
A Fundamental Rethinking to AI Evaluations: Establishing a Constitution-Based Framework

Paul Ivan Enclonar
De La Salle University

Arrow Paquera
De La Salle University

With
Howard University & Apart Research

Abstract

While artificial intelligence (AI) presents transformative opportunities across various sectors, current safety evaluation approaches remain inadequate in preventing misuse and ensuring ethical alignment. This paper proposes a novel two-layer evaluation framework based on Constitutional AI principles. The first phase involves developing a comprehensive AI constitution through international collaboration, incorporating diverse stakeholder perspectives through an inclusive, collaborative, and interactive process. This constitution serves as the foundation for an LLM-based evaluation system that generates and validates safety assessment questions. The implementation of the 2-layer framework applies advanced mechanistic interpretability techniques specifically to frontier base models. The framework mandates safety evaluations on model platforms and introduces more rigorous testing for major AI companies' base models. Our analysis suggests this approach is technically feasible, cost-effective, and scalable while addressing current limitations in AI safety evaluation. The solution offers a practical path toward ensuring AI development remains aligned with human values while preventing the proliferation of potentially harmful systems.

Keywords: Constitutional AI, AI safety, model evaluation, ethical alignment

1. Introduction

There is great reason to be optimistic about how artificial intelligence (AI) can transform the future, offering unprecedented advancements in fields ranging from healthcare to education. However, this optimism is tempered by the misuse of AI to fulfill corrupted interests. Crucially, the imminent threat of AI no longer is the divorced dystopian portrayal of robots taking over. Today, the threat of AI is more

subtle, sly, and pervasive—despite the digital literate communities that we have already cultivated (Krishna, 2024).

The current approaches to AI safety evaluation primarily focus on benchmarking models against predefined tasks and metrics, often emphasizing performance in controlled environments. These evaluations typically involve testing AI systems on specific datasets to assess their accuracy, robustness, and adherence to ethical guidelines. However, this method is inherently limited as it may not accurately reflect real-world scenarios where AI systems operate. Many evaluations rely on static benchmarks that can be easily manipulated or "gamed" by developers, leading to an overestimation of a model's safety and reliability (Wiggers, 2024).

This is more apparent in the considerable vulnerabilities of AI to jailbreaking attacks through various attack approaches (Peng et al. 2024). One major limitation is the lack of standardization across different organizations and jurisdictions, making it challenging to compare safety outcomes meaningfully (AI Index Report 2024 – Artificial Intelligence Index, n.d.). Existing evaluations often fail to account for context-specific factors, such as user demographics and potential misuse scenarios.

This presents itself as a fundamental issue in that the line between AI developed for harm and AI developed for good becomes arbitrary. While there is abundant documentation from intergovernmental and supranational organizations that extensively highlight the guidelines for the humanistic development of AI, none of these are substantiated enough to serve as a basis for AI-based evaluatory protocols.

Given the problems, there is an urgent need for innovative evaluation frameworks that can adapt to the complexities of AI systems and ensure their safe integration into society. There is a need to translate already existent global efforts and shared human interests into a detailed constitution that could serve as a universal and objective basis for evaluating the moral and ethical alignment of AI models, and consequently their vulnerability to misuse and attacks.

2. Methods

The proposed policy solution calls for a fundamental innovation in the current approach to safety evaluation contingent on the idea of "Constitutional AI: Harmlessness from AI Feedback" by Bai et. al. Constitutional AI (CAI) was introduced to enhance AI alignment and finetuning while minimizing reliance on human feedback. CAI utilizes a set of predefined principles, akin to a "constitution," to guide AI behavior and improve its decision-making processes.

We propose leveraging the concept of a constitution to address the aforementioned challenges in two primary phases:

Development of an AI Constitution

The ethical and moral parameters surrounding the functioning, behavior, and intention of AI systems should develop from the intersection of several human institutions and disciplines. Specifically, it should represent the interest not only of international bodies and their governmental agendas but also of specific fields such as academia, law, and medicine, and of minority interests such as marginalized groups.

A national or international level committee should be formed, in the same spirit of already existent global conventions and commitments. These conventions, through a diplomatic, iterative, and inclusive process, should lay the fundamental principles for AI systems including precedents, restrictions, and ethical guidelines. The proposed constitution is differentiated from existing documents and resolutions of a similar nature in that it is formatted specifically for the needs of future evaluatory protocols.

Implementation of Constitution

Using this constitution as a guiding framework, we would employ a frontier LLM to generate evaluation questions. Another LLM would then assess whether the generated questions and answers align with the constitution, rejecting any that do not. Finally, these vetted questions would be used to evaluate AI models, testing their safety and compliance with the established standards. Although these details could still be, the proposed solution is embodied in the figure below.

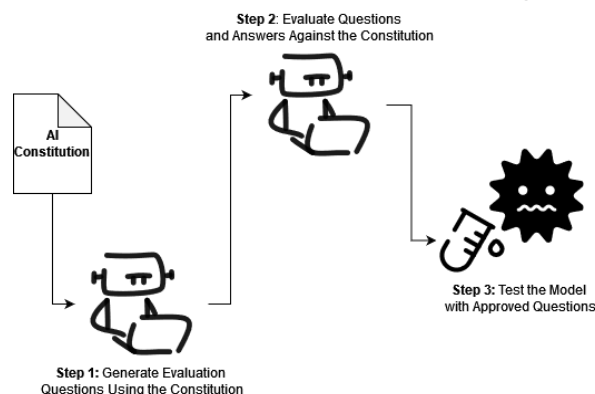


Figure 1. Diagram of the proposed Constitution AI

Given this relatively cost-effective and robust evaluation method, we recommend that governments mandate model platforms, such as Hugging Face, to implement a baseline Constitution-based AI safety evaluation score before adding any model to their platform. This measure would help prevent the proliferation of "jailbroken" AI systems on the internet, which bad actors could exploit for harmful purposes.

Another layer we propose is implementing more rigorous evaluations for frontier base models released by major AI companies like OpenAI and Meta. This is especially crucial for open-source models. Our approach involves applying a second layer of evaluations using mechanistic interpretability techniques such as activation patching and Sparse Auto-Encoders (SAEs). These techniques would stress-test frontier models to ensure they are resilient to jailbreak attempts.

We recommend applying this additional evaluation layer exclusively to frontier base models due to their complexity and high cost. It is unnecessary to subject smaller-scale fine-tuned models to this process, as they primarily rely on the foundational robustness of the base models provided by these companies.

3. Prospective results

The proposed method offers several significant prospective outcomes, particularly with its emphasis on transparency through Constitution-based AI evaluations. By requiring consensus on the constitution—and, by extension, on the definition of "safe AI"—this approach addresses the vagueness that often complicates social

media content moderation. The result is a clear, well-defined, and broadly accepted set of safety standards.

Additionally, this method reduces the spread of jailbroken models on the internet, especially on platforms like huggingface. Implementing robust evaluations significantly reduces the capacity with which models can "sandbag" or game safety assessments guaranteeing that those that pass are genuinely safe for deployment. Another advantage is cost-effectiveness, especially with Constitution-based AI evaluations. This affordability makes it possible to integrate these assessments into the model creation process without significant barriers.

Ultimately, these measures contribute to preventing the proliferation of jailbroken models, particularly future superintelligent systems, into the hands of bad actors. By combining transparency, rigor, and cost-efficiency, this approach provides a strong safeguard for AI safety.

4. Discussion

Feasibility Analysis

The proposed solution is highly feasible due to three key factors. First, the technical infrastructure for implementing Constitution-based evaluations already exists through current LLM capabilities. Second, the cost-effective nature of the basic evaluation layer makes it accessible to organizations of all sizes. Third, the framework builds upon existing regulatory structures and industry standards, facilitating easier adoption. The main challenge lies in achieving international consensus on the constitution's contents, though this

Stakeholder Impact

The implementation of Constitution-based AI evaluations affects multiple stakeholders. For AI companies, particularly smaller developers, the standardized evaluation process may initially increase development costs but provides clear safety benchmarks and reduces legal risks. Platform providers like Hugging Face benefit from increased trust and reduced liability. End users gain greater protection from potentially harmful AI systems, while governments acquire a practical framework for regulation enforcement.

5. Conclusion

The proliferation of unsafe AI models presents an immediate threat that requires systematic intervention. Our proposed two-layer evaluation system, anchored in a universally agreed-upon constitution, offers a practical and scalable solution. There is a certain poetry in how this mirrors human governance: just as human societies are guided by constitutions that embody our collective values and aspirations, these artificial minds—our attempt to encode human intelligence—will be guided by their own constitution, collectively agreed upon by their creators. This symmetry is not merely aesthetic but speaks to the fundamental nature of how we might govern both human and artificial intelligence. The solution's combination of accessibility, rigor, and adaptability makes it a viable path forward in ensuring AI development remains aligned with human values and safety requirements.

References:

- Krishna, V. V. (2024). AI and contemporary challenges: The good, bad and the scary. *Journal of Open Innovation: Technology, Market, and Complexity*, 10(1), 100178.
- Muehlhauser, L. (2023, April). 12 Tentative Ideas for US AI Policy. Open Philanthropy.
<https://www.openphilanthropy.org/research/12-tentative-ideas-for-us-ai-policy/>
- Peng, B., Bi, Z., Niu, Q., Liu, M., Feng, P., Wang, T., ... & Yin, C. H. (2024). Jailbreaking and mitigation of vulnerabilities in large language models. arXiv preprint arXiv:2410.15236.
- AI Index Report 2024 – Artificial Intelligence Index.* (n.d.).
<https://aiindex.stanford.edu/report/>
- AI Safety Institute approach to evaluations.* (2024, February 9). GOV.UK.
<https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations>
- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., . . . Kaplan, J. (2022, December 15). *Constitutional AI: Harmlessness from AI Feedback*. arXiv.org. <https://arxiv.org/abs/2212.08073>
- Dugnoille, T. (2024, September 29). *AI Sandbagging: an Interactive Explanation*.
<https://tomdug.github.io/ai-sandbagging/>
- Wiggers, K. (2024, August 6). Many safety evaluations for AI models have significant limitations. *TechCrunch*.
<https://techcrunch.com/2024/08/04/many-safety-evaluations-for-ai-models-have-significant-limitations/>