Chapter 11
Legal Frameworks Surrounding the Use of AI in Online Content Moderation

Pranjal Khare
https://orcid.org/0000-0002-9937-9588
O.P. Jindal Global University, India

Vishambhar Raghuwanshi
https://orcid.org/0009-0009-1365-1850
Manipal University Jaipur, India

ABSTRACT

The use of AI in online content moderation is a complex issue with significant ethical and legal implications. While AI offers the potential to efficiently identify and remove harmful content like hate speech and misinformation, it also raises concerns about censorship, biased algorithms, and the erosion of user trust. Striking a balance between free speech and user safety is crucial. Ethical frameworks and regulations are needed to guide the development and deployment of AI moderation tools, ensuring transparency, accountability, and fairness. However, the lack of global consensus and inconsistencies in national regulations hinder the development of a coherent international approach to AI governance. To address these challenges, this chapter will explore the legal framework and a global approach which is needed to establish standards for transparency, accountability, and fairness in AI-driven content moderation, ensuring that AI serves as a tool for good rather than harm.

DOI: 10.4018/979-8-3693-9904-0.ch011

INTRODUCTION

Computing intelligence has become indispensable in the current world, especially in managing content on social media. Social media is currently experiencing exponential growth in terms of population reach and has increasingly influenced the way people interact and acquire information and similarly has become the driving force for sharing useful as well as negative information. One of the most crucial problems of the contemporary network environment is the increase in the quantity of hate speech, fake news, and other various forms of undesirable content, which requires some new ways to regulate it. Considering the amount of content created every day, more platforms seek to use AI technologies for content moderation. This introduction will discuss the importance of AI to the regulation of free speech on social media platforms, in a bid to address questions about permissioned knowledge, and social media toxicity. This section shows how, depending on the context, AI can participate both to reopen expression or contribute to its closure This section will also demonstrate the importance of finding middle ground for the appropriate regulation of AI for moderating content (Canda, 2024).

The use of artificial intelligence in content moderation meets a staggering demand. Since there are billions of active users on social media sites like Facebook, Twitter, and YouTube, the social media firm finds itself in a very big challenge of tackling a very big number of events to sort out. This task is further complicated by high incidence of harmful material, with materials such as hate speech, and racist and extremist propaganda, harassment and misinformation. The volume and high velocity of such content ensures its impossibility of moderation by human beings alone, hence making way for the new age AI solutions that are also efficient. Due to great computing power, AI can be used to analyze potentially toxic content and then remove it before it spreads, making it an excellent tool for real-time moderation. This reliance on artificial intelligence is characteristic of a larger trend where platforms are self-monitoring content to manage a scope of information and to increase the quality of user experience through indicating safer content.

Notably, the application of AI in content moderation can be broken down into complex and practical technologies in the following manners. Some techniques, for instance, Natural Language Processing (NLP) techniques make the system input and deeply process text to be in a position to identify the language used in hate speech and identification of threats or harassments. Similarly image and video recognition software are tasked to scan image and video content in an attempt to identify obscene images or symbols linked to hate groups. Both technologies are firmly based on machine learning, which trains systems to be able to identify patterns in data, refine how they operate over time, and, for the best results, be able to update as the content they are presented with changes over time online. AI's adaptation in content