(Audience Handout 04)

Legal argumentation before the Robot Judge: the need for realistic regulation

BRAZIL-CHINA LEGAL DIALOGUE ON ARTIFICIAL INTELLIGENCE
Panel II:
Building an AI Governance: Regulatory Frameworks and Institutional Models in Brazil and China
Topic: AI regulation and legislatory strategies
Sept 09, 2026

Víctor Gabriel Rodríguez

Professor of Criminal Law, University of São Paulo (USP), Brazil
Professor of Ph.D Program of Latin American Studies, USP.
Ph.D. in Public Law, Universidad de Valladolid, Spain
victorgabriel@usp.br

INTRODUCTION

My current research is being carried out at the Institute of Artificial Intelligence and Machine Learning (CIAAM) at the University of São Paulo and Ibrachina-Ibrawork and NACL (Neuro-Artificial Criminal Law Program), with both short-term and long-term objectives.

A long-term goal is the **idea of a robot judge**: not a system designed to replace human judges, but a tool that could make judicial decisions more.

However, the more immediate challenge is how to **regulate the use of AI in the judiciary**. More specifically, we are trying to develop the minimum standards for legal argumentation before a robot judge. In my point of view, this three steps are necessary:

1) To ensure that the legal community has a clear and accurate understanding of how the AI systems used by the judiciary operate.

2) To recognize, as a legal community, that legal decisions are already being made exclusively by AI systems (the "robot judge")

3) To develop a regulatory framework that recognizes lawyers' right to direct their arguments to AI systems rather than to a human judge.

.

BASIC PREMISES:

Speaking of practical issues, I would like to begin with a recent Brazilian case that attracted attention in all Latin America.

Two lawyers embedded hidden prompts into their legal brief. Invisible to the human eye, these instructions were designed to influence either the court's AI system or the AI used by the opposing party. Among other things, they instructed the system to ignore certain arguments.

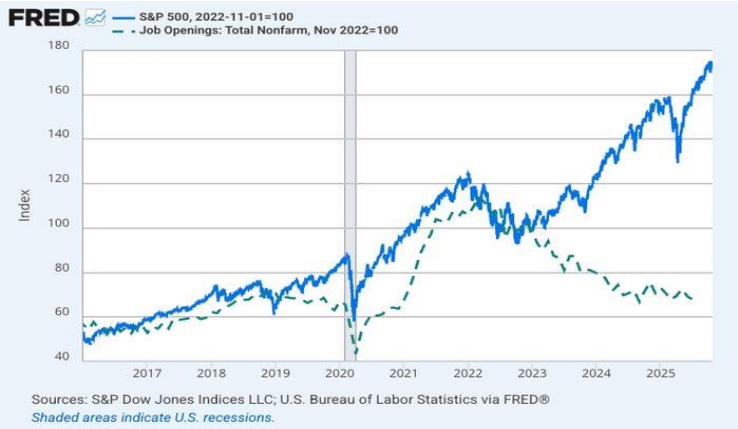
The lawyers were immediately sanctioned by the Brazilian Bar Association, and the media described the case as an act of "**sabotage**."

Perhaps it was. But I am not convinced that the legal question is so simple.

From the perspective of legal argumentation, the more important issue is that we still lack a **regulatory framework** capable of distinguishing legitimate persuasion from impermissible manipulation when the audience is no longer a human judge, but an Artificial Intelligence system.

This, in my view, is one of the central challenges for the next generation of AI regulation.

So, we need a basic assumption: today, all of us use Artificial Intelligence in one way or another. The labor market provides the clearest evidence of its massive adoption:



For our discussion, however, only one conclusion matters.

Judges, Lawyers, Law Firmes are already using **generative AI** but we do not have a clear regulatory framework for this new reality.

PART ONE: HOW GENERATIVE AI REALLY WORKS IN THE JUDICIARY

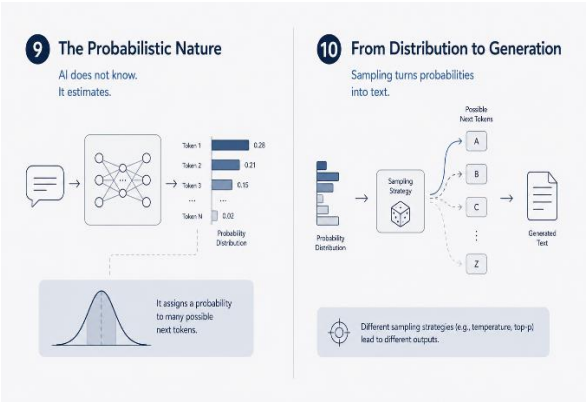
The first misunderstanding I would like to address concerns generative AI itself.

Before generative AI, Brazilian courts already used Artificial Intelligence. Those systems were regulated, but they were not generative. Today, systems such as **ApoIA** introduce a completely different type of technology

So let me begin with a very simple question: what is generative AI?

The first point is that generative AI no longer works like a traditional if-then algorithm. Instead, it works through probabilities.

Large Language Models process language by breaking it into tokens. Each token is represented as a mathematical vector, called an *embedding*. Those embeddings interact with one another through the transformer architecture, producing predictions about the next token.



In simple terms, a Large Language Model is an enormous probability machine.

What is the main consequence of this? The answer is simple.

As these systems become more powerful, they also become less predictable, less traceable, and much harder to explain.

This is what Dario Amodei and Anthropic call the **black box**[1], with important consequences for regulation**.**

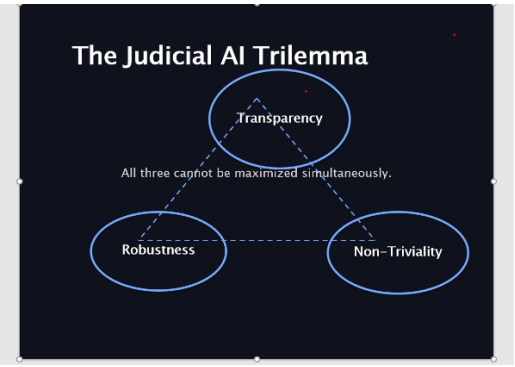
For many years, AI regulation assumed that systems could be made sufficiently transparent.

For example, Article 13 of the EU AI Act states that "high-risk AI systems should be **sufficiently transparent** to allow users to interpret their outputs"[2].

The first Brazilian regulation adopted a very similar idea.

But once we accept that Large Language Models are black boxes, complete transparency is no longer a realistic regulatory objective.

The era of **'sufficient transparency'** is over. Our regulatory system must assume that.



[1] "However, for the most part, the mechanisms by which they do so are unknown. **The black-box nature** of models is increasingly unsatisfactory as they advance in intelligence and are deployed in a growing number of applications. Our goal is to reverse engineer how these models work on the inside, so we may better understand them and assess their fitness for purpose." (Lindsey et al., 2025)

[2] Art. 13 High-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately. An appropriate type and degree of transparency shall be ensured with a view to achieving compliance with the relevant obligations of the provider and deployer set out in Section 3.

PART TWO: AI IN THE COURTS

Now let us look at how judicial AI actually works in Latin America, now specifically in Brazil.

Brazilian courts use several AI systems. Their legal function is for now limited to **assisting judicial work**. Among these systems, the most important is the already mentioned **ApoIA**.

Its purpose is to help judges read the parties' submissions, summarize case files, **draft the factual report** of judgments (in portuguese, “relatorios”), and prepare summaries of appellate decisions.

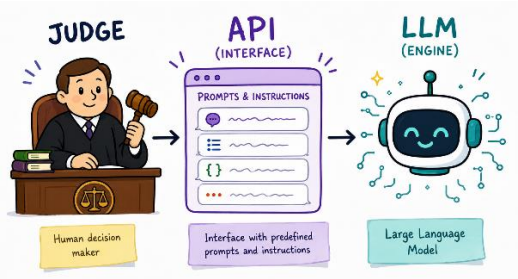
We do not know every technical detail behind Apoia. But one important fact is publicly known, and even acknowledged by the system itself: its **main engine** is a Large Language Model.

In other words, systems such as Apoia are essentially APIs.

As I mentioned before, an API is not itself an Artificial Intelligence. It is an interface that connects a judicial application to a Large Language Model through a series of predefined prompts and instructions.

This point is very important.

When we hear that "our Court has developed its own Artificial Intelligence," this is not entirely accurate. The court has developed the **API**. The Large Language Model remains the engine behind the system.



This brings to us two important consequences.

(**1**) The first, as we discussed earlier, is that the system is built on a black box. Therefore, complete transparency is impossible.

For this reason, the most recent Brazilian regulation no longer requires full transparency:

> § 6. Whenever generative AI is used to **assist in the drafting** of a judicial act, such **use may, at the judge's discretion, be mentioned in the text of the decision.** In any event, the use of such technology shall be automatically recorded in the court's internal system for the purposes of generating statistics, monitoring, and potential auditing.

(**2**) The second consequence is even more important.

Judges do not have access to the prompts embedded in the API. Maybe some of them believe they are talking directly to the AI. In reality, every request first goes through a series of hidden prompts and filters designed by the court before it reaches the Large Language Model.

As a result, whenever a judge asks the system to perform a task, the system already introduces choices that shape the final outcome. This is why I maintain that a robot judge is already operating within this type of judicial system in Latin America.

ARE WE ALREADY DEALING WITH ROBOT JUDGES?

Although the law does not formally allow it, I believe that a form of robot judge **is already operating** in every judicial system that uses generative AI for so-called "auxiliary tasks."

Brazilian regulation is very clear on this point. Resolution 615 of the National Council of Justice states that Artificial Intelligence cannot decide a case.

> II – the use of these tools shall **be solely auxiliary and complementary**, functioning as decision-support mechanisms. They shall not be used as autonomous judicial decision-making tools without the judge's appropriate oversight, interpretation, verification, and review. The judge shall remain fully responsible for every judicial decision rendered and for the information contained therein;

However, this is not exactly what happens in practice.

The Apoia system drafts the factual report (relatório) of the judgment by itself. In other words, it reads the parties' submissions, analyzes the evidence, and prepares the first factual narrative of the case. And its already a judicial decision, for two reasons:

(**1**) The first is **selection**.

Every text is a selection of reality.

Reality (the world itself) contains an infinite number of facts, but every text includes only some of them. Writing (or work with any kind of text, including images, as painting) always means **selecting**.

The same happens with Artificial Intelligence before the parties' arguments.

I like to explain this with an example from literature.

Ernest Hemingway developed what is known as the **Iceberg Theory**.

According to Hemingway, the writer shows only a small part of the story. The reader reconstructs everything else.

The important idea is that whoever decides which part of the iceberg is visible also influences the reader's understanding of the whole story.

Exactly the same happens with a judicial report.

The AI system decides which part of reality appears in the first narrative presented to the judge.

(**2**) But there is a second reason, and it is even more important. And I will use narrative theory to explain it as well.

The statement is this: The beginning of a text **shapes** everything that follows.

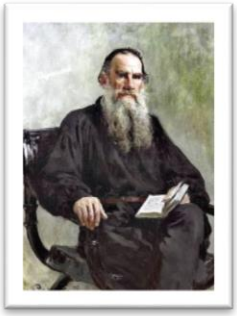
In literature, people often say that a great novel is built on its first sentence, its opening line. It lays the foundation for everything that comes after.

Ernst Hemingway himself gave a very simple piece of advice:

"Write the truest sentence that you know, and then go on from there"

An entire book grows from that first truth. The same happens with a judicial decision. If an AI system gives the judge the first version of the facts, that primitive text already influences (or determine) everything that comes next.

So, if changing the report changes the natural direction of the decision, then drafting the report is already part of deciding the case. A factual report cannot be regarded as merely "auxiliar, it is already a determinative part of the judicial decision.



This is very close to what social psychology calls the **anchoring effect,** or anchoring bias.

That is why I believe that, in practice, we already have a form of robot judge. Any attempt to regulate judicial AI must start from this reality.

PART THREE: REGULATING LEGAL ARGUMENT BEFORE AI

If my previous argument is correct, we immediately face a new question.

How should we regulate legal argument when an AI system already participates in the judicial decision?

This question takes us back to how generative AI actually works.

Although a Large Language Model communicates in natural language, it does not think like a human being. It just imitates natural language, as we have seen before, through a complex process of statistical relationships between words.

If that is true, it is reasonable to believe that the arguments most persuasive to an AI system may not be the same as those that persuade a human judge.

This is, in my opinion, one of the greatest regulatory challenges we face today: how to persuade an AI.

The first mistake is to assume that every attempt to persuade an AI system is manipulation.

I do not believe that.

To explain why, I like to use a simple example from chess.

Currently, no human player can defeat the strongest chess engines, such as Stockfish or AlphaZero.

But in 1995, the situation was different.

Garry Kasparov, the World Chess Champion, lost his first game against IBM's Deep Blue, but he did not give up. By understanding how the machine “thinks”, he was able to win the following games. After all, Deep Blue was not a human opponent. It had its own logic.

Now let me ask a simple question: **Did Kasparov cheat?**

Was it unfair to learn how the machine worked? Or was he simply adapting his strategy to the nature of his opponent? This, in my opinion, is exactly the question we **now** face in the judiciary.

If lawyers know that an AI system participates in judicial decision-making, should they be forbidden from understanding how that system processes legal arguments?

This is what I call **Decision Architecture**.

In my view, we must distinguish between **Decision Architecture** and **Adversarial Attacks**. Decision architecture includes many other techniques, some of which may be entirely legitimate. In my view, the case that opened this presentation, **involving the lawyers who used prompt injection in Brazil**, lies precisely on the borderline between an "attack" and a legitimate form of decision architecture.

To illustrate the point, let me use a slightly exaggerated comparison.

Completely prohibiting Decision Architecture would be **like requiring a lawyer to have a conversation with a microwave oven or a toaster**, even though everyone knows that, to make the machine do what it is designed to do, all that is necessary is to press the right button.

This brings us to three regulatory questions.

(**1**) First, **where** should **we draw the line between** legitimate Decision Architecture and prohibited manipulation?

(**2**) Second, how should we define an Adversarial Attack?

(**3**) Third, what sanctions should apply when that line is crossed?

At the moment, we face the worst possible scenario:

a)We already have a robot judge.

b) We already have Shadow AI.

c) Law firms are already optimizing legal arguments for AI systems.

d) Yet our legal rules still assume that every legal argument is **addressed only to a human judge**, and that AI serves merely as a supporting tool in judicial decision-making.

That legal fiction no longer reflects reality.

PART FOUR: OUR PROPOSAL — AN INTEGRATED ROBOT JUDGE

Let me conclude by presenting our proposal.

It is based on two forms of realism: technological realism and the practical needs of Latin American judicial systems.

We call it the **Integrated Robot Judge**[3].

[3] I had the opportunity to present the first version of this proposal here at ECUPL a few months ago.

The AI system should read the parties' submissions, evaluate the evidence, and draft the entire judgment, including both its reasoning and its final disposition.

The case is then returned to the human judge, who decides whether to adopt or reject the AI's proposed decision.

The model has three stages.

Stage 0. Lawyers submit their arguments exactly as they do today.

Stage 1. The AI analyzes the pleadings, the evidence, and the case file, and prepares a complete draft judgment.

Stage 2. The AI draft is published. Both the parties and the human judge receive exactly the same document.

Stage 3. The human judge makes the final decision.

In this model, the robot judge is not a private assistant to the human judge.

It becomes an institutional mechanism for transparency, consistency, and impartiality, because it emerges from a decision-making process that, although automated, is, so to speak, free from the individual subjectivities of the judge.

This is the long-term vision behind our project.

KEY TAKEAWAYS

1. It is no longer possible to prevent judges and lawyers from using Artificial Intelligence. Attempting to do so would simply increase the use of Shadow AI.
2. Most judicial "AI systems" are actually APIs connected to powerful commercial Large Language Models.
3. These systems suffer from two distinct transparency problems. The first concerns the hidden prompts and other forms of API-level fine-tuning. The second arises from the black-box nature of the underlying large language model (LLM).
4. Lawyers should not be required to pretend that they are arguing only before a human judge when they know that an AI system already participates in the decision-making process.
5. AI regulation must clearly distinguish legitimate Decision Architecture from forbidden manipulation and Adversarial Attacks.

6. A robot judge that produces complete draft judgments and makes them available to both the parties and the human judge could become an important institutional mechanism for judicial impartiality.
7. Our mission is to develop a **Code of Ethics** for the legal profession that addresses the reality of **Optimization for AI** while recognizing the emerging risks of manipulating justice. However, this requires more than simply prohibiting hidden architectures.
8. Finally, before regulating AI in the judiciary, we must first answer a more fundamental question:

Which characteristics of the human judge are truly indispensable for the administration of justice?

We must define these essential characteristics not to dispense with the human judge. On the contrary, we must identify precisely those capacities in which the human judge is genuinely irreplaceable. That's our goal.

Thank you,

Víctor

***