Chapter 6
Responsible Intelligence: A Framework for Managing AI Hallucinations in the Metaverse Systems

Gagandeep Singh
https://orcid.org/0000-0001-8644-4064
Lovely Professional University, India

Jasdeep Singh Walia
https://orcid.org/0000-0001-9010-082X
Lovely Professional University, India

ABSTRACT

The integration of multimodal generative AI into metaverse environments has intensified hallucination-driven distortions, creating synthetic outputs that appear credible within immersive spaces and influence judgment, behaviour and decision accuracy. Continuous presence in avatar-based ecosystems heightens perceptual reality drift and cognitive misalignment. The absence of responsible intelligence mechanisms for detection, monitoring and mitigation leaves immersive system users exposed to synthetic narrative manipulation, fabricated identity signals and cross-modal inconsistency. Fragmented regulation and limited provenance verification deepen risks as hallucination cycles evolve without oversight. This study develops a governance framework centred on real-time authenticity assurance, distortion tracking and consistency verification. It advances a structured approach to protect immersive participants, reduce uncontrolled hallucination escalation and strengthen accountable metaverse safety architecture.

DOI: 10.4018/979-8-3373-7534-2.ch006

1. INTRODUCTION

The rapid growth of Metaverse and immersive Metaverse-integrated AI systems is remodelling all major areas by creating enormous potential for the creation of sophisticated innovations (Mao & Xu, 2025). The inequalities in its fair access and in its distribution present major challenges for metaverse participants. Artificial intelligence and multimodal hallucination-prone models' vulnerabilities are impacting metaverse participants disproportionally (Hermann et al., 2024). The cognitive decline and risk of dark AI patterns mainly arise due to an inability to use digital technology (Ferrara, 2023). The speed of creation of new AI tools has come at considerable perceptual reality drift to immersive environment users, particularly metaverse participants and people (Kallina et al., 2025; Drabiak et al., 2023). The social inequalities are being exacerbated by an unrestricted artificial intelligence that does not accept ethical principles and harms privacy and trust. Immersion eliminates the separation between interface and perception, which places users inside narrative structures. Such structures are generated by large language models, avatar communication agents and real-time generative simulation systems

Metaverse participants are more impacted by AI tech-based decision-making aid, and it increases the digital divide. The development of Metaverse-integrated AI systems must be purposefully designed to take care of algorithm-level manipulation through intelligent interfaces (Novelli et al., 2024). The violation of their interests can have damaging impacts on a range of life decisions. It can tamper with their ability to govern themselves and lower individual independence. Also, this makes huge compromises to data privacy in the case of the metaverse participants (Kallina et al., 2025). The unregulated AI hallucinations and their existing paradigms do not consider the perceptual reality drift of unregulated use of AI (Hewage et al., 2024; Tejaskumar Pujari et al., 2022).

The current chapter keeps track of the analysis regarding the fast integration of the technologies and the vulnerabilities of our ageing population. The objective of the present research work is to carry out research on the issues arising due to digital divides, ethics and the absence of proper governance. This chapter aims to study the role of Metaverse-integrated AI systems integrated with the lives of immersive system stakeholders and the dangers that Metaverse-integrated AI systems pose due to the fragmented regulatory regime. It seeks to outline a governance model that is inclusive to promote trustworthy AI, thereby ensuring its conscientious use. This study closely examines the deficiencies, focusing on identifying the deficiencies in the damages caused by AI from the disconnected regulations and the complete participation of all stakeholders. This study investigates the effects of high-intensity metaverse exposure on their mental health.