influence outcomes.[15]

[14] Westlaw and Lexis are two of the most widely used legal research platforms in the world. They provide access to extensive databases of case law, statutes, regulations, legal commentary, and secondary sources. These tools help legal professionals quickly find relevant legal materials and analyze them through features like advanced search, citation tracking, and editorial summaries. While both serve similar functions, they differ slightly in interface, search capabilities, and editorial content.

[15] Carolyn McKay, Predicting risk in criminal procedure: actuarial tools, algorithms, AI and judicial decision-making, Current Issues in Criminal Justice (2020), 32:1, 22-39.

The current frontier is increasingly transformative, as jurisdictions such as Estonia and Argentina experiment with robot-judge projects and automated decision-making tools, pushing the boundaries of AI's role in adjudication.[16] In these settings, AI systems are not merely supporting human decision-makers; they are beginning to perform quasi-adjudicative and administrative functions themselves. This represents a conceptual shift from predictive justice—where algorithms advise human judges—to a new paradigm in which AI may increasingly participate in or even replace certain aspects of human judgment.[17]

Taken together, this historical trajectory illustrates how AI's role has evolved from facilitating legal research to actively shaping, and in some cases generating, judicial outcomes. This evolution carries important cognitive and normative implications, which will be examined in the sections that follow. As AI systems assume more active roles in judicial decision-making, their cognitive impact on human judges becomes an increasingly significant concern.

1.2 Cognitive Heuristic and Nudging Mechanisms

As AI-generated outputs are increasingly integrated into judicial workflows, they shape not only what information judges receive but also how and when they encounter it. Judicial decision-making is not immune to well-known cognitive heuristics, especially under conditions of time pressure and information overload.[18] AI tools often interact with these heuristics in ways that exacerbate their influence.[19] Three cognitive effects are particularly relevant:

[16] Francesca Ceresa Gastaldo, 'The Automaton-Judge: Some Reflections on the Future of AI in Judicial Systems' (2024) 14 Sortuz. Oñati Journal of Emergent Socio-legal Studies 399-408.
[17] Tania Sourdin, 'Judge v Robot? AI and Judicial Decision-Making' (2018) 41 University of New South Wales Law Journal <https://www.unswlawjournal.unsw.edu.au/article/judge-v-robot-artificial-intelligence-and-judicial-decision-making/> accessed 6 June 2025.
[18] Regina de Brito Duarte and Joana Campos, 'Looking For Cognitive Bias In AI-Assisted Decision-Making' INESC-ID, Instituto Superior Técnico, Lisbon, Portugal, 2024 p.1-3.
[19] ibid.

- **Anchoring Bias:** Judges may disproportionately rely on the first piece of information presented, such as a risk score or sentencing recommendation, when forming their final judgment.[20] Even if judges consciously try to counterbalance AI inputs, unconscious anchoring effects may still skew outcomes. Anchoring bias can arise in AI-assisted decision-making when AI-generated outputs serve as anchors, influencing human judgments.[21]
- **Automation Bias:** This refers to the tendency to over-trust algorithmic outputs—a phenomenon known as automation bias—which is particularly pronounced when the human decision-maker faces high cognitive load or lacks domain-specific expertise. Under such conditions, decision-makers may defer excessively to AI outputs, presuming their statistical superiority.[22]
- **Framing Effect:** The way AI systems present information can significantly influence human decision-making outcomes, depending on how options are presented. In AI-assisted contexts, variations in framing, such as presenting outcomes in positive or negative terms, have been shown to shape users' perceptions and choices.[23]

These effects are not isolated but often reinforce one another in subtle ways. While Duarte and Campos studied automation bias and anchoring bias as distinct cognitive phenomena, their findings suggest that AI-generated outputs can simultaneously serve as both trusted sources of information and potent cognitive anchors. In contexts where decision-makers place high trust in AI outputs, this trust may increase the likelihood that initial AI-provided values or recommendations disproportionately shape subsequent judgments. Although the interaction

[20] B. Englich, T. Mussweiler, F. Strack, Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making, Personality and Social Psychology Bulletin 32 (2006) 188–200.
[21] Birte Englich, Thomas Mussweiler and Fritz Strack, 'Playing Dice With Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making' (2006) 32 Personality and Social Psychology Bulletin 188.
[22] Duarte and Campos (n 18) 5.
[23] ibid 4.

between these biases was not directly tested, the experimental results underscore how both mechanisms can operate in tandem when individuals engage with algorithmic decision-support systems.[24]

Beyond unintended amplification, some AI systems are intentionally designed to use human thinking patterns. AI systems do not simply reflect cognitive biases; they can be designed to exploit them. As Schmauder et al. argue, algorithmic nudging uses predictable patterns in human cognition to influence behavior, often without the individual being aware of the manipulation.[25] While such nudges may appear innocuous in consumer and online contexts, AI systems increasingly exploit predictable patterns in human cognition to influence behavior, often without users' awareness, through subtle design choices in interfaces, such as how information is presented or defaults are configured. Understanding these mechanisms requires an interdisciplinary approach that combines cognitive science and AI research.[26]

1.3 Cognitive Fatigue, Deference, and Decision-Making Under Pressure

The impact of cognitive biases and algorithmic nudges is not uniform across all decision contexts. Duarte and Campos demonstrate that such biases are more likely to arise under high cognitive load.[27] In judicial settings, high cognitive load can often manifest as time pressure and emotional stress, making judges particularly susceptible to these influences. In modern judicial environments, judges increasingly operate under conditions of significant workload pressures and limited time for deliberation. Adjudicative decision-making is susceptible to a range of contextual factors that can undermine the quality of judicial reasoning, including

[24] ibid 4–5.
[25] Christian Schmauder and others, 'Algorithmic Nudging: The Need for an Interdisciplinary Oversight' (2023) 42 Topoi 799, 799–803.
[26] ibid 800.
[27] Duarte and Campos (n 18).

decision fatigue, emotional influences, and the cumulative effects of high-volume caseloads.[28] In this context, AI-based decision-support tools are frequently adopted to assist judges in managing their workload and streamlining certain aspects of decision-making.[29] While such systems offer pragmatic advantages in overloaded court systems, their increasing use as coping mechanisms may raise concerns regarding the potential erosion of judicial autonomy and the risk of undue reliance on algorithmic outputs. The susceptibility to such influences is further shaped by the cognitive demands placed on judicial decision-makers. Duarte and Campos demonstrate that cognitive biases, such as automation bias, are particularly likely to arise under conditions of high cognitive load or when the decision-maker lacks sufficient domain-specific expertise.[30] In such contexts, reliance on algorithmic outputs may increase, as individuals tend to defer to AI recommendations when faced with complex information processing demands. While Duarte and Campos do not specifically study judicial decision-making, their findings highlight cognitive mechanisms that can plausibly affect judges operating under similar pressures in overloaded court systems.

AI outputs that present predetermined answers may risk constraining human interpretive flexibility. Overreliance on AI can suppress individual creativity and intuition in problem-solving and decision-making contexts.[31] Moreover, the 'black box effect'—where AI's formal, opaque presentation style can obscure underlying limitations and biases—further complicates the user's ability to critically engage with AI recommendations.[32] In judicial contexts, similar dynamics may plausibly arise when judges interact with AI-generated recommendations

[28] Tania Sourdin, 'Judge v Robot? AI and Judicial Decision-Making' (2018) 41 University of New South Wales Law Journal 1129 <https://www.unswlawjournal.unsw.edu.au/article/judge-v-robot-artificial-intelligence-and-judicial-decision-making/> accessed 13 April 2025.
[29] ibid 1125–1126; 1131.
[30] Regina de Brito Duarte and Joana Campos, 'Looking For Cognitive Bias In AI-Assisted Decision-Making' INESC-ID, Instituto Superior Técnico, Lisbon, Portugal, 2024 p. 10.
[31] Al-Zahrani Abdulrahman M, 'Balancing Act: Exploring the Interplay Between Human Judgment and AI in Problem-Solving, Creativity, and Decision-Making' (2024) 2 IgMin Research 145, 148.
[32] ibid 153.

presented as authoritative or statistically optimal. Especially under conditions of cognitive load or procedural time constraints, the formal and data-driven presentation of such outputs may inadvertently narrow interpretive flexibility, thereby influencing judicial reasoning in subtle yet consequential ways.

The effectiveness of AI-assisted decision-making depends not only on the quality of AI outputs, but also on the decision-maker's mental model of the system—how accurately they understand its strengths, limitations, and when to defer to it.[33] Without a well-calibrated understanding, users may either over-rely on AI recommendations or discount valuable insights, leading to suboptimal outcomes.[34] Carter and Liu similarly emphasize that inadequate mental models can exacerbate anchoring effects and distort the interaction between human judgment and algorithmic advice.[35] In judicial contexts, where cognitive load and procedural pressures are prevalent, such miscalibrated trust can inadvertently shift the balance from AI as an advisory aid toward AI shaping core elements of judicial reasoning. Designing AI systems that support the development of accurate mental models—by communicating uncertainty and clarifying the scope of algorithmic competence—is therefore essential to preserving both accuracy and judicial autonomy.

1.4 Case Study: AI Nudging in Sentencing and Parole Decisions

The integration of AI into the criminal justice system has become increasingly visible in recent years, particularly in the areas of sentencing and parole, where decisions carry profound implications for individual liberty and social justice.[36] Among the most prominent applications

[33] Mark Steyvers and Aakriti Kumar, 'Three Challenges for AI-Assisted Decision-Making' 724–726.
[34] ibid 726.
[35] Lemuria Carter and Dapeng Liu, 'How Was My Performance? Exploring the Role of Anchoring Bias in AI-Assisted Decision Making' (2025) 82 International Journal of Information Management 102875.
[36] Carolyn McKay, 'Predicting Risk in Criminal Procedure: Actuarial Tools, Algorithms, AI and Judicial Decision-Making' (2020) 32 Current Issues in Criminal Justice 22, 22–23.

of AI in this context are risk-assessment tools such as COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), a proprietary system that purports to predict a defendant's risk of recidivism and informs judicial decisions concerning pre-trial detention, sentencing severity, and parole eligibility.[37] In practical terms, COMPAS outputs—typically presented as "low," "medium," or "high" risk scores—are regularly consulted by judges when setting sentences or parole conditions.[38]

While algorithmic tools such as COMPAS are often introduced to promote consistency, objectivity, and efficiency in sentencing, the Wisconsin Supreme Court's decision in *State v. Loomis* highlights significant normative and legal concerns. The Court ultimately allowed the use of COMPAS at sentencing but explicitly warned that its use must be "circumscribed" and accompanied by disclaimers outlining its limitations.[39] These warnings stressed that COMPAS relies on group data rather than individualized assessments, that its proprietary nature precludes transparency, and that "studies have raised questions about whether [it] disproportionately classify minority offenders as having a higher risk of recidivism".[40] Despite these caveats, the sentencing judge in Loomis stated: "You're identified, through the COMPAS assessment, as an individual who is at high risk to the community," using that classification to justify a lengthy custodial sentence.[41] This tension between formal judicial caution and practical reliance illustrates how algorithmic outputs may not only shape sentencing outcomes, but also subtly influence judicial reasoning, raising deeper questions about fairness, transparency, and the preservation of human discretion in criminal justice.

[37] ibid 27–28; Utsav Bahl and others, 'Algorithms in Judges' Hands: Incarceration and Inequity in Broward County, Florida' (SocArXiv, 1 May 2023) 246–247 <https://osf.io/326tw> accessed 8 June 2025.
[38] Bahl and others (n 37) 253; McKay (n 36) 27.
[39] Eric L Loomis, 'In the Supreme Court of the United States'.
[40] ibid.
[41] ibid.

As multiple studies demonstrate, the framing of AI-generated risk categories can systematically shape judicial perceptions and decisions[42], raising critical questions about how such tools interact with human cognitive biases and potentially alter the nature of judicial reasoning itself.

Empirical studies underscore these concerns. Users perceive algorithmic outputs as more accurate and less biased than equivalent human suggestions, even when both offer identical content.[43] Real-world analyses similarly reveal this dynamic: in Broward County, Florida, following the introduction of COMPAS, judges deferred incarceration more frequently for defendants labeled "low risk" and imposed harsher outcomes for those labeled "high risk" compared to pre-COMPAS patterns, even though underlying racial disparities in risk classification persisted.[44]

Duarte and Campos documented that participants systematically over-relied on AI recommendations in general decision-making tasks, particularly under conditions of high cognitive load or limited expertise.[45] These findings highlight the broader difficulty of calibrating trust in AI systems: human decision-makers may either ignore or excessively defer to AI outputs, depending on task complexity, cognitive load, and familiarity with algorithmic systems.[46] The *State v. Loomis* ruling itself implicitly recognized this risk, as the Wisconsin Supreme Court mandated that COMPAS reports must include explicit disclaimers to reduce over-reliance and ensure informed judicial use.[47] Building on this, the risk is compounded by the presentation formats of AI outputs, which can subtly make cognitive biases stronger and

[42] Christian Schmauder and others, 'Algorithmic Nudging: The Need for an Interdisciplinary Oversight' (2023) 42 Topoi 803.
[43] ibid 800–801; Carter and Liu 2–3; Duarte and Campos 5–6.
[44] Bahl and others (n 37).
[45] Duarte and Campos (n 18) 5–6.
[46] ibid.
[47] McKay (n 36); Lauren Kirchner, 'Wisconsin Court: Warning Labels Are Needed for Scores Rating Defendants' Risk of Future Crime'.

nudge judicial outcomes in ways that may not align with normative expectations of deliberative justice.[48]

The convergence of anchoring, automation bias, and framing effects suggests that the design of AI tools used in judicial contexts must be critically examined. As Schmauder et al. highlight, algorithmic nudging can exploit human cognitive biases in subtle and often opaque ways, necessitating careful attention to system design and the transparency of AI-human interaction.[49] Transparency alone is insufficient. Recent analyses of AI risks in judicial settings emphasize the importance of promoting judicial awareness and providing targeted training to reduce cognitive vulnerabilities and help judges think critically about AI outputs.[50] Without such measures, AI systems risk displacing rather than augmenting human judicial reasoning.

In the end, this chapter affirms the central premise of this thesis: AI systems are not neutral tools but epistemic actors that structure the space of legal reasoning.[51] If left unexamined, they risk transforming judicial decision-making from a human-centered, deliberative process into a mode of automated inference—one that subtly undermines judicial autonomy and transparency.[52]

A growing concern in AI-assisted adjudication is the risk of emotional distancing. As Sourdin warns, excessive reliance on digital tools may encourage judges to view litigants as data points rather than as individuals with complex social and moral realities.[53] This abstraction effect is magnified when AI interfaces prioritize statistical outputs over narrative nuance, reducing rich

[48] Schmauder and others (n 25) 800, 803–804.
[49] Schmauder and others (n 25).
[50] RaisuL Sourav, 'Relying on AI in Judicial Decision-Making: Justice or Jeopardy?'; Doron Teichman and Eyal Zamir, 'Judicial Decision-Making: A Behavioral Perspective' in Eyal Zamir and Doron Teichman (eds), Doron Teichman and Eyal Zamir, *The Oxford Handbook of Behavioral Economics and the Law* (Oxford University Press 2014) <https://academic.oup.com/edited-volume/34475/chapter/292525706> accessed 14 April 2025.
[51] Schmauder and others (n 25).
[52] Teichman and Zamir (n 50); Sourav (n 50).
[53] Sourdin (n 17).

human contexts to algorithmic risk scores.[54] In criminal justice settings, such distancing may blunt empathy and moral engagement, thereby undermining the expressive function of sentencing.[55] While judicial impartiality is essential, legal reasoning is not devoid of human values. The use of AI systems that suppress narrative detail or decontextualize facts risks eroding the judge's ability to respond to individual circumstances with appropriate compassion, skepticism, and moral discernment.[56] This further underscores the pressing need for human-centered AI design—one that not only respects but actively preserves the deliberative, narrative, and moral dimensions of judging.[57]

The analysis in this chapter has demonstrated that AI-assisted judicial tools do not simply enhance efficiency; they fundamentally reshape the cognitive architecture of judicial reasoning. By embedding biases such as anchoring, automation, and framing effects into the decision-making process, these tools risk converting subtle cognitive nudges into structural distortions of justice. Moreover, the risk of emotional distancing—where judges may unconsciously adopt a detached, data-driven view of litigants—further highlights the normative tensions AI introduces into domains that demand human empathy, discretion, and deliberative judgment.

While current design interventions and judicial training offer some mitigation, they remain incomplete in addressing the deeper epistemic transformation AI tools are affecting in judicial practice. The reality that AI systems act as epistemic actors underscores the urgent need for legal frameworks that can safeguard judicial autonomy, transparency, and legitimacy. Against this backdrop, the next chapter will examine the legal and ethical challenges that arise from the

[54] Sourav (n 50).
[55] ibid.
[56] ibid; Sourdin (n 28).
[57] Schmauder and others (n 25); Sourav (n 50).

integration of AI nudging mechanisms into judicial contexts—exploring how regulatory principles must evolve to confront these complex and emerging risks.

CHAPTER 2: LEGAL AND ETHICAL CHALLENGES OF AI NUDGING

2.1 Normative Concerns: Transparency, Accountability, and Legitimacy

Building on Chapter 1's demonstration that AI systems act as epistemic actors that structure judicial cognition through subtle nudging effects, this chapter explores how such cognitive dynamics translate into broader legal, ethical, and institutional risks for the judiciary. The integration of AI systems into judicial decision-making raises fundamental normative concerns regarding transparency, accountability, and legitimacy. These concerns echo the cognitive risks discussed in Chapter 1 but extend them to the institutional and public dimensions of justice.

Transparency is essential for safeguarding procedural fairness and enabling contestability. Without the ability to understand and interrogate the reasoning behind AI-generated outputs, both judges and affected parties are deprived of key legal safeguards. According to the *EU Ethics Guidelines for Trustworthy AI*, the procedural dimension of fairness requires that "the decision-making processes should be explicable" in order to enable effective contestability and accountability.[58] Without such explicability, both judges and affected parties are deprived of legal safeguards, and accountability risks becoming merely formal.

This challenge is not theoretical. Proprietary risk assessment tools such as COMPAS, widely used in US criminal justice, introduce serious due process risks precisely because their internal workings remain opaque.[59] The *Wisconsin v. Loomis* decision exemplifies this danger: while permitting the use of COMPAS scores, the court required that judges be explicitly warned of their limitations, including lack of transparency and potential bias.[60]

[58] 'Ethics Guidelines For Trustworthy AI'. European Comission, 2019
[59] McKay (n 15).
[60] Kirchner (n 47).

Public trust in AI-assisted justice is equally crucial. Empirical research by Watamura et al., conducted in Japan's lay judge system, indicates that participants' trust and perceived fairness of AI-assisted decisions can be influenced by perceptions of transparency and the system's ability to reflect human-like reasoning, particularly in cases involving mitigating circumstances.[61]

Finally, judicial legitimacy rests on persuasion: decisions must provide reasons that resonate with human audiences, not merely technically correct outputs.[62] If AI systems fail to meet this standard, their deployment may inadvertently erode public confidence in the justice system.

Ensuring transparent, accountable, and publicly comprehensible AI interventions is therefore not a technical luxury but a normative imperative for the rule of law.

2.2 The Black Box Problem and Due Process

One of the most significant threats posed by AI-assisted judicial tools concerns the "black box" nature of many such systems and their potential to undermine due process protections. While the promise of AI lies in improving efficiency and consistency, opaque decision-making processes jeopardize fundamental legal rights.

Due process requires that individuals understand the basis of decisions that affect them and have a meaningful opportunity to challenge those decisions. Yet many AI tools employed in the justice system, particularly proprietary risk-assessment algorithms, fail this basic test. Tools like COMPAS embed non-transparent reasoning into judicial determinations, preventing both defendants and their counsel from interrogating or contesting the underlying logic. [63]

[61] Eiichiro Watamura, Yichen Liu and Tomohiro Ioku, 'Judges versus AI in Juror Decision-Making in Criminal Trials: Evidence from Two Pre-Registered Experiments' (2025) 20 PLOS ONE e0318486.
[62] Eugene Volokh, 'CHIEF JUSTICE ROBOTS'.
[63] McKay (n 36) 32–33.

The *Wisconsin v. Loomis* case makes this risk explicit: the court permitted the use of COMPAS scores only with explicit warnings about their limitations and lack of transparency.[64]

The problem extends beyond technical legal standards to broader concerns about legitimacy. It is not enough for AI to produce formally correct outcomes; judicial decisions must also be intelligible and persuasive to human audiences. Black-box systems, by contrast, risk eroding public confidence in the fairness of AI-assisted justice.[65]

In addition, the black box problem has been identified as a "major concern" in recent empirical literatüre,[66] where opacity is seen as undermining accountability and trust. If litigants and the public cannot comprehend how AI influences judicial outcomes, the very legitimacy of such systems comes into question.

Overall, addressing the black box problem is not just about enhancing system transparency; it is about preserving the very foundations of due process and public trust in the judiciary. If AI-assisted tools cannot be explained and contested, they risk transforming adjudication into an opaque process that undermines both accountability and legitimacy.

2.3 Fairness, Bias, and Individualization in Criminal Justice

Fairness, bias, and individualization remain central concerns in the deployment of AI-assisted judicial tools. Beyond technical accuracy, the perceived fairness of legal outcomes—particularly procedural transparency and the system's capacity to incorporate human-like moral and emotional reasoning—plays a crucial role in shaping public trust. Empirical findings indicate that AI is often regarded as lacking this capacity, which can undermine confidence in

[64] Kirchner (n 47).
[65] Volokh (n 62) 1138–1139.
[66] Watamura, Liu and Ioku (n 61) p.1.

its judgments, especially in contexts requiring detailed understanding of mitigating circumstances.[67]

The opacity and impersonality of algorithmic decisions pose significant challenges to judicial legitimacy. As McKay explains, proprietary risk tools may obscure how risk scores are calculated, making them unknowable even to courts and defendants. Similarly, Pfeiffer et al. discuss how algorithmic models like COMPAS lack the capacity to adapt to individual circumstances and often reinforce existing biases, thereby undermining procedural fairness and public trust.[68]

Ensuring that AI systems offer explainable and contestable outputs is vital to uphold procedural fairness. The EU's Ethics Guidelines for Trustworthy AI, explicitly link explainability and the right to contest AI decisions to fundamental fairness principles.[69] In judicial contexts, such transparency is essential to maintaining public confidence and perceived legitimacy.

2.4 The Promise and Pitfalls of Predictive Consistency

AI-based sentencing tools are often introduced with the promise of enhancing consistency, objectivity, and efficiency in judicial decision-making. A frequently cited benefit is their potential to reduce inter-judge disparity and mitigate the influence of subconscious biases.[70] Studies suggest that algorithmic risk assessments may offer more consistent, timely, and transparent outputs than subjective human judgments.[71] However, this promise must be carefully weighed against emerging concerns about fairness, transparency, and the preservation of judicial discretion. AI-based sentencing tools carry a significant risk of reinforcing existing

[67] ibid 3–4, 10.
[68] Jella Pfeiffer and others, 'Algorithmic Fairness in AI: An Interdisciplinary View' (2023) 65 Business & Information Systems Engineering 209, 1, 9–10; McKay (n 15) 32–35.
[69] Pfeiffer and others (n 68) 7.
[70] Eugene Volokh, 'CHIEF JUSTICE ROBOTS'; McKay (n 36).
[71] McKay (n 36); Watamura, Liu and Ioku (n 61).

societal biases. The ProPublica investigation into the COMPAS risk-assessment system demonstrated that Black defendants were substantially more likely than White defendants to be incorrectly classified as high risk, a disparity with serious consequences for sentencing and parole decisions.[72] More broadly, machine learning systems that are trained on historically biased data can systematically reproduce and amplify such biases in future decisions.[73]

This raises fundamental normative concerns about the use of AI in judicial contexts. Drawing a crucial distinction, Morison and McInerney emphasise that while standardisation may be acceptable in administrative decisions, judicial decisions—such as sentencing and parole—require human interpretation, empathy, and legitimacy grounded in democratic values.[74] As McKay observes, replacing such deeply discretionary functions with opaque, data-driven tools risks undermining both procedural fairness and the expressive function of the law.[75] Moreover, lack of transparency and perceived depersonalisation in algorithmic judgments may weaken trust in judicial institutions.[76]

To navigate these tensions, policymakers and courts must proceed with caution. This includes mandating transparency and ensuring explainability—essential procedural safeguards to prevent opacity and enhance accountability in AI-assisted judicial decision-making.[77] Additionally, it is crucial to define appropriate levels of human oversight and to preserve the autonomy and discretion of judicial actors, particularly in contexts such as sentencing and

[72] Pfeiffer and others (n 68) 1.
[73] Jella Pfeiffer and others, 'Algorithmic Fairness in AI: An Interdisciplinary View' (2023) 65 Business & Information Systems Engineering 209, 1.
[74] Morison and McInerney (n 1) 8–10.
[75] McKay (n 36) 22–25.
[76] Pfeiffer and others (n 68) 1–2.
[77] ibid 7; McKay (n 36) 25–26.

parole where interpretive judgment and moral reasoning are indispensable.[78] Without such measures, the integration of AI risks sacrificing justice in the pursuit of efficiency.[79]

These concerns are particularly acute in the judicial domain, where the legitimacy of decisions relies not only on outcomes but also on the visibility of reasoning and the perceived fairness of the process.

2.5 Comparative Legal Perspectives

Approaches to the integration of AI into judicial systems vary significantly across jurisdictions, reflecting divergent legal cultures, regulatory philosophies, and institutional priorities. In the United States, AI-assisted tools such as risk-assessment algorithms have been deployed in a largely fragmented and ad hoc manner. This piecemeal approach contrasts sharply with the more structured regulatory frameworks emerging in other regions. The European Union's AI Act establishes a harmonised legal framework for high-risk AI systems, including those used in judicial and law enforcement contexts, with an emphasis on transparency, accountability, and respect for fundamental rights.[80] In China, by contrast, the development of 'Smart Courts' reflects a state-led, centralised strategy to integrate AI into judicial decision-making, aiming to enhance efficiency, consistency, and public trust in the courts.[81]

The European Union has adopted a notably cautious stance. Under the AI Act, AI systems used in certain high-risk contexts — including judicial and law enforcement applications — are subjected to stringent requirements regarding transparency, accountability, and human

[78] Morison and McInerney (n 1) 2-3,8-10; McKay (n 15) 23–25.
[79] Pfeiffer and others (n 68) 1–2; McKay (n 15) 25–26; Morison and McInerney (n 1) 3–10.
[80] 'Regulation - EU - 20241689 - EN - EUR-Lex' Recitals 1, 6–8.
[81] Benjamin Minhao Chen and Zhiyu Li, 'How Will Technology Change The Face of Chinese Justice?' [2020] Columbia Journal of Asian Law 1–4.

oversight.[82] This reflects a broader European commitment to preserving fundamental rights and ensuring that AI integration aligns with democratic values and the rule of law.[83]

By contrast, China's approach to judicial AI has been marked by the rapid and ambitious deployment of technologies within its *smart court* system. This rapid uptake of AI is further enabled by the judiciary's institutional positioning within the political structure of the Party-state, where courts are not fully insulated from external influence and judicial independence remains contested.[84] Courts increasingly use AI not only to assist judges but to partially automate decision-making processes, particularly in high-volume, low-value cases such as online trade disputes and minor civil claims.[85] While this model enhances efficiency and promotes consistency, it also raises fundamental questions about the capacity of such systems to uphold individualized justice and maintain public legitimacy.[86] As Sourdin highlights, legal culture plays a crucial role in shaping how judicial AI is received: transplanting highly automated models into jurisdictions with strong traditions of judicial discretion and public reason-giving—such as common law systems—may provoke resistance and produce unintended consequences.[87]

These comparisons show that using AI in courts is not just about technology. It raises important questions about justice, fairness, and how legal decisions should be made. Each country needs to design its rules carefully, so that AI supports—not replaces—its own legal traditions and democratic values.

[82] 'Regulation - EU - 20241689 - EN - EUR-Lex' (n 80) Recital 44, Art. 9–14.
[83] Pfeiffer and others (n 68) 13.
[84] Macquarie Law School, Macquarie University and others, 'THE USE OF AI IN JUDICIAL DECISIONMAKING: THE EXAMPLE OF CHINA' (2023) 2022 International Journal of Law, Ethics, and Technology 1, 4–5 2008.
[85] Simon Chesterman, 'All Rise for the Honourable Robot Judge? Using AI to Regulate AI' [2022] SSRN Electronic Journal 2–3 <https://www.ssrn.com/abstract=4252778> accessed 13 April 2025.
[86] ibid 1.
[87] Sourdin (n 17) 1115–1116.

2.6 Common Law vs. Civil Law Approaches to Judicial Discretion

The integration of AI into judicial processes interacts differently with the traditions and expectations of common law and civil law systems, particularly regarding judicial discretion. In civil law jurisdictions, where legal reasoning tends to be more codified and constrained by statutory interpretation, there may be greater institutional openness to AI tools that promote consistency and efficiency. In such contexts, judicial discretion is narrower, and the ideal of standardized decision-making is often viewed as compatible with systemic legitimacy.[88]

Conversely, common law systems place a stronger emphasis on individualized reasoning and the expressive dimensions of judicial decision-making. Judicial discretion is understood not merely as a technical gap-filler but as an essential component of the law's adaptability and moral authority. Transplanting AI models optimized for efficiency into common law contexts risks eroding these deliberative and narrative aspects of judging.[89] In addition, empirical studies indicate that public trust in AI-assisted judgments is closely linked to expectations of transparency and fairness in judicial processes,[90] although how these expectations vary across legal traditions remains an open question.

AI tools that encourage automation bias may further complicate this dynamic. Proprietary risk-assessment tools like COMPAS can subtly narrow judicial discretion by framing decisions around algorithmically generated risk profiles.[91] Similarly, while AI may enhance performance in administrative domains, its application in discretionary judicial contexts must be approached with care to preserve the deliberative integrity of the courts.[92]

[88] Gastaldo (n 16) 402–404.
[89] Sourdin (n 17) 1115–1116.
[90] Watamura, Liu and Ioku (n 61) 3–4.
[91] McKay (n 15) 22–25.
[92] Morison and McInerney (n 1) 2–3, 8–10.

In the end, the use of AI in the judiciary depends heavily on the legal culture of each system. To protect the courts' legitimacy, AI tools must work alongside, rather than replace, judges' ability to make thoughtful and independent decisions.

CHAPTER 3 — REGULATORY AND POLICY RESPONSES TO AI IN THE JUDICIARY

3.1 Overview of Existing Regulatory Frameworks

Given that AI systems increasingly act as epistemic actors that influence judicial reasoning, as demonstrated in previous chapters, regulatory responses must be evaluated not only for their capacity to ensure transparency and oversight, but also for their ability to safeguard the cognitive autonomy of human judges. Regulatory responses to AI-assisted judicial tools vary considerably across jurisdictions, reflecting distinct legal cultures, governance models, and institutional capacities. Crucially, effective regulation extends beyond the mere adoption of high-level legal instruments; it also depends on the presence of practical mechanisms—such as oversight structures, training protocols, and transparency guarantees—that shape how AI is implemented in daily judicial practice.

The European Union has thus far adopted one of the most structured regulatory frameworks for AI. Under the proposed AI Act, AI systems used in judicial contexts fall within the category of "high-risk" systems and are therefore subject to stringent obligations regarding transparency, human oversight, and accountability.[93] Complementing such formal frameworks, expert reports on judicial AI use in France have underscored important operational safeguards. For instance, a 2021 report from the Fondation Jean Jaurès stresses that judges must fully understand how algorithmic decision-support tools interpret facts, lest they risk being bound by recommendations they cannot substantiate in law or in fact[94] justice. This guidance highlights the need to preserve legal contestability and to avoid opaque algorithmic influences on judicial reasoning.

[93] Pfeiffer and others (n 73) 210.
[94] Basdevant, A., Jean, A. & Storchan, V., Mécanisme d'une justice algorithmisée, Fondation Jean Jaurès, Paris, 2021.

In contrast, the United States has adopted a fragmented, state-level approach. While federal regulatory coherence remains lacking, soft law instruments have begun to emerge. The National Center for State Courts (NCSC) has issued voluntary ethical guidelines advocating for transparency, auditability, and informed human oversight in the use of AI by courts.[95] Nevertheless, the reliance on proprietary systems—such as COMPAS—has raised significant accountability concerns, as the opaque and trade-secret-protected nature of such algorithms prevents meaningful scrutiny by judges, defendants, and the public.[96]

China's smart court initiatives represent a distinctive regulatory model—one characterised by rapid technological deployment under centrally coordinated state control. Since 2014, China has introduced a series of smart courts aimed at transforming the judiciary through AI, with robot judges, automatic reason-generation frameworks, and national e-evidence platforms now operating in courts such as the Suzhou Intermediate Court, Beijing Internet Court, and Hangzhou Internet Court[97] At the same time, a 2023 UNESCO global survey reports that 44% of judicial professionals worldwide are integrating AI tools into their daily tasks, yet 91% have received no formal training or guidelines on their responsible use.[98] This global mismatch between AI uptake and institutional readiness raises concerns that are particularly salient in China's rapidly evolving smart court ecosystem, where judicial autonomy and algorithmic influence remain critical areas of debate. Such concerns are particularly acute in China's context due to its fast-paced implementation of AI tools without corresponding institutional readiness.

At the transnational level, cross-border governance of AI is still in its infancy. A recent report by the UN Chief Executives Board highlights the need for globally coherent regulatory

[95] 'AI Foundations in the Courts | National Center for State Courts' <https://www.ncsc.org/resources-courts/ai-foundations-courts> accessed 10 June 2025.
[96] McKay (n 36) 32.
[97] Macquarie Law School, Macquarie University and others (n 84) 4–5.
[98] Sourav (n 53) UNESCO, 2023, cited in Sourav, 2024, p. 3.

frameworks, pointing out that current state and regional approaches lack harmonization and are often siloed.[99] Moreover, some scholars argue for the establishment of a new international AI regulatory mechanism—one that could develop core interoperable standards and facilitate multi-jurisdictional cooperation.[100] Meanwhile, algorithmic nudging—subtle design choices that shape human behavior—continues to pose specific regulatory challenges unsupported by mere disclosure obligations.[101] Similarly, as Bolander cautions, true explainability demands interactive interfaces that allow meaningful user interrogation—not just superficial transparency statements.[102]

In the end, safeguarding judicial integrity in the face of algorithmic governance will require practical oversight mechanisms—including regular security audits, transparent public testing, and redundancy through multiple AI designs—to reduce risks such as hacking, hidden biases, and systemic vulnerabilities.[103] Without such safeguards, regulatory frameworks risk becoming largely symbolic, offering insufficient substantive protection against potential harms.[104] In light of this, future regulatory efforts should not only establish binding legal standards but also incorporate robust operational safeguards to ensure that AI systems support—rather than erode—the human-centred foundations of judicial decision-making.[105]

[99] United Nations Chief Executives Board for Coordination, United Nations System White Paper on AI Governance (May2024) https://unsceb.org/sites/default/files/202405/UN%20System%20White%20Paper%20AI%20Governance.pdfaccessed 10 June 2025.
[100] Olivia J Erdélyi and Judy Goldsmith, 'Regulating AI: Proposal for a Global Solution' (arXiv, 22 May 2020) 5–8 <http://arxiv.org/abs/2005.11072> accessed 10 June 2025.
[101] Schmauder and others (n 25) 801.
[102] Thomas Bolander, 'What Do We Loose When Machines Take the Decisions?' (2019) 23 Journal of Management and Governance 849, 851.
[103] Volokh (n 62) 1171–1176.
[104] ibid.
[105] ibid 1175–1176.

3.2 Policy Proposals in Literature and Practice

The integration of AI tools into judicial decision-making has spurred a growing body of policy proposals aimed at balancing innovation with legal and ethical safeguards. While the literature rightly emphasises best practices in adaptive governance, this section argues that truly effective AI governance in the judiciary must go beyond flexibility and transparency—it must ensure genuine contestability, reduce subtle cognitive and emotional risks, and establish robust institutional oversight mechanisms.

Adaptive governance models have been widely recommended as an essential component of responsible AI regulation. The Nuffield Foundation, for instance, underscores the need for regulatory frameworks that are both flexible and capable of evolving alongside rapid technological advancements.[106] Risk-based governance approaches that can adapt to emerging risks and evolving societal expectations are also highlighted in the recent literature.[107] However, adaptive models must guard against a key blind spot: the risk of algorithmic nudging.[108] Subtle design choices in AI interfaces can steer human decision-making—including judicial reasoning—in ways that remain largely invisible even under transparency mandates.[109]

Therefore, adaptive governance must explicitly include mechanisms to detect and counteract nudging effects.[110]

Mandatory transparency and explainability standards remain a cornerstone of responsible AI governance. Explainability should not be confined to technical validation alone; it must also

[106] Jess; Nyrup Whittlestone Rune; Alexandrova, Anna, *Ethical and Societal Implications of Data and AI: A Roadmap for Research* (Nuffield Foundation 2019).
[107] Pouya Kashefi, Yasaman Kashefi and AmirHossein Ghafouri Mirsaraei, 'Shaping the Future of AI: Balancing Innovation and Ethics in Global Regulation' (2024) 29 Uniform Law Review 524, 524.
[108] Schmauder and others (n 25) 801.
[109] ibid.
[110] ibid.

ensure usability for human actors—such as judges and litigants—who interact with AI-supported decisions.[111] Yet explainability alone is insufficient to safeguard the integrity of judicial reasoning. I argue that policy frameworks should additionally require simulatability: the capacity for human judges to mentally model and anticipate AI outputs under varying conditions, thereby fostering deeper understanding and trust. Furthermore, contestability must be ensured, so that AI-generated recommendations can be meaningfully interrogated and challenged within judicial proceedings. Without such features, transparency risks becoming a superficial checkbox, providing the appearance of accountability without enabling substantive oversight.

Moreover, cognitive training for judges is essential but should be broadened. Teichman and Zamir highlight how various cognitive biases and contextual effects—such as compromise and contrast effects, anchoring, and the influence of irrelevant information—can shape judicial decision-making, potentially undermining the quality and fairness of judgments.[112] Building on this, I contend that judicial training must also address risks of automation complacency and potential emotional distancing. Some studies show that users often perceive algorithm-operated systems as more objective and less judgmental, which can foster excessive trust and reliance on AI outputs.[113] In judicial contexts, this dynamic may further contribute to reduced critical engagement and attentiveness to the narrative and moral dimensions of legal reasoning—an outcome that training programs should explicitly seek to counteract.

Finally, governance proposals must incorporate institutional oversight mechanisms. Regular audits, redundancy measures, and public testing have been identified as essential safeguards for

[111] Bolander (n 102) 851.
[112] Teichman and Zamir (n 50) s 3.
[113] Asbjørn Ammitzbøll Flügge, Thomas Hildebrandt and Naja Holten Møller, 'Street-Level Algorithms and AI in Bureaucratic Decision-Making: A Caseworker Perspective' (2021) 5 Proceedings of the ACM on Human-Computer Interaction 1, 6–7.

AI-supported judicial decision-making.[114] I propose that such mechanisms should be mandatory for AI tools used in judicial contexts, with results subject to public reporting and independent review. Without institutional checks, even well-intentioned transparency and training measures may fail to prevent systemic distortions of judicial reasoning.[115]

In sum, while existing policy proposals provide a solid foundation, a more ambitious and critical governance vision is needed. Adaptive governance must confront the risks of algorithmic nudging;[116] I argue that transparency must evolve into simulatability and contestability. Judicial training should address not only cognitive vulnerabilities[117], but also emotional dimensions of judicial engagement. Institutional oversight must be formalised and rigorously enforced.[118] Only through this multi-layered approach can AI governance in the judiciary genuinely safeguard the autonomy, integrity, and human-centred nature of legal decision-making.

3.3 Proposed Reforms and Normative Design Criteria

As AI-assisted judicial tools continue to evolve, there is an urgent need for normative design criteria that not only safeguard judicial integrity and public trust but also proactively reduce subtle risks that may otherwise erode the human-centred foundations of adjudication. Recent literature identifies several priorities, yet this section argues that future governance must go further—embedding mechanisms for contestability, audibility, and emotional resilience into both AI system design and institutional oversight.

First, the design of AI tools for judicial contexts must prioritise human-centric principles not only in terms of usability but also in terms of preserving narrative and expressive justice.

[114] Volokh (n 62) 1174–1176.
[115] Sourav (n 50) 5.
[116] Schmauder and others (n 25) 801.
[117] Teichman and Zamir (n 50) ss 3–7.
[118] Sourav (n 50) 5.

Judicial decision-making is characterised by deliberative and interpretive reasoning, which involves empathy, moral judgement, and contextual understanding.[119] Explainability must therefore be tailored to the practical needs of legal actors, facilitating meaningful engagement with the reasoning behind AI-supported decisions.[120] Building on this, AI design should explicitly support the capacity for narrative reasoning and moral engagement, rather than merely presenting decontextualised statistical outputs. Over-reliance on data-driven tools may lead judges to perceive litigants as mere data points, thereby undermining the moral and humanistic dimensions of adjudication.[121] Moreover, AI systems are often perceived as neutral and objective, which can foster over-trust and may contribute to emotional distancing and disengagement.[122] Human-centric design must thus encompass interface and interaction features that preserve the judge's role as an empathetic and morally engaged decision-maker.

Second, while some scholars have proposed limiting AI tools in judicial contexts to an advisory-only role,[123] I argue that advisory status alone is insufficient unless coupled with requirements for systematic audibility and institutionalised contestability. As shown in Chapter 1, even advisory-only AI systems can subtly shape judicial cognition through anchoring, automation bias, and framing effects. Without robust contestability and mechanisms to foster calibrated trust, advisory status alone cannot fully safeguard judicial discretion. Public testing and transparency mechanisms are essential to enable external auditing and oversight of AI outputs.[124] Without such procedural safeguards, advisory-only AI may still produce subtle

[119] Morison and McInerney (n 1) 3, 19, 25.
[120] Bolander (n 102) 851.
[121] Sourav (n 50) 5.
[122] Ammitzbøll Flügge, Hildebrandt and Møller (n 113) 6.
[123] Chesterman (n 85).
[124] Volokh (n 62) 1176.

forms of automation bias and path dependency, whereby judges unconsciously defer to seemingly authoritative outputs.[125]

Finally, future-proof governance must move beyond broad calls for adaptability and explicitly operationalise mechanisms for AI impact assessments and periodic review cycles. Co-regulatory frameworks that leverage collaboration between state and non-state actors have been proposed as effective strategies to foster adaptive governance.[126] In addition, forward-looking safeguards have been emphasised as essential to address the evolving nature of AI-related risks.[127] However, risks such as algorithmic nudging and latent biases can evolve dynamically and may elude static governance models.[128] Accordingly, I propose that regulatory regimes should mandate AI impact assessments prior to deployment, combined with annual review requirements to detect and address emergent risks. These mechanisms must be accompanied by transparent public reporting and independent oversight to sustain legitimacy and trust.

In sum, responsible AI governance in the judiciary must adopt a multi-layered and forward-facing approach. It must embed human-centric design principles that preserve narrative reasoning and moral engagement; enforce advisory-only usage through robust audit and contestability structures; and institutionalise dynamic oversight mechanisms to future-proof AI deployment against evolving risks. Only through such comprehensive reforms can AI be harnessed to enhance—rather than undermine—the human-centred foundations of judicial decision-making.

[125] Francesca Ceresa Gastaldo, 'The Automaton-Judge: Some Reflections on the Future of AI in Judicial Systems' 116.
[126] Erdélyi and Goldsmith (n 100) 5–6.
[127] Whittlestone (n 106).
[128] Schmauder and others (n 25) 801.

CONCLUSION

The integration of AI into judicial decision-making presents not merely a technical evolution, but a profound epistemic and normative shift. As this thesis has demonstrated, AI-powered decision-support tools are not neutral instruments; they actively shape the cognitive environment within which judges reason and decide. Through mechanisms such as anchoring, automation bias, and algorithmic nudging, AI systems can subtly, and at times systematically, influence judicial outcomes. Moreover, their black-box characteristics challenge foundational principles of transparency, accountability, and contestability—principles that underpin the legitimacy of adjudication.

Across jurisdictions, responses to these risks have varied. While the European Union has embraced a rights-driven and precautionary approach, the United States' fragmented governance landscape highlights the dangers of insufficient oversight. China's model illustrates both the promise and the perils of rapid AI adoption in judicial contexts. These comparative insights underscore that effective governance of judicial AI must be both adaptive to evolving technological risks and sensitive to the legal-cultural contexts in which such tools are deployed.

To this end, the thesis argues that AI integration in the judiciary must be guided by a multi-layered governance model. First, human-centric design must be prioritised, ensuring that AI systems support—not supplant—the deliberative, narrative, and moral dimensions of judging. Second, procedural safeguards must move beyond transparency alone to guarantee meaningful contestability, including simulatability and institutionalised auditing. Third, judicial training must explicitly address both cognitive and emotional risks associated with AI reliance. Finally, dynamic oversight mechanisms—encompassing periodic AI impact assessments and robust co-

regulatory structures—are essential to ensure that AI governance remains responsive to emerging risks.

In short, AI should augment rather than erode the human-centred foundations of justice. Achieving this balance demands more than technical fixes; it requires a normative recalibration of how legal systems conceptualise the role of AI in adjudication. As courts increasingly encounter AI-driven tools, the principles articulated here—anchoring judicial autonomy, preserving human discretion, and institutionalising accountability—offer a path toward ensuring that the pursuit of efficiency does not come at the expense of justice itself. In the end, as this thesis has argued, AI is not a passive technical instrument but an epistemic actor that reshapes the architecture of judicial reasoning—an influence that must be governed with care if courts are to remain trusted custodians of justice.

BIBLIOGRAPHY

Al-Zahrani, Abdulrahman M, 'Balancing Act: Exploring the Interplay Between Human Judgment and Artificial Intelligence in Problem-Solving, Creativity, and Decision-Making' (2024) 2 *IgMin Research* 145

Olaborede, Adebola and Meintjes-van Der Walt, Lirieka, 'Cognitive Bias Affecting Decision-Making in the Legal Process' (2021) 41 *Obiter* 806

Ammitzbøll Flügge, Asbjørn, Hildebrandt, Thomas and Holten Møller, Naja, 'Street-Level Algorithms and AI in Bureaucratic Decision-Making: A Caseworker Perspective' (2021) 5 *Proceedings of the ACM on Human-Computer Interaction* 1

Bahl, Utsav and others, 'Algorithms in Judges' Hands: Incarceration and Inequity in Broward County, Florida' (SocArXiv, 1 May 2023)

Basdevant, A, Jean, A and Storchan, V, *Mécanisme d'une justice algorithmisée* (Fondation Jean Jaurès 2021)

Bolander, Thomas, 'What Do We Lose When Machines Take the Decisions?' (2019) 23 *Journal of Management and Governance* 849

Carter, Lemuria and Liu, Dapeng, 'How Was My Performance? Exploring the Role of Anchoring Bias in AI-Assisted Decision Making' (2025) 82 *International Journal of Information Management* 102875

Ceresa Gastaldo, Francesca, 'The Automaton-Judge: Some Reflections on the Future of AI in Judicial Systems' (2024) 14 *Sortuz. Oñati Journal of Emergent Socio-legal Studies* 399

Chen, Benjamin Minhao and Li, Zhiyu, 'How Will Technology Change The Face of Chinese Justice?' (2020) *Columbia Journal of Asian Law*

Chesterman, Simon, 'All Rise for the Honourable Robot Judge? Using Artificial Intelligence to Regulate AI' (2022) *SSRN Electronic Journal*

Duarte, Regina de Brito and Campos, Joana, 'Looking For Cognitive Bias In AI-Assisted Decision-Making' (INESC-ID, Instituto Superior Técnico, 2024)

Englich, Birte, Mussweiler, Thomas and Strack, Fritz, 'Playing Dice With Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making' (2006) 32 *Personality and Social Psychology Bulletin* 188

Erdélyi, Olivia J and Goldsmith, Judy, 'Regulating Artificial Intelligence: Proposal for a Global Solution' (arXiv, 22 May 2020)

European Commission, 'Ethics Guidelines For Trustworthy AI' (2019)

European Union, 'Regulation - EU - 2024/1689 of the European Parliament and of the Council of 13 June 2024 on Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)

and Amending Regulations (EU) 2017/1128 and (EU) 2022/2065 and Directives 2002/58/EC and (EU) 2016/1148' (2024) *Official Journal of the European Union*

Fregosi, Caterina and Cabitza, Federico, 'A Frictional Design Approach: Towards Judicial AI and Its Possible Applications' (HHAI-WS 2024: Workshops at the Third International Conference on Hybrid Human-Artificial Intelligence (HHAI), 2024)

Kashefi, Pouya, Kashefi, Yasaman and Ghafouri Mirsaraei, AmirHossein, 'Shaping the Future of AI: Balancing Innovation and Ethics in Global Regulation' (2024) 29 *Uniform Law Review* 524

Kirchner, Lauren, 'Wisconsin Court: Warning Labels Are Needed for Scores Rating Defendants' Risk of Future Crime' *ProPublica* (11 August 2016)

Loomis, Eric L, 'In the Supreme Court of the United States' *Brief in Opposition* (2017)

Macquarie Law School, Macquarie University and others, 'THE USE OF ARTIFICIAL INTELLIGENCE IN JUDICIAL DECISIONMAKING: THE EXAMPLE OF CHINA' (2023) 2022 *International Journal of Law, Ethics, and Technology* 1

McKay, Carolyn, 'Predicting Risk in Criminal Procedure: Actuarial Tools, Algorithms, AI and Judicial Decision-Making' (2020) 32 *Current Issues in Criminal Justice* 22

Morison, John and McInerney, Tomás, 'When Should a Computer Decide? Judicial Decision-Making in the Age of Automation, Algorithms and Generative Artificial Intelligence' in S Turenne and M Moussa (eds), *Research Handbook on Judging and the Judiciary* (Edward Elgar-Routledge forthcoming 2024)

National Center for State Courts, 'AI Foundations in the Courts' https://www.ncsc.org/resources-courts/ai-foundations-courts (accessed 10 June 2025)

Pfeiffer, Jella and others, 'Algorithmic Fairness in AI: An Interdisciplinary View' (2023) 65 *Business & Information Systems Engineering* 209

Schmauder, Christian and others, 'Algorithmic Nudging: The Need for an Interdisciplinary Oversight' (2023) 42 *Topoi*799

Sourav, RaisuL, 'Relying on AI in Judicial Decision-Making: Justice or Jeopardy?' (2024)

Sourdin, Tania, 'Judge v Robot? Artificial Intelligence and Judicial Decision-Making' (2018) 41 *University of New South Wales Law Journal* 1114

Steyvers, Mark and Kumar, Aakriti, 'Three Challenges for AI-Assisted Decision-Making' (2024) 19 *Perspectives on Psychological Science* 722

Teichman, Doron and Zamir, Eyal, 'Judicial Decision-Making: A Behavioral Perspective' in Eyal Zamir and Doron Teichman (eds), *The Oxford Handbook of Behavioral Economics and the Law* (Oxford University Press 2014)

Taylor, Luke, 'Colombian Judge Says He Used ChatGPT in Ruling' *The Guardian* (3 February 2023)

United Nations Chief Executives Board for Coordination, *United Nations System White Paper on AI Governance* (May 2024)

Volokh, Eugene, 'CHIEF JUSTICE ROBOTS' (2019) 105 *Virginia Law Review* 1163

Watamura, Eiichiro, Liu, Yichen and Ioku, Tomohiro, 'Judges versus Artificial Intelligence in Juror Decision-Making in Criminal Trials: Evidence from Two Pre-Registered Experiments' (2025) 20 *PLOS ONE* e0318486

Whittlestone, J N Rune, Alexandrova, Anna and Waseem, Zana, *Ethical and Societal Implications of Data and Artificial Intelligence: A Roadmap for Research* (Nuffield Foundation 2019)