

Peer Review

Review of: "AI Ethics by Design: Implementing Customizable Guardrails for Responsible AI Development"

Nythamar de Oliveira¹

1. Philosophy, Pontifícia Universidade Católica do Rio Grande do Sul, Brazil

This is a highly original article focusing on the development of an ethical guardrail framework for AI systems rather than ethical principles, practical issues, or consequences for AI systems and users. By stressing the importance of customizable guardrails that align with diverse user values and underlying ethics, the authors have successfully addressed most of the major challenges posed for AI ethics and proposed a framework integrating rules, policies, and AI assistants to ensure responsible AI behavior. This is, by all criteria, one of the most enlightening papers on the existing state-of-the-art guardrails for research in AI ethics.

It seems that the major normative challenge for research in AI Ethics continues to be the impossibility of simply resorting to a list of principles or procedures that do not necessarily apply to specific contexts (social, economic, cultural, etc.) or moral demands. Just as it happens with principlism, there are also moral demands that are typical for a given social or cultural situation (thinking of virtues, consequences, deontological principles). Hence, the proposal of developing an ethical guardrail framework for AI systems rather than ethical principles seems to be more realistic and feasible.

Declarations

Potential competing interests: No potential competing interests to declare.