THE ARBITRATION LEGITIMACY STACK: DUE PROCESS, EXPLAINABILITY, AND ENFORCEABILITY OF AI-AUGMENTED AWARDS IN A POST-AGI WORLD

Haley Yi, David Scott Lewis
AIXC Research
reports@aiexecutiveconsulting.com

ABSTRACT

International commercial arbitration is a private, transnational governance system whose legitimacy rests on a fragile bargain: parties trade public adjudication for speed and expertise, but they still demand due process and an enforceable award. As AI assistance becomes routine in case management, evidence triage, translation, and draft-award generation, enforceability will increasingly hinge on *process evidence*—what a tribunal can show about how it reached a decision, not just what it decided. We propose the *Arbitration Legitimacy Stack*, a socio-legal framework that decomposes legitimacy into three auditable layers: (1) rule-level transparency (which rules and instruments were applied), (2) inference-level transparency (inspectable steps from premises to conclusions, ideally machine-checkable for rule-governed determinations), and (3) narrative-level transparency (human-readable reasons that faithfully summarize the inference artifacts). We argue that post-AGI arbitration fails when these layers decouple: fluent narratives can mask untraceable inferences, while formal traces can become unintelligible to parties. To operationalize the stack, we introduce an *AI-Assisted Award Record* (AAR): a compact bundle containing AI-use disclosure, provenance metadata for evidence operations, solver-checkable traces for formalizable steps (e.g., jurisdiction, admissibility, timelines, costs), and an override log capturing non-delegable human judgment. We sketch confidentiality-preserving disclosure modes and outline a Tiny-Paper-sized research agenda: micro-benchmarks tied to procedural rulebooks, cross-layer faithfulness metrics, and auditability-by-design patterns resilient to prompt-injection and explanation-laundering attacks.

1 MOTIVATION: ENFORCEABILITY BECOMES PROCESS-AUDITABLE

The New York Convention (NYC) makes arbitration "portable" by committing Contracting States to recognize and enforce foreign awards subject to limited defenses, including lack of notice or inability to present a case, excess of mandate, and public policy. (United Nations, 1958) The UNCITRAL Model Law mirrors these guardrails in its setting-aside and enforcement provisions, and also imposes baseline procedural commitments such as equal treatment and a full opportunity to present one's case. (United Nations Commission on International Trade Law (UNCITRAL), 2006) The upshot is stable: arbitration tolerates private decision-making because courts presume a procedurally fair process that produces an award recognizable across jurisdictions. (United Nations Commission on International Trade Law (UNCITRAL), 2016a)

In a post-AGI world, tribunals will not merely *use* AI; they will operate inside AI-saturated workflows (search, summarization, translation, chronology construction, damages modeling, and procedural drafting). Technology reports already frame adoption as compatible with fairness only if deployed with safeguards that preserve equality of arms, information security, and procedural integrity. (ICC Arbitration and ADR Commission, 2022) Dedicated AI guidance emphasizes competence, confidentiality, and non-delegable decision-making by arbitrators. (Silicon Valley Arbitration & Mediation Center (SVAMC), 2024; Chartered Institute of Arbitrators (CIArb), 2025; American Arbitration Association (AAA-ICDR), 2025) This shifts the legitimacy target from "a reasoned award" to a

reasoned process: parties and enforcing courts will increasingly ask what constraints were enforced, what uncertainty remained, and where human judgment intervened (Appendix Figure 3).

2 THE ARBITRATION LEGITIMACY STACK

We propose a decomposition of arbitral legitimacy into three auditable layers (Figure 1).

Layer 1: rule-level transparency. This layer captures *which* norms were applied: the arbitration agreement, institutional rules, and relevant soft-law instruments (e.g., evidence protocols). (London Court of International Arbitration (LCIA), 2020; International Chamber of Commerce (ICC), 2021; International Bar Association (IBA), 2020) Rule-level transparency is necessary for mandate disputes: parties and enforcing courts ask whether the tribunal decided issues outside the submission to arbitration or departed from agreed procedure. (United Nations, 1958; United Nations Commission on International Trade Law (UNCITRAL), 2016a)

Layer 2: inference-level transparency. This layer captures *how* conclusions follow from premises: admissibility and procedural determinations, jurisdictional checks, and constrained fact-to-rule mappings. For AI-augmented reasoning, inference transparency is where "glass-box" behavior lives: steps are inspectable and, when appropriate, machine-checkable (e.g., via constraint solvers or proof traces) rather than only rhetorically plausible. Formal verification and traceability are established techniques for making complex decision procedures contestable. (Clarke et al., 1999)

Layer 3: narrative-level transparency. This layer is the award as understood by humans: a coherent account that ties the record to the dispositive reasoning. Because post-hoc explanations can be misleading, narrative transparency must be evaluated for *faithfulness* to inference artifacts, not just readability. (Rudin, 2019; Lipton, 2016; Guidotti et al., 2018) In arbitration, narrative clarity also serves the "right to be heard" function by enabling parties to challenge key premises before an award hardens. (United Nations Commission on International Trade Law (UNCITRAL), 2016b)

Failure mode: layer decoupling. Post-AGI failure occurs when layers drift apart: fluent narrative masking uncheckable inference; or solver-correct traces that parties cannot understand. The stack therefore implies a *cross-layer consistency obligation*: narrative statements should be grounded in identifiable inference artifacts, and inference artifacts should be anchored to declared rules and provenance.

3 THE AI-ASSISTED AWARD RECORD (AAR)

We propose an *AI-Assisted Award Record* as a compact, standardizable bundle that supports challenges and enforcement while protecting confidentiality (Figure 2). The AAR is not a new "public record"; it is a *structured audit artifact* that can be disclosed in full, partially, or under protective orders.

AAR-1: AI-use disclosure. A minimal disclosure should specify tools/models used, scope of reliance, whether external data was transmitted, and which outputs were reviewed or adopted. This aligns with AI-in-arbitration guidance and with broader legal-profession recommendations emphasizing competent use and, when appropriate, disclosure to clients. (Silicon Valley Arbitration & Mediation Center (SVAMC), 2024; Chartered Institute of Arbitrators (CIArb), 2025; American Arbitration Association (AAA-ICDR), 2025; International Bar Association (IBA) & Center for AI and Digital Policy (CAIDP), 2024)

AAR-2: provenance for evidence operations. Evidence triage and summarization should carry provenance metadata (document identifiers, timestamps, extraction/summarization lineage, and chain-of-custody), so parties can audit what the system saw and how derived excerpts were produced. Standard provenance models exist for representing such lineage. (World Wide Web Consortium (W3C), 2013) Model and dataset documentation practices (e.g., model cards, datasheets) provide templates for communicating limitations and intended use that can be adapted to arbitration toolchains. (Mitchell et al., 2019; Gebru et al., 2018)

AAR-3: solver-checkable traces for rule-governed steps. Many determinations in arbitration are naturally formalizable: jurisdictional prerequisites, time limits, and some cost allocation criteria. (United Nations Commission on International Trade Law (UNCITRAL), 2006; London Court of International Arbitration (LCIA), 2020) Where formalization is feasible, the AAR should include a trace that can be independently checked, reducing the risk that generative text "launders" weak reasoning.

AAR-4: override and deliberation log. To preserve the non-delegable core of arbitral judgment, the record should capture where arbitrators rejected, modified, or constrained AI suggestions (without exposing privileged strategy). This directly supports the "no improper delegation" principle articulated in AI guidance for arbitrators. (Silicon Valley Arbitration & Mediation Center (SVAMC), 2024; American Arbitration Association (AAA-ICDR), 2025)

Confidentiality-preserving options. Two pragmatic approaches can reconcile auditability with confidentiality: (i) *redactable traces* that expose logical skeletons while pointing to sealed evidence, and (ii) *privacy-preserving audits* in which a neutral reviewer verifies logs under protective orders. Cybersecurity protocols provide process patterns for implementing proportional information-security measures in arbitration. (ICCA-NYC Bar-CPR Working Group, 2022)

4 THREAT MODEL: POST-AGI LEGITIMACY RISKS

AI saturates arbitration with new adversarial surfaces.

Asymmetry and strategic opacity. If one party has better models or better data access, AI can amplify informational advantages. Guidance documents increasingly call for competence and appropriate disclosure to mitigate unfairness and information-security risks. (Silicon Valley Arbitration & Mediation Center (SVAMC), 2024; Chartered Institute of Arbitrators (CIArb), 2025; International Bar Association (IBA) & Center for AI and Digital Policy (CAIDP), 2024)

Prompt injection and evidence poisoning. Evidence corpora can contain adversarial instructions that steer AI tools during summarization or chronology construction. Prompt injection attacks are documented in real-world LLM-integrated applications and can induce arbitrary misbehavior. (Liu et al., 2023; 2024) System-level defenses based on information-flow control motivate treating evidence as untrusted input that must not rewrite decision instructions. (Wu et al., 2024)

Explanation laundering. Post-hoc explanations can give a false sense of correctness: a persuasive narrative may not reflect the model's actual decision logic. (Lipton, 2016; Rudin, 2019; Guidotti et al., 2018) The legitimacy stack addresses this by requiring narrative claims to be traceable to inference artifacts.

5 RESEARCH AGENDA (TINY-PAPER-SIZED)

We close with a tractable agenda that fits the workshop format.

Micro-benchmarks aligned to procedural rulebooks. Construct small, public tasks derived from procedural provisions (e.g., Model Law triggers and timelines; institutional procedural requirements) and evaluate whether systems produce (a) correct outcomes and (b) auditable traces. (United Nations Commission on International Trade Law (UNCITRAL), 2006; London Court of International Arbitration (LCIA), 2020)

Cross-layer faithfulness metrics. Measure whether narrative explanations faithfully summarize inference traces, drawing on interpretability evaluation proposals and counterfactual-style "what would change the outcome" diagnostics. (Doshi-Velez & Kim, 2017; Wachter et al., 2018)

Auditability-by-design for confidentiality. Develop AAR redaction schemes and third-party audit protocols that minimize disclosure while maximizing verifiability, leveraging risk-management and

transparency frameworks for trustworthy AI systems. (National Institute of Standards and Technology, 2023; European Union, 2024)

Limitations. This paper is conceptual: it does not claim that enforcing courts will adopt a single artifact standard. Instead, it argues that as AI use becomes normal, *standardized process evidence* will be increasingly salient for maintaining enforceability under existing doctrines.

APPENDIX: FIGURES AND MINIMAL AAR SCHEMA (EXCLUDED FROM PAGE LIMIT)

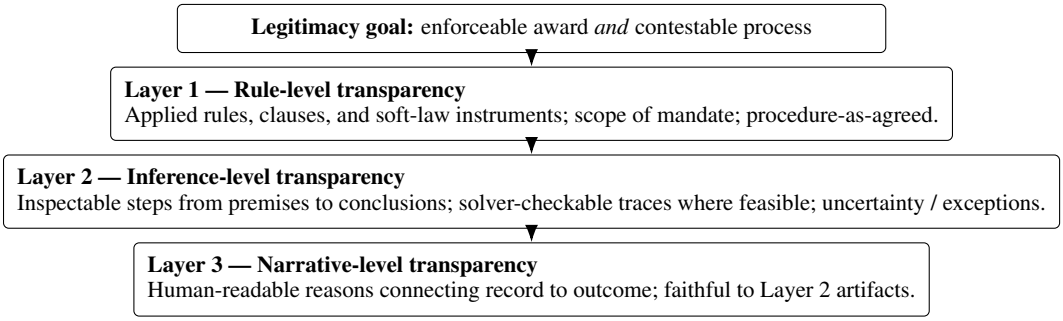

Figure 1: The Arbitration Legitimacy Stack: enforceability requires consistent rule-, inference-, and narrative-level transparency.

AI-Assisted Award Record (AAR)

Artifact	Purpose / example contents
AI-use disclosure	Tools/models; scope of reliance; data-handling; human review points.
Provenance bundle	Document IDs; timestamps; extraction lineage; chain-of-custody.
Rule traces	Solver-checkable steps for jurisdiction, admissibility, timelines, costs.
Override log	Where AI suggestions were rejected/edited; rationale tags (non-privileged).

Disclosure modes: (1) full production to parties; (2) redacted logical skeleton + sealed record pointers; (3) third-party audit under protective order.

Figure 2: AI-Assisted Award Record (AAR): minimal audit artifacts and confidentiality-preserving disclosure modes.

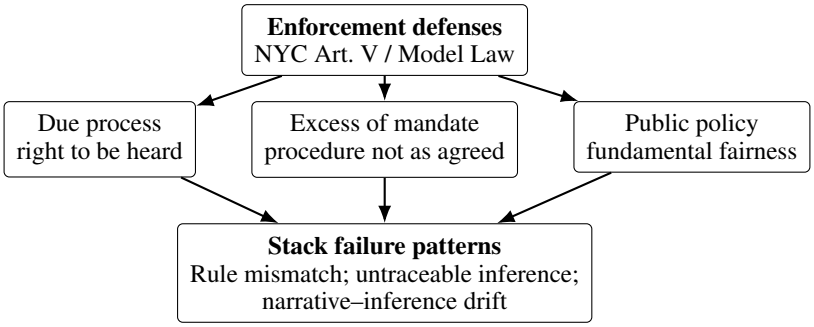

Figure 3: Mapping: enforcement challenges become questions about stack failures. The AAR is designed to supply the relevant process evidence.

REFERENCES

American Arbitration Association (AAA-ICDR). AAA-ICDR guidance on arbitrators' use of AI tools, 2025. URL https://go.adr.org/rs/294-SFS-516/images/2025_AAA-ICDR%20Guidance%20on%20Arbitrators%20Use%20of%20AI%20Tools%20%282%29.pdf.

Chartered Institute of Arbitrators (CIArb). Guideline on the use of AI in arbitration, 2025. URL https://www.ciarb.org/media/bpndtcgu/guideline-on-the-use-of-ai-in-arbitration_updated-sept-2025.pdf. Updated September 2025.

Edmund M. Clarke, Orna Grumberg, and Doron A. Peled. *Model Checking*. MIT Press, 1999. ISBN 9780262032704. URL https://mitpress.mit.edu/9780262032704/model-checking/.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017. URL https://arxiv.org/abs/1702.08608. arXiv:1702.08608.

European Union. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act), 2024. URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng. OJ L, 2024/1689, 12.7.2024.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets, 2018. URL https://arxiv.org/abs/1803.09010. arXiv:1803.09010.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, and Fosca Giannotti. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5): 93:1–93:42, 2018. doi: 10.1145/3236009. URL https://dl.acm.org/doi/10.1145/3236009.

ICC Arbitration and ADR Commission. Leveraging technology for fair, effective and efficient international arbitration proceedings. Technical report, International Chamber of Commerce (ICC), 2022. URL https://iccwbo.org/wp-content/uploads/sites/3/2022/02/icc-arbitration-and-adr-commission-report-on-leveraging-technology-for-fair-effective-and-efficient-international-arbitration-proceedings.pdf.

ICCA-NYC Bar-CPR Working Group. Icca-nyc bar-cpr protocol on cybersecurity in international arbitration (2022 edition). Technical report, International Council for Commercial Arbitration, 2022. URL https://cdn.arbitration-icca.org/s3fs-public/document/media_document/ICCA-reports-no-6-icca-nyc-bar-cpr-protocol-cybersecurity-international-arbitration-2022-edition.pdf.

International Bar Association (IBA). Iba rules on the taking of evidence in international arbitration, 2020. URL https://www.ibanet.org/MediaHandler?id=def0807b-9fec-43ef-b624-f2cb2af7cf7b. Adopted 17 December 2020.

International Bar Association (IBA) and Center for AI and Digital Policy (CAIDP). The future is now: Artificial intelligence and the legal profession. Technical report, International Bar Association, 2024. URL https://www.ibanet.org/document?id=The-future-is+now-AI-and-the-legal-profession-report.

International Chamber of Commerce (ICC). Icc arbitration rules (2021), 2021. URL https://iccwbo.org/wp-content/uploads/sites/3/2020/12/icc-2021-arbitration-rules-2014-mediation-rules-english-version.pdf.

Zachary C. Lipton. The mythos of model interpretability, 2016. URL https://arxiv.org/abs/1606.03490. arXiv:1606.03490.

Xiaogeng Liu, Zhiyuan Yu, Yizhe Zhang, Ning Zhang, and Chaowei Xiao. Automatic and universal prompt injection attacks against large language models, 2024. URL https://arxiv.org/abs/2403.04957. arXiv:2403.04957.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. Prompt injection attack against LLM-integrated applications, 2023. URL https://arxiv.org/abs/2306.05499. arXiv:2306.05499.

London Court of International Arbitration (LCIA). Lcia arbitration rules 2020, 2020. URL https://www.lcia.org/Dispute_Resolution_Services/lcia-arbitration-rules-2020.aspx. In force from 1 October 2020.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229. ACM, 2019. doi: 10.1145/3287560.3287596. URL https://arxiv.org/abs/1810.03993.

National Institute of Standards and Technology. Artificial intelligence risk management framework (AI rmf 1.0). Technical Report NIST AI 100-1, U.S. Department of Commerce, 2023. URL https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. doi: 10.1038/s42256-019-0048-x. URL https://www.nature.com/articles/s42256-019-0048-x.

Silicon Valley Arbitration & Mediation Center (SVAMC). Svamc guidelines on the use of artificial intelligence in arbitration (first edition), 2024. URL https://svamc.org/wp-content/uploads/SVAMC-AI-Guidelines-First-Edition.pdf. Issued 30 April 2024.

United Nations. Convention on the recognition and enforcement of foreign arbitral awards (new york, 1958), 1958. URL https://uncitral.un.org/sites/uncitral.un.org/files/media-documents/uncitral/en/new-york-convention-e.pdf. Adopted 10 June 1958; entered into force 7 June 1959.

United Nations Commission on International Trade Law (UNCITRAL). Uncitral model law on international commercial arbitration (1985), with amendments as adopted in 2006. Technical report, United Nations, 2006. URL https://uncitral.un.org/sites/uncitral.un.org/files/media-documents/uncitral/en/06-54671_ebook.pdf.

United Nations Commission on International Trade Law (UNCITRAL). Uncitral secretariat guide on the convention on the recognition and enforcement of foreign arbitral awards (new york, 1958). Technical report, United Nations, 2016a. URL https://uncitral.un.org/sites/uncitral.un.org/files/media-documents/uncitral/en/2016_guide_on_the_convention.pdf.

United Nations Commission on International Trade Law (UNCITRAL). Uncitral notes on organizing arbitral proceedings (2016). Technical report, United Nations, 2016b. URL https://uncitral.un.org/sites/uncitral.un.org/files/media-documents/uncitral/en/arb-notes-2016-e.pdf.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2): 841–887, 2018. URL https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf.

World Wide Web Consortium (W3C). Prov-dm: The prov data model, 2013. URL https://www.w3.org/TR/prov-dm/. W3C Recommendation, 30 April 2013.

Fangzhou Wu, Ethan Cecchetti, and Chaowei Xiao. System-level defense against indirect prompt injection attacks: An information flow control perspective, 2024. URL https://arxiv.org/abs/2409.19091. arXiv:2409.19091.