Towards Responsible AI Governance in the Brazilian Judiciary

Bruno Fonseca
University of São Paulo
bcastrodias@usp.br

Abstract

This paper examines the governance of artificial intelligence in the Brazilian judiciary, analyzing over 140 AI systems implemented under the National Council of Justice's Justice 4.0 program and the evolution from principle-based to prescriptive regulatory frameworks. While Brazil has advanced toward international standards of responsible AI, persistent gaps in oversight, accountability, and ethical safeguards highlight the challenges of aligning judicial automation with principles of justice, fairness, and transparency.

Artificial Intelligence (AI) systems are increasingly being used in judicial systems worldwide, including Brazil, where over 140 AI tools have been implemented since the early 2020s under the National Council of Justice's (NCJ) "Justice 4.0" program. These systems are used for tasks like document classification, case outcome prediction, and procedural triage, with notable examples being Victor (Supreme Federal Court) and Athos (Supreme Court of Justice), which are used for tasks like legal document classification, jurisprudence recommendation, and workload triage. While promising increased efficiency, these tools raise concerns about transparency, fairness, and accountability.

AI deployment introduces risks such as bias amplification (Barocas and Selbst 2016), opacity in decision-making (Burrell 2016), and public trust issues (Zuboff 2019). In the judicial context, AI can unintentionally reinforce discriminatory patterns or limit judicial discretion. Addressing these challenges requires not only technical safeguards but robust governance frameworks that consider legal and ethical constraints. Concerns about AI non-alignment—when systems operate in ways that don't align with human values or judicial norms—are also critical (Binns 2018).

Responsible AI, which emphasizes transparency, accountability, fairness, and human oversight, is essential to maintaining procedural fairness and institutional credibility in the judiciary (Jobin, Ienca, and Vayena 2019; Floridi et al. 2018). AI systems must not only perform tasks efficiently but also align with the values of justice and impartiality.

A national survey conducted by the NCJ in 2023 highlights responsible AI practices, based on 140 AI models reported. The most frequently supported practice is promoting user awareness about AI (102 responses, 73% of total), followed by training public servants and judges (90 responses, 64%). Respondents also emphasized the importance of clear guidelines and policies (74 responses, 53%) and mechanisms for AI explainability (56 responses, 40%). Public availability of source code and system functions was considered important by 25 respondents (18%). Fewer respondents supported creating a dedicated IT department for ethical concerns (5 responses, 4%) or providing models and training datasets (3 responses, 2%). These findings highlight a strong focus on transparency, education, and ethical considerations but also reveal concerns about AI systems aligning with judicial norms. Addressing these concerns requires governance that balances technical benefits with ethical implications, ensuring AI supports, rather than replaces, human judgment.

Based on the initial stage of the research, which compared the evolution of NCJ resolutions, Brazil's judicial AI governance has progressed from the principle-based NCJ Resolution 332/2020 to the more prescriptive Resolution 615/2025, introducing robust governance mechanisms such as a model risk-classification system, privacy-by-design and privacy-by-default requirements, secure data curation, mandatory algorithmic impact assessments, and contractual and operational rules for generative AI. The new framework reinforces human oversight throughout the AI lifecycle, mandates standardized public reporting, and involves additional institutions like the Bar Association and Public Prosecutor's Office, signaling convergence with international responsible AI standards.

Despite these advances, significant gaps persist: external audits remain at the discretion of courts without binding civil society participation; technical criteria are vague; there are no standardized metrics for impact, explainability, or bias mitigation; and enforcement lacks clear sanctions or graduated compliance mechanisms. The risk-classification system underestimates the role of auxiliary legal services in judicial decision-making, while environmental impacts and ethical benchmarks are entirely absent, limiting the development of a culture of algorithmic integrity.

This research examines how responsible AI is being implemented in Brazil's judiciary by analyzing the governance strategies, technical infrastructures, and legal frameworks that guide its development. As one of the most institutionally advanced cases of AI use in the public sector globally, the Brazilian judiciary offers a unique opportunity to explore how automated systems interact with legal and ethical norms. The study employs document analysis, semi-structured interviews with key stakeholders, and case studies of systems such as *Victor* and *Athos.*

Focusing on both the techniques and limits of applying responsible AI, the research investigates how principles such as justice, fairness, transparency, and human oversight are translated into practice. It pays particular attention to the risks posed by misalignment, including the reinforcement of existing biases, reduction of judicial discretion, and potential conflicts with constitutional values. The next empirical phase of the research involves conducting in-depth, semi-structured interviews with judges, court administrators, and high-level officials from Brazil's judicial councils, including the National Council of Justice (NCJ) and regional appellate courts. These interviews aim to capture institutional perceptions, implementation challenges, and normative expectations surrounding the use of AI in judicial decision-making and administration. By engaging directly with key actors responsible for developing, deploying, or overseeing AI systems, the study will enrich its analysis of how responsible AI principles are interpreted and operationalized in practice, and will identify institutional tensions, gaps in oversight, and opportunities for improvement in AI governance within the judiciary.

By grounding these issues in Brazil's experience, the study contributes to global academic and policy debates on AI governance and provides insights into the real-world application of responsible AI in complex institutional settings.

References

Barocas, S.; and Selbst, A. D. 2016. Big Data's Disparate Impact. *California Law Review* 104(3): 671–732. doi.org/10.2139/ssrn.2477899.

Binns, R. 2018. On Being "Responsibly Aligned": The Ethics of AI. *Ethics and Information Technology* 20: 81–98. doi.org/10.1007/s10676-018-9457-6.

Burrell, J. 2016. How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms. *Big Data & Society* 3(1). doi.org/10.1177/2053951715622512.

Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* New York: St. Martin's Press.

Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; Schafer, B.; Valcke, P.; and Vayena, E. 2018. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28(4): 689–707. doi.org/10.1007/s11023-018-9482-5.

Jobin, A.; Ienca, M.; and Vayena, E. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence* 1(9): 389–399. doi.org/10.1038/s42256-019-0088-2.

Zuboff, S. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* New York: PublicAffairs.