Chapter 7
The Dark Side of AI:
Risks, Ethics, and Safeguarding Human Interests

Arun Agrawal
https://orcid.org/0000-0001-7233-6660
Institute of Technology and Management, Gwalior, India

Jagveer Singh
Institute of Technology and Management, Gwalior, India

ABSTRACT

Artificial intelligence (AI) is a cornerstone of modern technology, driving advances in automation, decision-making, and innovation in many fields. While AI offers significant benefits, it also raises concerns that it could harm humanity if misused or poorly designed. This article takes a closer look at the ways in which artificial intelligence can negatively impact humans, from invasions of personal privacy and financial security to impacts on mental health and physical safety. By analyzing real-world case studies, this discussion highlights the ethical dilemmas surrounding AI adoption. The paper further proposes strategies and frameworks to ensure that AI is aligned with human values, and advocates responsible development and regulation to reduce risks. Ultimately, this research strikes a balance between harnessing the potential of AI and preventing unintended consequences, fostering a future where AI serves humanity rather than destroys it.

DOI: 10.4018/979-8-3693-9904-0.ch007

1. INTRODUCTION

Artificial intelligence (AI) has become one of humanity's greatest assets, with the potential to revolutionize industry, healthcare, education, and governance. From enhancing decision-making to automating daily tasks, artificial intelligence offers significant benefits (Mhlanga, 2022). However, its rapid development has also raised concerns about unintended consequences of misuse or malfunction. These concerns include the erosion of human autonomy, invasion of privacy, increasing inequality, and even physical harm (Blauth, Gstrein, & Zwitter, 2022)

This Chapter explores how artificial intelligence can work against human interests, whether through deliberate misuse or unintentional design flaws. We examine how bias, lack of transparency, and a lack of ethical oversight in AI algorithms can lead to harmful outcomes. In addition, malicious applications of AI, such as surveillance, cyberattacks, and autonomous weapons, are also considered key areas where AI may pose significant risks. This Chapter provides real-life examples of AI-driven systems leading to discriminatory outcomes, job losses, and threats to personal privacy.

To address these risks, this Chapter highlights the importance of putting in place ethical frameworks and regulatory measures to ensure the responsible development and deployment of AI. To prevent discrimination, we advocate for greater transparency in AI systems, more rigorous testing for bias and error, and the inclusion of diverse perspectives in AI development. The Chapter concludes by proposing strategies to align AI with human values and ensure that technological advances support rather than harm human well-being. Addressing the risks and challenges of AI can help realize its potential while preventing unintended harm.

2. THE POTENTIAL FOR AI TO WORK AGAINST HUMAN INTERESTS

While AI is designed to perform tasks that help and enhance humans, it does not inherently cause harm. The following sections explore key areas where AI may have an impact on humanity.

2.1. Job Displacement and Economic Inequality

One of the most pressing concerns about the rise of artificial intelligence (AI) is its potential to displace jobs and thereby exacerbate economic inequality (Farahani & Ghasemi, 2024). As artificial intelligence becomes increasingly capable of automating both routine and complex tasks, many industries are experiencing significant changes in their workforce needs (Tschang & Almirall, 2021). Jobs that