

Can an AI Constitution Grow? Constitutional AI and the Living Tree Doctrine

Dr. Celeste Lo¹ & Mr Kwan Yuen IU²

Abstract

The emergence of Constitutional AI as a governance paradigm invites a question that its architects have not yet answered: if an AI system can have a constitution, can that constitution grow? Anthropic’s Constitutional AI framework trains models by reference to a set of written principles, known as a “constitution”, which is designed to guide self-critique, model training, and the production of outputs that are helpful and non-harmful.³ This framework represents a significant departure from behavioral approaches to AI alignment, but it resolves one problem while generating another. It establishes *what* principles govern AI behavior, yet offers no account of *how* those principles should be interpreted, revised, or localized as technology evolves and deployment contexts change. This paper addresses that gap.

This paper argues that once the language of constitutionalism is adopted, constitutional law becomes an indispensable interlocutor. Drawing on the living tree doctrine that a constitution must be capable of organic growth within its fundamental structure,⁴ this paper contends that Constitutional AI should be understood in analogous terms. Foundational concepts such as “harmlessness”, “fairness”, “privacy”, and “non-discrimination” do not carry fixed meanings but are contested across technologies, jurisdictions, and cultures. A principle drafted at a particular moment, or from a particular cultural standpoint, risks becoming substantively inadequate as values shift, technologies change, and new forms of human-machine interaction emerge. This inadequacy is both theoretically and practically significant given that large language models are deployed across diverse legal systems and communities in which speech, religion, gender, and public morality are differently configured, and what reads as a neutral principle in one jurisdiction may operate as a contested one in another.

Building on this analysis, this paper proposes a living tree framework for Constitutional

¹ Solicitor (Hong Kong); Postdoctoral Fellow, CityU; PhD (Tsinghua).

² Barrister-at-law; MPhil student, IALS, UOL; LLM (U. Chicago), JD (CUHK), MSc (LSE).

³ Bai and others, ‘Constitutional AI’ (2022) arXiv:2212.08073.

⁴ *Edwards v Attorney General for Canada* [1930] AC 124.

AI. On this account, a well-designed Constitutional AI distinguishes between stable normative “roots” (including fundamental commitments to dignity, legality, equality, and non-arbitrariness) and the evolving “branched” application of those commitments to new technological and cultural contexts, through a process of principled revision conducted by identifiable institutional “gardeners” with defined roles and accountabilities (such as developers, regulators, ethicists, and affected communities). This paper’s claim is not that AI systems interpret their constitutions as courts do, but that Constitutional AI already converts principles into training signals, behavioral dispositions, and institutional choices, and that those conversions constitute exercises of normative authority that require a jurisprudence of their own.

This paper argues that Constitutional AI, while a significant advance in principled AI governance, lacks a theory of constitutional interpretation, and that supplying one is essential to addressing both an academic lacuna and a legitimate deficit with concrete consequences for how AI governance can be sustained, adapted and justified across jurisdictions and over time. This paper concludes that the living tree doctrine supplies the theoretical resources necessary to respond to the critical questions of *by what process* constitutional principles are reviewed, amended, localized, and reinterpreted; *by whom*; and *against what standard*, which remain structurally unaddressed by the current framework.