IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 12, May 2025
Impact Factor: 7.67

Latent Space Ethics : Controlling AI Content Generation Before Output

Alanjoe George Paul Mendonca and Flavia D Gonsalves
MCA, MET-ICS, Mumbai University, Mumbai, India
Assistant Professor, MET-ICS, Mumbai, India
mca23_1434ics@met.edu and flaviag_ics@met.edu

Abstract: *The rise of AI-generated content has raised growing concerns around bias, misinformation, and the creation of harmful material. Traditional moderation systems often act after content is generated, relying on filtering or blocking tools. This research introduces a proactive solution by embedding ethical principles directly into the latent space of generative models—intervening before content reaches the output stage.*

Our methodology applies mathematical constraints and ethically guided loss functions within the latent space of large language and image models. By modifying latent vectors during training, we encourage the model to internalize ethical AI principles, guiding it away from unethical conceptual directions before generation occurs. This reduces dependence on external content moderation tools and shifts ethical awareness to the core of the model's decision-making process.

Initial results show a noticeable decline in the production of biased, offensive, or misleading outputs across both text and multimedia, while maintaining a high level of creative freedom. This suggests that enforcing ethical boundaries need not limit originality. We also examine the sociocultural complexity of defining "ethics" and highlight the importance of avoiding over-constraint.

This study marks a significant step toward AI safety, contributing to the design of systems capable of generating responsible, culturally aware, and autonomous content from the outset.

Keywords: Latent space, ethical AI, content moderation, creative freedom, AI safety

DOI: 10.48175/IJARSCT-27814