Ethical Guardrails for AI: A Framework for Fairness and "Make Your Own Ethics"

This paper presents a novel framework for implementing customizable ethical guardrails in artificial intelligence (AI) systems, focusing on aligning AI behavior with diverse user values and addressing critical fairness challenges in algorithmic systems. The framework, first proposed by Šekrst et al. (2024), introduces a flexible structure of rules and policies designed to mitigate algorithmic biases, prevent data privacy violations, and enhance fairness through user-centered configurations. By accommodating ethical pluralism and promoting transparency, this approach ensures that AI systems can dynamically adapt to various societal, organizational, and individual ethical standards.

The proposed framework integrates three types of rules – static, natural-language, and classifier-based – which can be customized and combined into hierarchical policies. These policies govern both user input and AI output, enabling comprehensive control over interactions. Importantly, the design explicitly addresses emerging threats such as prompt injections, where adversarial inputs manipulate AI systems to bypass ethical safeguards, as highlighted by Branch et al. (2022). It also builds upon insights from Winfield et al. (2019) regarding ethical design for autonomous systems and the taxonomy of ethical agency, as well as Russell's (2019) principles for aligning AI with human values.

Furthermore, this paper critically evaluates existing guardrail systems, including NeMo Guardrails and LlamaGuard, highlighting their limitations in scalability, flexibility, and user accessibility. The proposed solution addresses these challenges by offering a user-friendly interface for ethical customization, ensuring broader applicability across industries such as healthcare, finance, and education, where fairness is paramount. To resolve conflicts arising from ethical pluralism, the framework employs strategies such as weighted averaging, hierarchical precedence, and contextual triggering, ensuring consistent and fair decision-making. Additionally, the use of customizable rules provides an avenue for continuous improvement, allowing end-users to iteratively refine ethical guidelines in response to evolving societal norms and regulations.

By bridging technical innovations with philosophical foundations, this framework advances the discourse on fairness in machine learning. It emphasizes the need for interdisciplinary collaboration to design AI systems that are not only technically robust but also ethically sound, paving the way for a more equitable integration of AI into society.

References

1. Branch, H. J., Rodriguez Cefalu, J., McHugh, J., et al. (2022). Evaluating the susceptibility of pre-trained language models via handcrafted adversarial examples. *arXiv preprint arXiv:2209.02128*. https://arxiv.org/abs/2209.02128
2. Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics, 21*(4), 403–418. https://doi.org/10.1007/s10892-017-9252-2
3. Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. New York, NY: Viking.
4. Šekrst, K., McHugh, J., & Rodriguez Cefalù, J. (2024). AI ethics by design: Implementing customizable guardrails for responsible AI development. *arXiv preprint arXiv:2411.14442*. https://arxiv.org/abs/2411.14442
5. Winfield, A. F., Michael, K., Pitt, J., & Evers, V. (2019). Machine ethics: The design and governance of ethical AI and autonomous systems [Scanning the issue]. *Proceedings of the IEEE, 107*(3), 509–517. https://doi.org/10.1109/JPROC.2019.2900622