Kluwer Arbitration Blog

What is Constitutional AI and Why Does it Matter for International Arbitration?

Sophie Nappert (3 Verulam Buildings & ArbTech) and Fernanda Carvalho Dias de Oliveira Silva, Benjamin Malek · Saturday, June 7th, 2025

Large language models ("LLMs") are increasingly reshaping legal work. According to the 2024 Wolters Kluwer Future Ready Lawyer Survey Report, 76% of in-house legal departments and 68% of law firms now use artificial intelligence ("AI")–powered tools at least once a week.

In arbitration, AI has an emerging role in various stages of the process—from dispute prevention to arbitrator selection, procedural management and drafting. These developments promise to improve efficiency and reduce costs. However, they also raise important concerns about accuracy, bias, and accountability—see, for example, previous blog posts regarding risks of evidence manipulation and the use of deepfake videos in the context of remote hearings.

One noteworthy proposal for mitigating such AI-related risks is "Constitutional AI" ("CAI"), introduced by researchers at Anthropic. CAI embeds a set of clearly defined normative principles —akin to a constitution—into the AI model itself, shaping how it responds to user prompts.

In the context of international arbitration, this built-in framework could help align AI tools with core procedural values such as fairness, neutrality, accountability, and transparency. This article explains the key features of CAI, how it differs from conventional LLMs, and its potential implications for arbitration. It also forms part of a broader editorial initiative exploring the future of AI in arbitration—details on how to contribute are included below.

CAI Briefly Explained

Over the last few years, there have been several attempts to train AI systems in a way that reduces harm. One of those attempts, reinforcement learning from human feedback ("RLHF"), relied on humans to supervise all aspects of the AI's behaviors. But this dependence creates challenges: as pointed out by Anthropic researchers, RLHF requires tens of thousands of human feedback labels that often remain private, and even when shared, are difficult to interpret at scale. The approach is also costly, time consuming, and often impractical as AI models take on tasks that surpass human capabilities. Moreover, systems trained in this way can become overly evasive, refusing to answer controversial questions without offering any explanation.

CAI offers a different path. It trains AI models to be helpful, harmless, and honest without relying

extensively on human feedback at every step of the training process. CAI replaces much of the human oversight required during training with a written set of normative principles—the "constitution"—which the AI uses to guide its behavior. These principles, explained in plain language (like "avoid harmful content" or "be respectful"), serve as internal guardrails. Rather than being told what is right or wrong in thousands of individual examples, the model learns to assess its own responses in light of these overarching principles.

The training process begins with the AI model being presented with challenging or intentionally harmful prompts (e.g., hateful content or instructions on illegal activities), known as "red teaming." The AI's responses are then evaluated against its constitution. Whenever a response conflicts with the guiding principles, the model is shown how to adjust its answer. By comparing the original and revised outputs, the AI learns why the change was needed and how to avoid the undesired approach in the future. A few carefully chosen examples further reinforce this revision process. The model also undergoes a fine-tuning process so that it naturally produces constitution-aligned responses whenever prompted. Later stages of training involve a secondary AI model that helps identify and reward the best constitution-aligned responses. This step further reduces reliance on human reviewers. In some cases, the AI is also prompted to outline its reasoning step-by-step, often referred to as its "chain-of-thought," enhancing the transparency of its decisions. Further explanation about this process can be found here, here, and here.

Integrating CAI into International Arbitration

As mentioned above, CAI has the potential to strengthen key principles of international arbitration. The following examples show how this could happen.

Embedding Arbitration Principles into AI Systems: The Example of Evidence

One of CAI's key contributions is its capacity to serve as a higher regulatory layer within AI systems, instilling them with a foundational understanding of arbitration principles.

Consider tasks like procedural or evidentiary summaries. An untrained AI model may inadvertently draw legal conclusions or assume witness credibility, functions that are traditionally reserved for the arbitral tribunal. For instance, a conventional AI model might summarize a witness testimony as follows: "Witness A testified that Party B knowingly manipulated financial figures during the meeting, directly supporting the fraud allegations."

In contrast, a CAI-trained model can be guided to maintain neutrality, clearly attribute contested assertions, and distinguish between fact and interpretation. Instead of implying wrongdoing, it could produce a safer summary as follows: "Witness A stated that during the March 5 meeting, Party B presented financial data inconsistent with earlier reports. Witness A expressed concern these discrepancies may have been intentional. This testimony may be relevant to Party A's allegations but does not by itself establish wrongdoing."

To achieve this kind of balanced output, the CAI framework is embedded directly into the AI model's training. The framework includes principles such as avoiding adversarial or accusatory language, clearly attributing contested assertions, and refraining from making judgments or

expressing opinions. The AI is also guided to separate facts from interpretation and remain within procedural bounds. For example, when handling evidentiary material, the AI could flag contradictions without speculating on intent, present conflicting views without favoring one over the other, and avoid implying that any fact is decisive unless the tribunal has expressly found it to be so.

Personalizing Fairness and Enhancing Party Autonomy

The CAI framework could enable parties to jointly define a "constitution" comprising procedural principles, jurisdiction-specific rules, or ethical guidelines that reflect their shared expectations for fairness. This constitution could include principles on how to handle evidentiary inconsistencies, when to flag—but not resolve—contradictions, or how to maintain neutrality when summarizing disputed facts. By doing so, the AI can operate transparently and predictably, with its outputs traceable to a set of principles agreed upon by the parties, rather than emerging from an opaque, unaccountable system.

This "constitution" can be applied during the supervised fine-tuning phase, where the AI model learns to revise its outputs in line with the agreed principles, or during real-time use through structured prompts and chain-of-thought reasoning. Where the parties wish to refine the AI's behavior further, they can provide annotated datasets that exemplify acceptable treatment of issues such as confidentiality, tone, or procedural framing. This allows the AI to learn from their specific preferences while still adhering to common core values.

This process would start by setting clear governing principles, phrased as instructions for the CAI. The parties would then choose a capable base AI model (an "instruction-tuned LLM") and train it using examples of arbitration tasks, refining its responses to strictly follow those agreed-upon principles. This training (Supervised Learning, or SL-CAI) would teach the AI to follow the agreed rules closely, akin to showing best practices to a new junior lawyer using carefully selected case examples and precedents.

Next, the CAI would learn to evaluate and improve its own outputs by comparing potential responses against the constitutional principles. This could be done through a reward system that would encourage the CAI to generate responses that best adhere to the desired legal standards. Such reward systems include Reinforcement Learning from Human Feedback, a technique used to provide human feedback or evaluate the actions of another AI agent with a reward model, and Proximal Policy Optimization, which aims to stabilize and improve the training process by taking small, controlled steps when updating the agent's policy.

It is important to stress that, as things currently stand, there are fundamental tenets of the dispute resolution process that are likely to still require human judgment for the foreseeable future. For example, the Oxford Internet Institute's researchers point out that fairness cannot be fully automated. Still, CAI may enable a safer use of LLMs in arbitration in a way that engages parties with the technology and builds procedural fairness.

Regarding access and feasibility to this kind of model, further thought needs to be given to ways in which the arbitral process can accommodate ease of access to high quality LLM models.

Harmonizing International Procedural Standards

A shared AI "constitution" can help ensure consistency in how AI tools are used in arbitration across different legal systems. By encoding procedural norms, institutional rules, and jurisdiction-specific standards into the AI model's behavior, CAI also enables AI systems to anticipate and mitigate potential conflicts.

Practically, the model could be trained on institutional rules (such as ICC, LCIA, or UNCITRAL), applicable legislation, and sample procedural documents. With a structured critique-and-revision process, the AI would learn to adjust its responses in line with those norms.

For example, if the AI suggests disclosing a sealed or privileged exhibit, a CAI-trained system—guided by a principle such as "disclosure restraint"—would flag the issue, revise the output, and explain the reasoning for rejecting disclosure. Similarly, by being exposed to protective orders or confidentiality clauses through prompting, the AI can learn to identify when a request risks violating privacy obligations and respond in a non-evasive but principled manner.

CAI can also be prompted to reveal its reasoning through step-by-step "chain-of-thought" outputs, allowing human reviewers to understand not just what the AI recommends, but why. Each step in the reasoning process can be compared against the underlying principles encoded in the AI's constitution, enabling a form of second-order review. This traceable logic ensures that legal judgment remains grounded in an interpretable, procedural context rather than opaque or unexamined conclusions.

Conclusion and Call for Papers

As AI continues to reshape arbitration, CAI presents an opportunity to embed procedural fairness, transparency, and party autonomy into the fabric of AI-assisted decision-making. Of course, its implementation raises complex questions around regulation, accountability, and the evolving role of human oversight in arbitral processes.

To further explore these and other pressing issues involving AI and arbitration, **Kluwer Arbitration**, in collaboration with Editors **Sophie Nappert, Benjamin Malek**, and **Fernanda Carvalho Dias de Oliveira Silva**, invites scholars and practitioners to contribute to our upcoming **AI and Arbitration 2025** publication. We seek forward-thinking papers that critically assess the intersection of AI and arbitration, addressing challenges, ethical concerns, and future possibilities.

Interested contributors should submit a **400-word proposal** (original, non-AI-generated) to **aiandarbitration@outlook.com**. We look forward to your ideas as we shape the future of arbitration in the age of AI.

This entry was posted on Saturday, June 7th, 2025 at 12:55 pm and is filed under Arbitration, Artificial Intelligence, Constitutional AI, Technology
You can follow any responses to this entry through the Comments (RSS) feed. You can leave a response, or trackback from your own site.