WHITE PAPER

System 1 Lawyering

The Cognitive Risk of AI-Assisted Legal Work and the Case for Structured Supervision

Shivam Shukla
Advocate, Allahabad High Court

March 2026

ABSTRACT

AI-generated legal content presents a cognitive risk that existing discussions of hallucination and fabrication do not adequately capture. Drawing on Kahneman's dual-process theory of cognition, this paper demonstrates that the fluency and structural coherence of AI-generated output systematically activates the advocate's fast, intuitive cognitive layer (System 1), suppressing the effortful verification that professional obligations demand (System 2). The paper identifies four cognitive mechanisms through which this substitution produces professional failure, maps each to documented judicial findings from Indian courts in 2025-2026, and demonstrates that a structured supervisory method operating at the level of cognitive process—not platform features—is the necessary countermeasure. The Supervised Intelligence Method is analysed as a cognitive intervention that forces System 2 engagement at each stage where professional judgment is required.

Keywords: dual-process theory, AI-assisted legal practice, cognitive bias, professional responsibility, supervised intelligence, legal governance, hallucination risk

SECTION I

The Problem Beyond Hallucination

The prevailing account of AI risk in legal practice centres on hallucination: the generation of fictitious citations, non-existent judgments, and fabricated statutory provisions. This account is not wrong. It is incomplete. Hallucination is a symptom. The underlying pathology operates at the level of the advocate's cognition.

When a large language model produces a legal memorandum that reads fluently, cites plausible-sounding authorities, structures arguments in recognisable doctrinal form, and presents conclusions with apparent analytical rigour, it does not merely create a risk of factual error. It creates a cognitive environment in which the advocate is systematically less likely to detect that error—even when detection is well within the advocate's professional competence.

This is not a claim about AI's limitations. It is a claim about how human cognition responds to AI-generated output. The risk is not that the machine fails. The risk is that the advocate's cognitive defences are structurally weakened by the very quality of the machine's output.

The distinction matters because it determines what kind of countermeasure is required. If the problem is hallucination, the solution is better AI. If the problem is cognitive, the solution must operate at the level of the advocate's decision-making process. This paper argues for the second position.

SECTION II

Dual-Process Theory and Professional Cognition

Daniel Kahneman's dual-process framework, developed across four decades of research with Amos Tversky and others, distinguishes two modes of cognitive operation.

System 1 operates automatically, rapidly, and with minimal effort. It recognises patterns, generates intuitive judgments, and produces impressions of coherence. It is the cognitive system that reads a well-structured paragraph and registers it as ‘correct’ before any deliberate analysis has occurred. System 1 excels at fluency detection—the assessment of whether information ‘feels right’ based on its presentation, familiarity, and internal consistency.

System 2 is effortful, deliberate, and slow. It performs logical analysis, checks premises against evidence, and evaluates whether an apparently coherent conclusion is actually supported by its stated foundations. System 2 is the cognitive system that a professional obligation of verification demands. It is also the system that requires active engagement—it does not activate automatically.

The central finding of decades of cognitive research is that System 1 operates as the default. System 2 engages only when System 1 encounters difficulty, surprise, or an explicit demand for deliberation. When the input is fluent, structured, and internally coherent, System 1 produces a judgment of validity—and System 2 often ratifies that judgment without independent scrutiny.

Kahneman termed this the *law of least effort*: cognitive systems default to the processing mode that requires the least expenditure of attentional resources.

2.1 The Professional Significance

Legal practice has always required System 2 dominance at critical decision points. The advocate who reads a judgment must assess not merely whether the prose is coherent but whether the ratio supports the proposition for which it is cited. The advocate who drafts a petition must verify not merely that the statutory provision exists but that its text says what the argument requires it to say. These are System 2 operations: effortful, deliberate, resistant to the surface impression.

Before the availability of generative AI tools, the advocate's drafting process inherently engaged System 2. Research required reading judgments. Drafting required constructing arguments from first principles. The cognitive effort was embedded in the workflow. The advocate could not produce a submission without having performed the analytical work that verification demands.

AI-generated legal content fundamentally alters this cognitive environment.

SECTION III

AI Output as a System 1 Trigger

Large language models produce output that is optimised, by architecture and training, for precisely the characteristics that activate System 1 acceptance: fluency, structural coherence, and pattern conformity. This is not a design flaw. It is the operational objective of the technology. The model generates text that reads as though a competent professional produced it—because it was trained on the patterns of competent professional output.

The result is that AI-generated legal content bypasses the cognitive friction that traditionally forced System 2 engagement. The advocate receives a memorandum that looks correct, reads correctly, and is structured correctly. System 1 registers coherence. The effortful verification that System 2 would perform—locating the cited judgment, reading the ratio, confirming the statutory text—is not triggered, because nothing in the surface presentation signals a need for it.

3.1 Four Cognitive Mechanisms

The substitution of System 1 acceptance for System 2 verification operates through at least four identifiable cognitive mechanisms, each of which has produced documented professional failure in Indian courts.

Fluency-Validity Substitution.

Cognitive research demonstrates that processing fluency—the ease with which information is absorbed—correlates with perceived truth. Statements that are easier to read are rated as more likely to be true, independent of their actual content. AI-generated legal text is engineered for maximum fluency. The advocate's System 1 registers the fluency and substitutes it for a judgment

of validity. The fabricated citation that appears in proper format, with a plausible case name and a correct-looking neutral citation, is processed as valid by System 1 precisely because it conforms to the expected pattern.

Anchoring on AI-Generated Structure.

When an advocate receives a structured legal memorandum from an AI system—with headings, sub-issues, cited authorities, and a conclusion—the structure itself becomes an anchor. Kahneman's anchoring effect demonstrates that initial information disproportionately influences subsequent judgment, even when the anchor is arbitrary. The AI-generated structure anchors the advocate's analysis: the issues identified become the issues pursued, the authorities cited become the authorities considered, and the analytical framework becomes the framework adopted. The advocate edits within the AI's structure rather than reconstructing the argument from the advocate's own analysis of the law and facts.

Coherence as Proof.

System 1 evaluates information by constructing the most coherent story available from the data presented. Kahneman described this as WYSIATI—What You See Is All There Is. When AI-generated output presents a coherent narrative—issue, authority, application, conclusion—System 1 accepts the narrative as complete. The advocate does not ask what is missing because the presented story is internally consistent. The missing fact, the unreported overruling decision, the statutory amendment not captured in the training data—these absences do not register because the presented material forms a coherent whole.

Effort Substitution.

When faced with a difficult question, System 1 frequently substitutes an easier question and answers that instead. The difficult question—'Does this authority actually support this proposition when read in full?'—is substituted with the easier question—'Does this authority look like it supports this proposition based on the AI's summary?' The advocate answers the easier question, System 1 registers satisfaction, and the difficult question is never engaged. This mechanism explains why partial verification—confirming that a cited judgment exists without reading it—is both common and professionally inadequate.

SECTION IV

Judicial Evidence: Cognitive Failure in Indian Courts

The cognitive mechanisms described above are not theoretical predictions. They have produced documented judicial consequences in Indian courts within the twelve months preceding this paper.

4.1 Gummadi Usha Rani v. Sure Mallikarjuna Rao

SLP(C) 7575/2026, Supreme Court of India, 27 February 2026. A trial court judge used an AI tool to draft a judicial order. The order cited four judgments. None existed. The Supreme Court held

that this constitutes misconduct—not mere error but a failure of the professional obligation that attaches to every judicial act.

Cognitively, this is a textbook instance of fluency-validity substitution compounded by effort substitution. The AI-generated draft presented citations in correct format. System 1 registered pattern conformity. The effortful step—locating each judgment and reading it—was not performed because the surface presentation did not trigger System 2 engagement. The judge answered the easy question (‘Do these citations look correct?’) rather than the hard question (‘Do these judgments exist and do they say what is claimed?’).

4.2 Blue Star Aluminium & Door House v. Federal Bank

WP(C) 43123/2025, Kerala High Court, 10 December 2025. The court observed that several writ petitions before it appeared to be AI-generated—exhibiting formal structure but absent material facts. Advocates were unable to answer the court's questions about their own pleadings.

This is anchoring made visible. The advocates adopted the AI-generated structure without reconstructing it around the specific facts of their clients' cases. The structure became the submission. When questioned, the advocates could not explain the analytical choices embedded in their own pleadings—because those choices were the AI's, accepted at the level of System 1 without System 2 reconstruction.

4.3 Deepak Bahry v. Heart & Soul Entertainment

2026:BHC-AS:828, Bombay High Court, 7 January 2026. Costs were imposed for AI-generated submissions containing fabricated citations. The court recorded the failure of verification as a professional conduct failure.

The cognitive pattern here is coherence-as-proof operating alongside effort substitution. The submission presented a coherent argument supported by citations that appeared genuine. System 1 accepted the coherent story. The advocate did not perform independent verification because the presented material formed an internally consistent narrative—the WYSIATI mechanism prevented the advocate from asking what might be missing or wrong.

SECTION V

Why Platform Improvements Cannot Solve a Cognitive Problem

A common objection holds that the solution lies in improving AI systems: reducing hallucination rates, embedding verification into platforms, linking citations to verified databases. This objection misidentifies the problem.

Even a hypothetical AI system that never fabricates a citation still produces output that the advocate has not independently verified. The advocate who relies on a platform's internal verification has outsourced a professional obligation. The verification is the platform's, not the advocate's. The advocate's independent knowledge of the authority cited—its ratio, its application,

its relationship to the specific facts—remains absent.

More fundamentally, the cognitive mechanisms described in Section III do not depend on AI being unreliable. They depend on AI output being fluent and structured. Even perfectly accurate AI output activates System 1 and suppresses System 2. The advocate who accepts an accurate AI-generated memorandum without reconstruction has still failed to exercise independent professional judgment. The submission is the AI's work with the advocate's signature—regardless of whether it contains errors.

The cognitive problem therefore survives every platform improvement. It can only be addressed by an intervention that operates at the level of the advocate's decision-making process, not at the level of the AI's output quality.

SECTION VI

Structured Supervision as Cognitive Countermeasure

If the risk is cognitive, the countermeasure must be cognitive. A supervisory method that forces System 2 engagement at each stage where professional judgment is required addresses the problem at its source.

The Supervised Intelligence Method, as articulated in the literature on AI-assisted legal practice, provides one such framework. Its five stages are analysed here not as operational procedures but as cognitive interventions.

6.1 Stage 1 — Legal Framing (Human Only)

Cognitive function: Forces the advocate to define the legal question before any AI system is engaged. This pre-commits System 2 to the analytical framework. The advocate's own framing becomes the reference point against which all subsequent AI output is evaluated. Without this stage, the AI's framing becomes the anchor—and the advocate's System 2 never establishes independent analytical control.

6.2 Stage 2 — Pattern Expansion (AI-Assisted)

Cognitive function: Designates all AI output as presumptively unverified. This is not merely a procedural label. It is a cognitive instruction that prevents System 1 from registering AI output as validated. The presumption of unreliability creates the cognitive friction that System 1 would otherwise eliminate.

6.3 Stage 3 — Doctrinal Reconstruction (Human-Dominant)

Cognitive function: Requires the advocate to reconstruct the argument from the advocate's own analysis—not to edit the AI's output. This is the critical distinction between editing and reconstruction. Editing operates within the AI's structure (System 1 anchoring preserved). Reconstruction replaces the AI's structure with the advocate's own analytical framework (System 2 dominance established). The advocate who reconstructs cannot avoid engaging System 2,

because reconstruction requires the advocate to make every analytical choice independently.

6.4 Stage 4 — Verification (Human Only)

Cognitive function: Mandates independent confirmation of every citation against primary sources. This is the direct countermeasure to effort substitution. The method does not permit the advocate to answer the easy question (‘Does this citation look correct?’). It requires the hard question (‘Have I located, read, and confirmed this authority against the enacted text or reported judgment?’). The separation of verification of authorities from verification of statutory provisions ensures that neither category benefits from a general impression of completeness.

6.5 Stage 5 — Strategic Judgment (Human Only)

Cognitive function: Reserves all strategic decisions to the advocate. This prevents the coherence bias from determining which arguments are advanced and how they are sequenced. The AI system presents all arguments at equal weight—because it cannot assess the bench, the facts, or the tactical landscape. System 1 would accept this equal weighting as adequate. Stage 5 forces System 2 to impose strategic hierarchy based on the advocate's professional judgment.

SECTION VII

Implications for Institutional Governance

The cognitive analysis presented here carries three implications for the institutional governance of AI use in legal practice.

First, any institutional standard that addresses AI risk only at the level of output quality—hallucination rates, citation accuracy, platform reliability—is structurally insufficient. The cognitive risk survives every improvement in AI output quality. Institutional standards must address the advocate's cognitive process, not merely the AI's technical performance.

Second, disclosure requirements alone are inadequate. Requiring an advocate to disclose that AI was used in preparation of a submission addresses transparency but not competence. The advocate who discloses AI use but has not performed independent verification has satisfied a procedural requirement while failing a substantive professional obligation.

Third, any governance framework must be platform-independent. Cognitive risk does not vary by platform. The mechanisms described in this paper operate identically whether the AI system is a commercial chatbot, a legal-specific research tool, or an open-source model running on local infrastructure. A governance framework tied to specific platforms is both underinclusive and fragile. The framework must govern the advocate's process, which remains constant across all technological configurations.

SECTION VIII

Conclusion

The risk that AI poses to the quality of legal practice is not primarily technological. It is cognitive. The fluency and structural coherence of AI-generated legal content systematically activates the advocate's fast, intuitive cognitive processes and suppresses the effortful verification that professional obligations require. Every documented instance of AI-related professional failure in Indian courts—fabricated citations, unverified authorities, submissions the advocate cannot explain—is traceable to a specific cognitive mechanism: fluency-validity substitution, anchoring, coherence bias, or effort substitution.

The countermeasure must therefore be cognitive. A structured supervisory method that forces System 2 engagement at each stage where professional judgment is required addresses the problem at its source. Such a method does not restrict AI use. It governs it. The advocate who operates within a structured supervisory framework uses AI more effectively—not less—because every output is independently verified, every argument is independently reconstructed, and every strategic decision is independently made.

The professional obligations that such a method enforces are not new. They predate every AI system and will survive every AI system. What is new is the cognitive environment in which those obligations must now be discharged—an environment in which the advocate's own cognitive architecture is systematically deployed against the advocate's professional interests.

Recognising this is the first step. Building the institutional response is the next.

REFERENCES

[1] Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

[2] Tversky, A. & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157), 1124–1131.

[3] Stanovich, K.E. & West, R.F. (2000). Individual Differences in Reasoning: Implications for the Rationality Debate? *Behavioural and Brain Sciences*, 23(5), 645–665.

[4] Kahneman, D. & Frederick, S. (2002). Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment.* Cambridge University Press.

[5] Oppenheimer, D.M. (2008). The Secret Life of Fluency. *Trends in Cognitive Sciences*, 12(6), 237–241.

[6] Vaswani, A. et al. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems* (NIPS 2017).

[7] Gummadi Usha Rani v. Sure Mallikarjuna Rao, SLP(C) 7575/2026, Supreme Court of India (27 February 2026).

[8] Blue Star Aluminium & Door House v. Federal Bank, WP(C) 43123/2025, Kerala High Court (10 December 2025).

[9] Deepak Bahry v. Heart & Soul Entertainment, 2026:BHC-AS:828, Bombay High Court (7 January 2026).

[10] Supreme Court of India, White Paper on Artificial Intelligence and the Judiciary (November 2025).

This paper is an independent academic and professional contribution. It does not constitute legal advice or professional engagement. The author is a practising advocate and the analysis reflects his independent professional assessment.

© 2026 Shivam Shukla. All rights reserved.