Chapter 9
Human–AI Synergy in Ethical Content Moderation:
Navigating Fairness, Accountability, and Transparency Challenges

Hemlata Parmar
https://orcid.org/0009-0009-1438-5427
Manipal University Jaipur, India

Utsav Krishan Murari
https://orcid.org/0009-0007-1606-6775
Sharda University, India

ABSTRACT

The rapid growth of digital platforms has highlighted the need for robust content moderation systems that manage large content volumes while addressing human moderators' mental health. This study explores AI's evolving role in content moderation, emphasizing collaboration with human moderators for nuanced judgments. AI can identify harmful content trends, but human oversight is needed for contextual accuracy. However, AI raises ethical issues, such as bias from prejudiced training data, potentially harming marginalized groups. To improve accountability, the study suggests using diverse datasets, transparent decision-making, and explainable AI. Regulatory frameworks and mental health support for moderators are also essential for fair, transparent content moderation.

DOI: 10.4018/979-8-3693-9904-0.ch009

1. INTRODUCTION

Nowadays Artificial Intelligence(AI) is reinventing every single aspect of modern life from healthcare, education and finance to governance and entertainment. Generative artificial intelligence is developing faster than ever and the world started adopting it widely, which led many to debate whether generative AI can improve or sink society (The Economic Times, 2023). Simultaneously, AI also has incredible potential to provide breakthroughs and solve some of humanity's most challenging problems rapid diagnostics resulting in better health care; predictive analytics for climate change; broader access to high-quality education for underrepresented communities. But it also asks difficult questions on equality and discrimination, employment and privacy which if not managed correctly could push the minimizing of social injustices into overdrive (Anderson & Rainie, 2018).

While AI was previously only a technical technology discussion, it is now also social, ethical and economic concerns given the current rapid pace of AI development. And finally, scholars, policymakers and advocates are asking, who benefits from AI? Unless we ensure that the dividends of innovation, fueled by AI, be equally dispersed among our nations. How to ensure that groups at risk are not harmed, in determining how AI systems goodness can be ensured, and by providing other protections from harm? Addressing all of these problems will require social justice fundamentally embedded in a systems approach to AI research, development and deployment (Bhuptani, 2024).

1.1 Defining Social Justice in the Context of AI

Many of the discussions of the implications of AI have centered around social justice, which is defined as the equitable distribution of resources, opportunities and privileges within a society. In a just, equitable society, the progress of AI would not duplicate nor increase inequality; it is up to us to determine if and how this technological revolution becomes instead a tool for common good. It involves centering the experiences and rights of marginalized communities, promoting equity and accountability in algorithmic systems, and facilitating access to AI governance and policy processes (Buccella, 2022).

We know from history that technological revolutions have not only informed how we live but also embodied and exaggerated social inequities. For instance, the industrial revolution has been very beneficial for economy development but it had an unfair approach to human resources - thus raising inequalities between riches. There was good in a digital age even if it forged an invisible footprint, a widening gap between individuals with the wherewithal to enter and leverage the power of information technology and those standing at the curb looking on helplessly as little