Qeios

Peer Review

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

Mamia Agbese[1]

1. University of Jyväskylä, Finland

The idea of the paper is highly relevant, especially given the growing use of AI, particularly LLMs. While this is an important and timely topic, the paper itself struggles to clearly articulate its central message. It would benefit from a more focused and coherent presentation of the key points to better communicate the authors' main argument.

Introduction

The introduction provides a broad overview of the ethics of AI and some of the ethical challenges. However, it lacks a clear focus on the specific point the authors are trying to communicate. The discussion remains general and touches on AI in a wide context, with a brief mention of LLMs, but it does not provide a clear definition or narrow down to a focused topic. It would be helpful to refine the introduction to clarify the specific area of focus and to better guide the reader through the main argument.

It seems the study is specifically focused on LLMs; it would be helpful to describe them within the context of the study and then explain the key concepts in relation to this focus. Additionally, it would be useful to clarify whether the paper is addressing the philosophical aspects of AI or just mentioning it. For example, how do issues like algorithmic bias and the AI's ability to process vast amounts of data relate to the study's central argument? Providing more explicit connections between these concepts and the study’s goals would strengthen its focus and clarity.

Rationale:

doi.org/10.32388/JPUCKQ

The rationale for the study is unclear to me and can benefit from further clarification. Specifically, it is important to define what the study is about: Is it focused on policy revision, technical aspects, regulation, or argumentation? If it is about policy, which specific policy is being addressed? If technical, etc. Additionally, is the study examining the ethical challenges of AI in general or focusing specifically on LLMs? Or is the paper intended to discuss AI guardrails?

These concepts need to be clearly defined, specifying which area of AI it refers to. As it stands, the lack of clarity makes it difficult to understand the authors' central argument and objectives.

Is the purpose of the paper to shift ethical decision-making to the user and create a framework that aligns with their ethical needs and values? This central idea needs to be clearly communicated in both the introduction and throughout the body of the paper. As it stands, this purpose is neither explicitly mentioned nor fully addressed. It is important to ensure that this concept is integrated throughout the paper so that the reader can easily understand the focus of the study. Additionally, if this is indeed the crux of the study, the paper should clearly connect this idea to existing frameworks of NeMo Guardrails, LlamaGuard, and Guardrails AI, especially if they are being used as examples or benchmarks. While these tools offer important safeguards, the paper should clearly explain how they fall short in providing the necessary agility using citations and customization for different organizational needs, if that is a central argument.

Scope

The scope and context of the paper appear to be mixed, addressing both technical and moral aspects, which are often handled separately in academic discourse. If the focus is primarily on the technical aspects, the scope of the paper should reflect this clearly. On the other hand, if the paper's main focus is on the ethical or moral aspects, particularly in relation to AI, this should be explicitly communicated. If the intention is to show how the technical and ethical aspects influence one another, this relationship should be clearly explained. As it stands, the paper seems to blend both perspectives without clearly differentiating or explaining how they interrelate, which makes the overall argument somewhat confusing.

Citation:

Many of the issues raised in the paper lack adequate support from existing literature or practical examples. For instance, the claim that 'they may also not easily accommodate integration with diverse data sources, such as internal databases, CRM systems, or industry-specific knowledge repositories,

which is essential for creating truly context-aware and aligned AI assistants' would benefit from citation to relevant studies or examples to substantiate it. This issue of unsupported claims continues throughout the remainder of Chapter 3 and parts of 4. Incorporating citations from relevant literature or practical case studies would strengthen the argument and provide the necessary context for the points being made.

Proposed Solution-

The proposed solution could benefit from more clarity and focus, as some elements seem to be less connected and might need further alignment to create a more cohesive strategy.

The authors combine several factors in the proposed solution, but the synergy between them is not entirely clear. It's also unclear whether the solution is based on empirical research or a review of the literature. If it is based on empirical research, this is not communicated clearly; and if it is a review of literature, this aspect could be more explicitly outlined and the methods used to arrive at the solution clearly articulated.

The solution is somewhat vague, particularly in the claim that 'various rules can be combined into policies.' I find this statement unclear, as there are no citations provided to support it. To strengthen the argument, it would be helpful to include references or further explanation, as making broad statements without supporting evidence can weaken the credibility of the article.

This trend continues throughout the proposed solution and framework summary.

The article reads more like an essay and lacks the empirical support typically expected in this type of publication. If this is the intended format for the article, I recommend that the authors review similar published works to better understand how to clearly articulate their arguments and incorporate the necessary evidence. Based on the current state of the article, I would recommend rejection for now.

Declarations

Potential competing interests: No potential competing interests to declare.