


Chapter 11


AI Safety and Security

Mosiur Rahaman

 <https://orcid.org/0000-0003-0521-2080>

*International Center for AI and Cyber Security Research and Innovations,
Taiwan*

Princy Pappachan

 <https://orcid.org/0000-0001-6728-0228>

Department of Foreign Languages and Literature, Asia University, Taiwan

Sheila Mae Orozco

Northern Bukidnon State College, Philippines

Shavi Bansal

Insights2Techinfo, India

Varsha Arya

Department of Business Administration, Asia University, Taiwan

ABSTRACT

The chapter “AI Safety and Security” presents a comprehensive and multi-dimensional exploration, addressing the critical aspects of safety and security in the context of large language models. The chapter begins by identifying the risks and threats posed by LLMs, delving into vulnerabilities such as bias, misinformation, and unintended AI interactions, impacts like privacy concerns. Building on these identified risks, it then explores the strategies and methodologies for ensuring AI safety, focusing on principles like robustness, transparency, and accountability and discussing the challenges of implementing these safety measures. It concludes with an insight into long-term AI safety research, highlighting ongoing efforts and future directions to sustain AI system safety amidst rapid technological advancements and encouraging a collaborative approach among various stakeholders. By integrating perspectives

DOI: 10.4018/979-8-3693-3860-5.ch011

from computer science, ethics, law, and social sciences, the chapter provides an insightful and comprehensive analysis of current and future challenges in AI safety and security.

INTRODUCTION

A wide range of technologies that allow machines to carry out tasks that ordinarily require human intelligence are together referred to as artificial intelligence (AI). These tasks include language translation, speech recognition, sophisticated data interpretation, and decision-making (Joiner, 2018). Large Language Models (LLMs), such as OpenAI's GPT series, have become particularly important among the many AI implementations. Large-scale text datasets are used to train LLMs, which enable them to produce language that is logical and appropriate for the context given the input (Kalyan, 2024). Their capacity to estimate the likelihood of a word sequence enables them to construct complete phrases and paragraphs that can imitate dialogue, provide content, and respond to queries (Dwivedi et al., 2023).

It is impossible to adequately emphasise how crucial safety and security are for AI systems, especially as these tools are incorporated into more and more vital elements of daily life. The goal of AI safety is to guarantee that these technologies function within their intended parameters without resulting in unintentional harm (Díaz-Rodríguez et al., 2023). This includes stopping AI systems from becoming biased and acting accordingly, from choosing poorly, and from being tricked by malicious inputs. Conversely, security pertains to safeguarding AI systems from unapproved access and attacks that may result in data breaches or improper use (Frank, 2024). In addition to safeguarding people and their data, ensuring the safety and security of AI systems is essential for preserving public confidence in these technologies and encouraging their widespread use (Yelne et al., n.d.). Figure 1 illustrating the key aspects of LLMs and AI system safety and security.

28 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/ai-safety-and-security/354401

Related Content

Intrusion Detection and Analysis in IoT Devices Using Machine Learning Models

Ankit Kumar Jain, Pooja Kumari and Ritesh Gupta (2024). *Challenges in Large Language Model Development and AI Ethics* (pp. 384-409).

www.irma-international.org/chapter/intrusion-detection-and-analysis-in-iot-devices-using-machine-learning-models/354402

GuessXQ: A Query-by-Example Approach for XML Querying

Daniela Morais Fonte, Daniela da Cruz, Pedro Rangel Henriques and Alda Lopes Gancarski (2013). *Innovations in XML Applications and Metadata Management: Advancing Technologies* (pp. 57-76).

www.irma-international.org/chapter/guessxq-query-example-approach-xml/73173

Cybersecurity Fundamentals

Boy Firmansyah (2024). *Challenges in Large Language Model Development and AI Ethics* (pp. 280-320).

www.irma-international.org/chapter/cybersecurity-fundamentals/354399

Software Evolution with XVCL

Weishan Zhang, Stan Jarzabek, Hongyu Zhang, Neil Loughran and Awais Rashid (2005). *Software Evolution with UML and XML* (pp. 152-189).

www.irma-international.org/chapter/software-evolution-xvcl/29613

Formal Specifications of Software Model Evolution Using Contracts

Claudia Pons and Gabriel Baum (2005). *Advances in UML and XML-Based Software Evolution* (pp. 184-208).

www.irma-international.org/chapter/formal-specifications-software-model-evolution/4936