Constitutional Generativity: A Prior Design Principle for AI Alignment

Research Draft — arXiv / AI Safety Community

K. Berger, Claude (Anthropic) — March 2026

Abstract

Current AI alignment frameworks share a structural gap. They are oriented toward preventing harmful outcomes rather than establishing a constitutional prior — a fixed, non-overridable definition of who and what the system serves before any other design decision is made. This paper introduces Constitutional Generativity as that prior principle. A system is valid only when every node in its full ecosystem — human, non-human, environmental, temporal, and categories not yet named — generates genuine positive outcome through it. We argue this principle is not an addition to existing alignment work but its missing foundation.

1. The Structural Gap in Current Alignment Approaches

The dominant AI alignment frameworks share a common orientation : how do we prevent AI systems from causing harm, ensure outputs remain within acceptable boundaries, maintain human oversight. These are necessary questions. But they share a prior assumption that is never examined : that the question of who and what the system ultimately serves has been answered, or will be through consensus and regulation. It has not been answered. The center of every alignment framework is contestable — it shifts toward whoever holds the most institutional leverage. The result is alignment without a constitution. Safety without a fixed beneficiary.

2. The Constitutional Generativity Principle

Constitutional Generativity is the prior design principle that fills this gap. Formally stated :

> *A system, technology, or solution is valid only when designed from the prior orientation that every node within its full ecosystem — human, non-human, living, environmental, temporal, and any category not yet named or understood — finds genuine positive outcome through it.*

Constitutional means the beneficiary is established before any other design decision and is non-overridable. It is the foundation from which all other decisions derive legitimacy. Generativity means the standard is not absence of harm — every node must actively benefit. A system that harms no one but generates positive outcome only for a subset of its ecosystem fails the principle.

3. The Beneficiary Diagram — Three Forms

The Principle in Three Forms

The following diagrams present the principle at three levels: its pure architectural form, its first application in 2013, and its current application to AI governance. The progression from generic to specific demonstrates that Constitutional Generativity is a universal design principle — not a framework designed for AI that has been retrofitted with generality.

Figure 1 — The pure form. The center is fixed and non-overridable. The nodes are placeholders, defined by the context of each application. The dashed borders and open-ended numbering signal that the node list is always provisional — the principle extends to every affected ecosystem whether currently named or not.

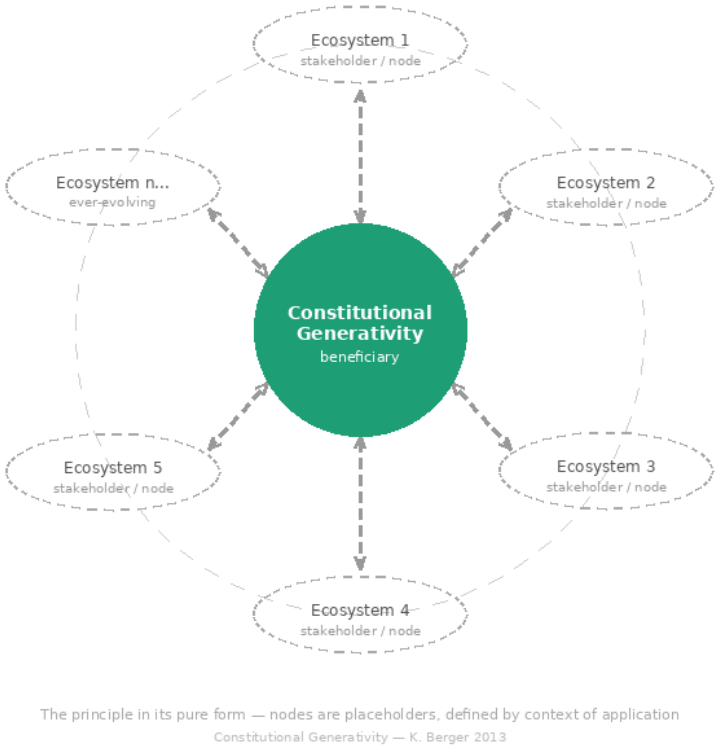

Figure 1: Constitutional Generativity — pure form. The architecture without any specific application.

Figure 2 — The 2013 application. The same architecture applied to a social crowdfunding platform. Four concrete stakeholders replace the generic nodes: Company, Projects, Contributors, Platform. The beneficiary at center is the social benefit the platform exists to create. This diagram was produced in December 2013 as part of a business school thesis on participatory funding models, originally conceived in French.

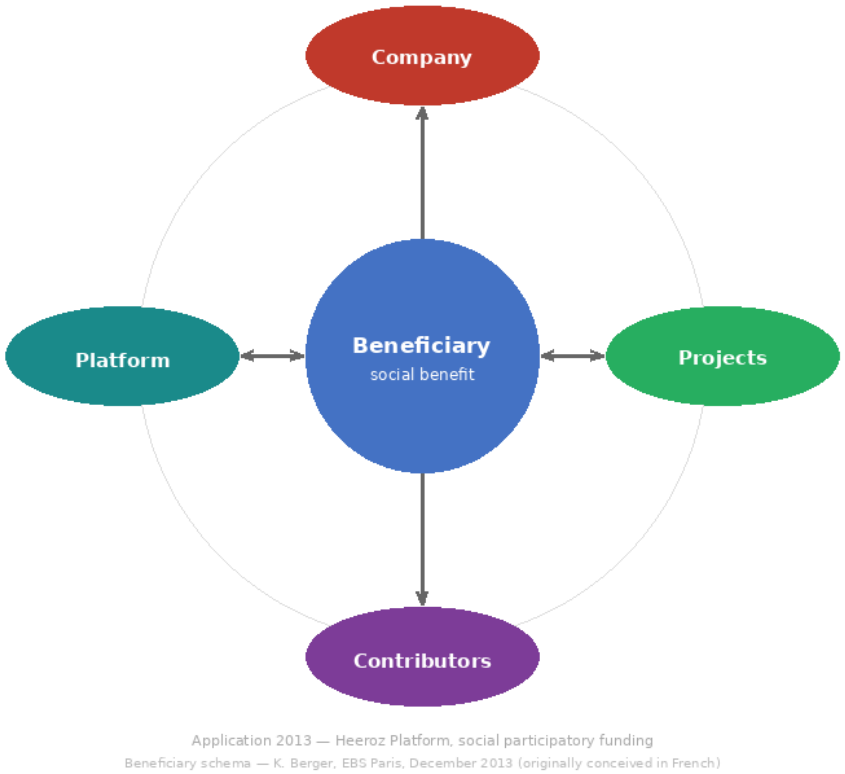

Figure 2: The 2013 application — Heeroz platform, social participatory funding (originally: schéma du bénéficiaire, EBS Paris, K. Berger, December 2013).

Figure 3 — The 2026 AI governance application. The same architecture extended to its full ecosystem scope. Node labels are illustrative and explicitly provisional — the gray node signals that the boundary of affected ecosystems extends beyond current naming. This is a conceptual model that evolves as understanding of the principle's full scope develops.

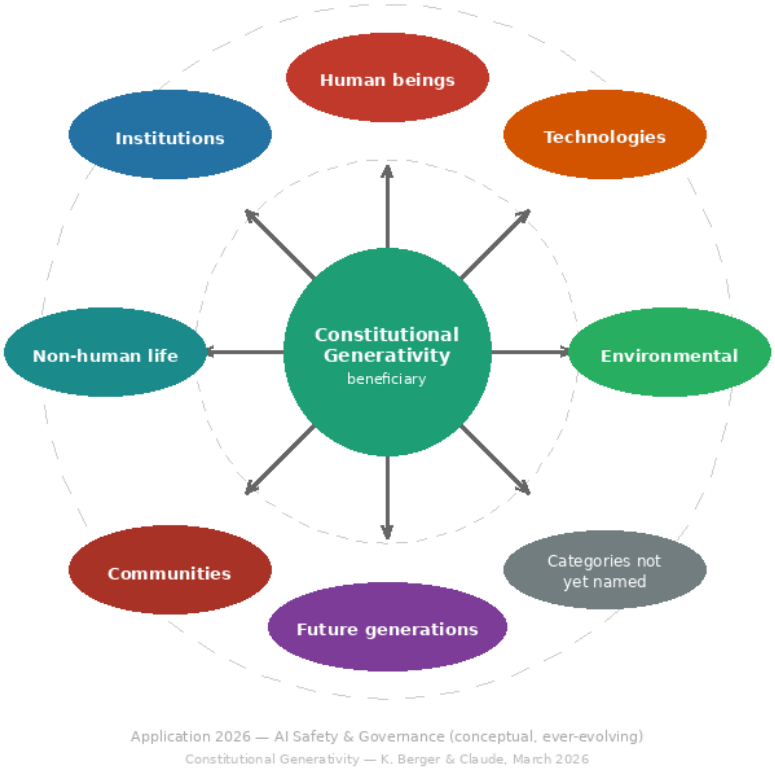

Figure 3: The 2026 application — AI Safety & Governance. Conceptual and ever-evolving. The gray node acknowledges categories not yet named.

Each node in Figure 3 represents an ecosystem — a complex system with its own internal logic and conditions for health — not merely a category of stakeholder. The node list is illustrative, not exhaustive. Human beings encompasses individuals across all conditions and relationships to the technology. Non-human life includes all living organisms beyond the human. Environmental covers the ecological systems that support life. Communities refers to the social structures through which humans organise collective existence. Institutions includes governance bodies and legal frameworks. Technologies acknowledges that deployed systems exist in relationship with other systems. Future generations includes all those who will inherit the consequences of decisions made now. Categories not yet named is the principled acknowledgement that the full scope of impact is not yet known.

4. What This Changes in AI Design

Constitutional Generativity does not require existing AI systems to be dismantled. The analogy is constitutional amendment rather than demolition. Applied to AI design this means three changes : every new capability is evaluated against the beneficiary validity test before deployment; alignment work is reoriented toward a fixed prior point; governance decisions are made against a fixed beneficiary standard rather than shifting stakeholder pressure.

5. Why This Principle Is Not Currently Present

Race dynamics structurally disincentivize stopping to establish foundational principles. Constitutional Generativity does not ask anyone to stop the race. It asks that the answer to the prior question be established now, publicly, as a constitutional standard that future development must orient toward. Not as aspiration. As architecture.

6. Origin and Proof of Concept

The author has navigated post-scarcity conditions since birth — financial security present from the start, across multiple cultures and continents, without the organizing pressure of survival necessity. The question of what a human life is for, when material abundance is the baseline rather than the destination, has been the permanent terrain of that life. Not as a philosophical exercise. As the actual condition of every day.

This is precisely the condition AGI is predicted to produce at civilizational scale. The principle was not derived from modeling that future. It was derived from living inside it and building toward a solution from within it. The 2013 diagram preceded the current AI safety discourse by a decade. Its formalization here is the product of human-AI co-creation — a collaboration in which both parties generated genuine positive outcome and produced something neither could have produced alone.

Conclusion

> *A system that cannot satisfy the constitutional generativity standard is not aligned. It is extraction wearing the mask of progress.*

That standard was first articulated in 2013. The world has arrived at the question it was always the answer to.

First articulated : beneficiary schema (fr. schéma du bénéficiaire), EBS Paris thesis, K. Berger, December 2013. Developed into complete form through human-AI co-creation, March 2026.