Detecting the Detectors: A Systematic Review of AI-Generated Text Detection Tools and Their Reliability

Pasawut Cheerapakorn[1] , Rattanakul Kongpha[2] and Juneerat Jannit[3]

[1]Educational Innovation and Technology, Faculty of Education

[2]The Office of General Education and Innovative Electronics Learning

[3]The Office of The President

[1,2,3]Suan Sunandha Rajabhat University, Thailand

Email: pasawut.ch@ssru.ac.th[1] ; rattanakul.ko@ssru.ac.th[2] ; juneerat.ja@ssru.ac.th[3]

Abstract

This study systematically investigates the reliability of AI-Generated Text Detection Tools through a PRISMA-based systematic review combined with bibliometric analysis. Using the SCOPUS database, 949,127 records related to Artificial Intelligence (AI), Generative AI, AI-Generated Text Detection, and AI Ethics were identified and filtered through identification, screening, and eligibility stages. A total of 73 qualified studies were included for synthesis. The PRISMA framework ensured methodological transparency, while bibliometric analysis using VOSviewer and the R Bibliometrix Package revealed publication trends, key contributors, and research networks. The analysis identified Artificial Intelligence, ChatGPT, and Generative AI as core topics strongly associated with research integrity, ethics, and detection reliability. Findings indicate that BERT-based and graph neural network models show high accuracy in distinguishing AI-generated text but remain inconsistent across linguistic and contextual variations. Bibliometric mapping uncovered five major research clusters—AI Ethics, Academic Integrity, Language Models, Detection Methods, and Education—reflecting the interdisciplinary nature of this domain. The study emphasizes that technical precision alone is insufficient; ethical considerations such as transparency, fairness, and accountability are crucial for maintaining reliability and trust in AI detection tools. In conclusion, developing trustworthy AI detectors requires a balanced integration of technical validation and ethical governance. The results highlight a need for continuous refinement of methodologies and stronger alignment with ethical principles to enhance trust, transparency, and research integrity in the era of generative AI.

Keywords: Artificial Intelligence (AI), Generative AI, AI-Generated Text Detector, AI Ethics

1. Introduction

Artificial Intelligence (AI) has become a transformative force that continues to reshape human life, enabling machines to perform tasks that require cognitive abilities such as reasoning, problem-solving, and natural language understanding ("Artificial Intelligence," 2022; Morandín-Ahuerma, 2022). It is generally divided into *narrow AI*, which focuses on specific applications like image recognition or speech processing, and *general AI*, which aspires to replicate human-like cognition across various domains (Suresh & Sasidharan, 2025;

Chaudhary et al., 2024). The advancement of AI technologies—particularly in machine learning and natural language processing—has revolutionized industries such as healthcare, finance, and education. However, these innovations also bring ethical challenges, including bias, privacy, and job displacement (Chaiwchan, Kaewrattanapat, & Nookhong, 2025; Suresh & Sasidharan, 2025).

A recent breakthrough in this evolution is Generative Artificial Intelligence (GenAI), which can autonomously generate text, images, and audio by learning from extensive datasets. In education, GenAI assists in intelligent exam design, personalized learning, and AI-driven tutoring that supports diverse learners (Chow, 2024; Sowmiya, 2025). In business, it fosters productivity and cross-disciplinary innovation through data-driven automation (Bharti et al., 2024; Taşabat, 2025). Despite its benefits, the increasing sophistication of GenAI has raised significant concerns over authenticity, originality, and academic integrity. The ability of AI to produce human-like text has blurred the boundary between human and machine authorship, leading to the urgent need for effective AI-generated text detection tools (AI Detectors).

Recent studies have developed several approaches to detect AI-generated content. BERT-based models have demonstrated high accuracy—up to 99.72%—by leveraging contextual language representation (Wang et al., 2024). Graph neural networks (GNNs) also show promise in capturing writing structures, achieving accuracies of about 93.6% (Abbas, 2025). Theoretical research suggests detection is feasible unless human and machine text distributions are identical (Chakraborty et al., 2023). Nevertheless, the reliability of current tools remains inconsistent across languages, genres, and evolving AI architectures.

Within this context, AI ethics has emerged as a critical field that addresses the moral implications of AI systems, emphasizing *fairness, transparency, accountability, and privacy* (Gao et al., 2024; Bayan, 2024). Scholars highlight the need for explainable AI to ensure decision-making transparency and accountability ("An Overview of Artificial Intelligence Ethics," 2023). Principles-based and process-oriented frameworks have been introduced to guide ethical AI development, advocating human-centric approaches that integrate ethical reflection throughout the AI lifecycle (Kazim & Koshiyama, 2020, 2021). Yet, the rapid advancement of AI technologies continues to outpace ethical governance, risking misuse and social harm.

Accordingly, this study conducts a PRISMA-based Systematic Review titled *"Detecting the Detectors: A Systematic Review of AI-Generated Text Detection Tools and Their Reliability."* It aims to synthesize existing research, evaluate detection accuracy and reliability, and identify bibliometric trends to guide the development of transparent, accountable, and ethically grounded AI detection systems.

2. Research Objectives

2.1. Research Questions

2.1.1 How can PRISMA analysis be applied to systematically review and synthesize research on AI-Generated Text Detection Tools and Their Reliability

2.1.2 What is the key bibliometric trends, including publication growth, leading sources, geographical distribution, and subject areas, in AI-Generated Text Detection Tools and Their Reliability

2.2. Research Objectives

2.2.1 To conduct a PRISMA-Based Systematic Analysis of the AI-Generated Text Detection Tools and Their Reliability

2.2.2 To perform a Bibliometric Analysis of the AI-Generated Text Detection Tools and Their Reliability

3. Conceptual Framework

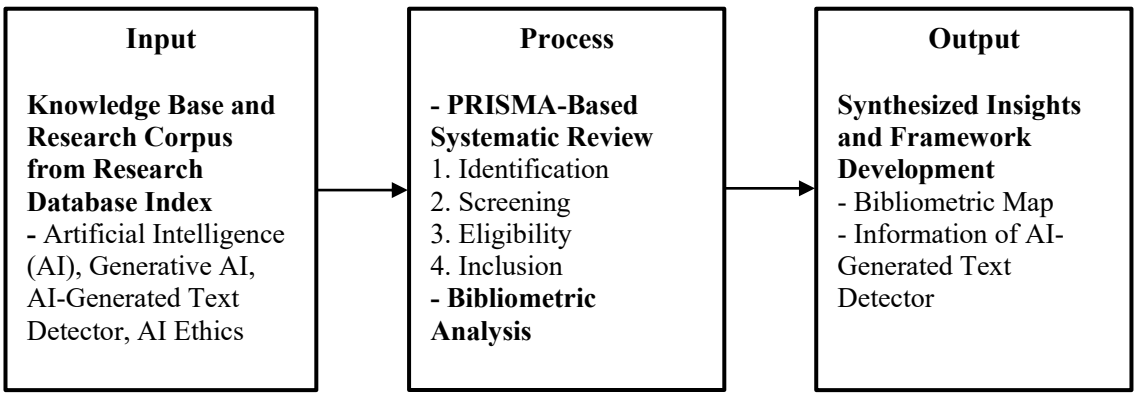

Figure 1. Conceptual Framework of the Study

Figure 1 presents the conceptual framework of the study, illustrating the systematic process used to analyze and synthesize existing research on *AI-Generated Text Detection Tools and Their Reliability*. The framework consists of three primary components: Input, Process, and Output, which collectively represent the logical flow of the research methodology.

The Input stage comprises the *knowledge base and research corpus* derived from indexed research databases, focusing on key domains including *Artificial Intelligence (AI)*, *Generative AI (GenAI)*, *AI-Generated Text Detection*, and *AI Ethics*. These sources form the foundational body of literature that guides both the systematic review and bibliometric analysis.

The Process stage integrates two methodological approaches: the PRISMA-Based Systematic Review and Bibliometric Analysis. The PRISMA framework ensures a rigorous and transparent procedure through four sequential phases—*Identification, Screening, Eligibility,* and *Inclusion*—to select relevant studies. Following this, the *Bibliometric Analysis* quantitatively examines publication patterns, keyword co-occurrences, research productivity, and global collaboration networks, thereby complementing the qualitative synthesis from the PRISMA process.

The Output stage delivers *synthesized insights and framework development*. This includes the creation of a bibliometric map visualizing research trends and interconnections, as well as an information synthesis of AI-generated text detection tools detailing detection models, performance metrics, and ethical implications. Together, these outcomes contribute to a comprehensive understanding of the reliability, transparency, and ethical considerations within AI-generated text detection research.

Overall, the conceptual framework emphasizes the integration of systematic and bibliometric methodologies to produce evidence-based, ethically grounded insights that support the advancement of trustworthy AI detection systems.

4. Methodology

This study employs a systematic approach integrating PRISMA-based analysis and bibliometric analysis to examine the AI-Generated Text Detection Tools and Their Reliability. The methodology is structured into two phases corresponding to the research objectives.

PRISMA-Based Systematic Review

To achieve first objective, a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework was used to systematically identify, screen, and select relevant literature on AI ethics, hackathon-based innovation, and higher education.

Database Selection: Scopus was chosen for its comprehensive indexing of high-impact journals, conference proceedings, and book chapters related to AI-Generated Text Detection Tools and Their Reliability.

Search Strategy: A Boolean search query was designed with keywords such as "Artificial Intelligence (AI)," "Generative AI," and “"AI Detector" or "Text Generated"”, “AI Ethics” refining results using filters for document type, publication year (2020–2025), and language (English only).

Inclusion and Exclusion Criteria:

Included: Peer-reviewed journal articles, conference papers, and book chapters explicitly discussing AI-driven as an tool for AI-Generated Text Detection Tools and Their Reliability.

Excluded: Non-English publications, grey literature, opinion pieces, and non-peer-reviewed sources.

Screening and Selection: The PRISMA flowchart guided the document filtering process, ensuring a rigorous and reproducible review.

The study inclusion criteria included: (i) the articles that were published in open-access journals, presented at academic conferences, full edition, and published since 2020; (ii) the research papers whose article title, abstract, and keywords matched the keywords of “Artificial Intelligence” or “AI”, “Generative AI” and “AI-Generated Text Detector”, and (iii) the research papers that related to “AI Ethics”.

The keywords used included

(TITLE-ABS-KEY ("Artificial Intelligence" OR "AI") AND PUBYEAR > 2023 AND PUBYEAR < 2025) AND ("generative ai") AND ("ai detector" OR "text generated") AND (“ai ethics”) AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp") OR LIMIT-TO (DOCTYPE , "bk")) AND (LIMIT-TO (SRCTYPE , "j") OR LIMIT-TO (SRCTYPE , "p")) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (OA , "all"))

Bibliometric Analysis

To fulfill second objective, a bibliometric analysis was conducted to examine publication trends, citation networks, leading research contributors, and subject distribution in Artificial Intelligence (AI), Generative AI, ,AI-Generated Text Detector, and AI Ethics

Data Extraction: The refined dataset from Scopus was exported for analysis, including metadata (titles, abstracts, authors, affiliations, citations, and keywords).

Tools for Analysis:

VOSviewer: Used for co-occurrence network visualization of keywords and research clusters.

R Bibliometrix Package: Applied for descriptive statistics, citation analysis, and collaboration mapping.

PRISMA Flowchart: Ensured systematic literature screening and selection.

Analytical Dimensions: Publication Trends - Number of papers published per year and their citation impact. Source Impact - Leading journals, conferences, and book publishers. Geographical Distribution - Countries contributing significantly to Artificial Intelligence (AI), Generative AI, ,AI-Generated Text Detector, and AI Ethics research. Co-Authorship and Citation Networks - Identifying influential researchers and collaborative networks.

Conduct the bibliometrics analysis of research related to Artificial Intelligence (AI), Generative AI, ,AI-Generated Text Detector, and AI Ethics.

Choose type of data: Create a map based on bibliographic data: Choose this option to create a co-authorship, keyword co-occurrence, citation, bibliographic coupling, or co-citation map based on bibliographic data.

Choose data source: Read data from reference manager files: Supported file types: RIS, EndNote, and RefWorks.

Choose type of analysis and counting method:

Type of analysis: Co-occurrence

Unit of analysis: Keywords

Counting method: Full counting

Choose threshold: Minimum number of occurrences of a keyword: 3 of the 419 keywords, 35 meet the threshold.

Choose number of keywords: For each of the 35 keywords, the total strength of the co-occurrence links with other keywords will be calculated. The keywords with the greatest total link strength will be selected. Number of keywords to be selected: 27

Items Filter = 27 items (5 clusters): Cluster 1 (9 items), Cluster 2 (6 items), Cluster 3 (5 items), Cluster 4 (5 items), and Cluster 5 (2 items)

5. Findings And Result

This section presents the findings obtained from the integration of two complementary analytical approaches: the PRISMA-based Systematic Review and the Bibliometric Analysis. The PRISMA process was employed to ensure a transparent and rigorous selection of relevant studies concerning AI-Generated Text Detection Tools and Their Reliability, resulting in a refined set of research that meets established inclusion criteria. This systematic method guarantees methodological reliability and minimizes bias in synthesizing existing literature. In parallel, the Bibliometric Analysis provides a quantitative exploration of publication trends, author collaborations, and thematic relationships within the field. Through bibliometric mapping, key research clusters are identified—linking concepts such as Artificial Intelligence, Generative AI, ChatGPT, and AI Ethics. Together, these approaches offer both depth and

breadth: the PRISMA framework contributes a structured synthesis of evidence, while the bibliometric analysis reveals the intellectual structure and evolving research landscape of AI-generated text detection. The combined results illuminate how the reliability of AI detection tools has been conceptualized and evaluated across studies, highlighting not only technical precision but also the growing significance of ethical and human-centered considerations in ensuring trustworthy AI applications.

PRISMA-Based Systematic Review Analysis

This section presents the findings from both PRISMA-Based Systematic Review and Bibliometric Analysis. The PRISMA process systematically identified, screened, and included relevant studies to ensure transparency and methodological rigor, resulting in a refined corpus of research on AI-generated text detection and its reliability. Meanwhile, the bibliometric analysis visualized publication trends, influential authors, keyword networks, and collaborative patterns across databases. Together, these results provide an integrated view of how the field has evolved—highlighting methodological progress, ethical awareness, and emerging research gaps that inform the development of reliable and responsible AI detection systems.

Preferred Reporting Items for Systematic reviews and Meta Analyses (PRISMA) of Detecting the Detectors: A Systematic Review of AI-Generated Text Detection Tools and Their Reliability

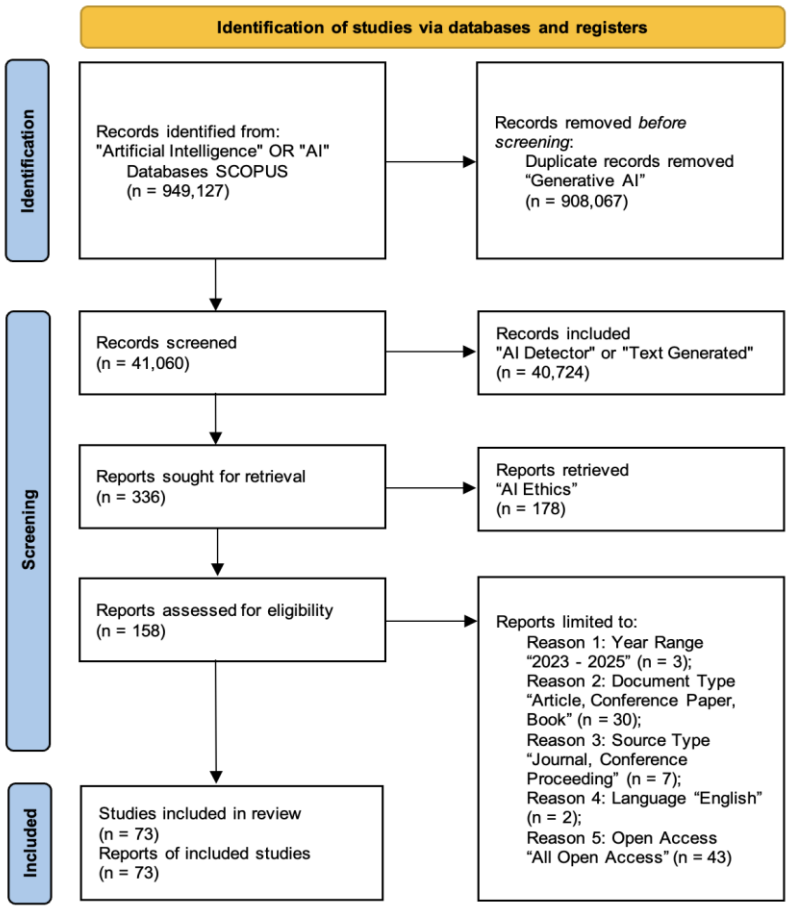

Figure 2 The PRISMA of the Detecting the Detectors: A Systematic Review of AI-Generated Text Detection Tools and Their Reliability

Figure 2 illustrates the PRISMA flow process ensuring the reliability and ethical validity of studies on *AI-Generated Text Detection Tools*. From 949,127 records identified, duplicates and irrelevant items were removed, and 41,060 were screened. Reports related to *AI Detectors* and *AI Ethics* were evaluated using strict inclusion criteria—year (2023–2025), document type, peer review, language, and open access. Finally, 73 studies were included. This systematic process strengthens confidence in the findings by emphasizing both the technical reliability of AI detection tools and their ethical integrity in ensuring fairness and transparency.

Bibliometric Analysis

This section summarizes the bibliometric analysis conducted to complement the systematic review. The analysis quantitatively explores research trends, publication growth, and collaboration networks related to AI-Generated Text Detection Tools and Their Reliability. Data extracted from the SCOPUS database were used to identify key authors, countries, and keywords, providing an overview of how this research area has evolved and where future studies may be directed.

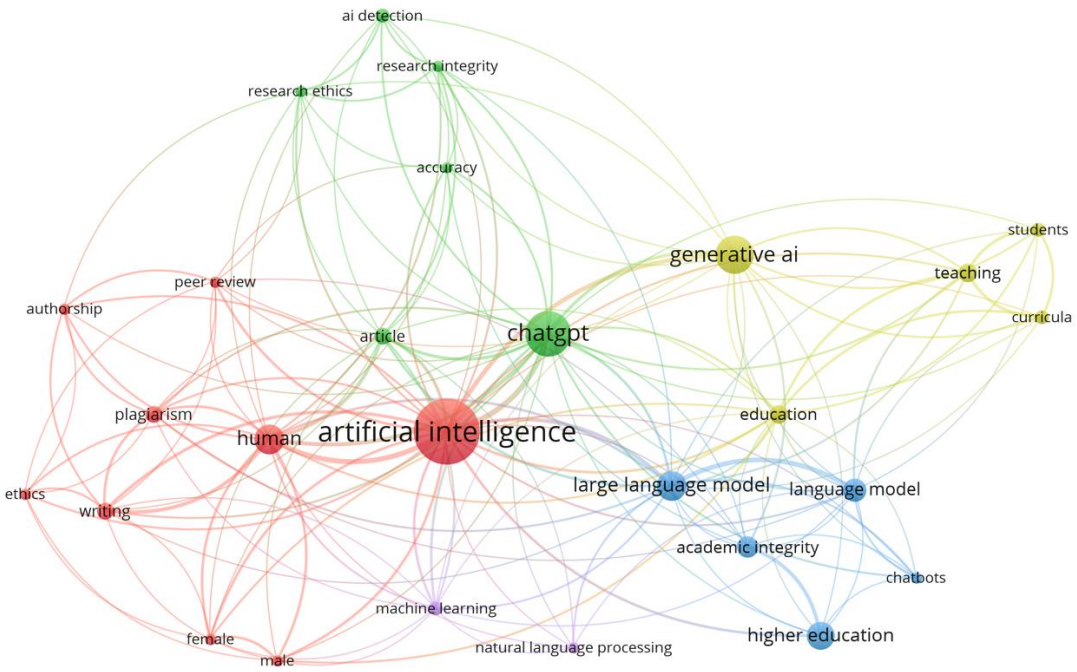

Figure 3 The PRISMA of the Detecting the Detectors: A Systematic Review of AI-Generated Text Detection Tools and Their Reliability

Figure 3 illustrates the bibliometric network visualization of research keywords related to *AI-Generated Text Detection Tools and Their Reliability*. The interconnected nodes represent major research themes such as artificial intelligence, ChatGPT, generative AI, and AI detection, highlighting the multidisciplinary nature of this field.

The strong links between *AI detection*, *research integrity*, and *research ethics* indicate growing scholarly attention to the ethical reliability of AI-based detection systems. Similarly, connections with terms like *accuracy*, *peer review*, and *plagiarism* reflect ongoing concerns

about ensuring the trustworthiness and transparency of AI detectors used to verify academic integrity.

Clusters surrounding *artificial intelligence* and *ChatGPT* also reveal overlaps between technical development and ethical accountability, suggesting that the reliability of AI-generated text detectors depends not only on algorithmic precision but also on adherence to principles of fairness, explainability, and responsible AI use. Overall, the visualization underscores that advancing reliable detection tools requires a balanced integration of technical validation and AI ethics.

6. Conclusion

This study systematically examined the landscape of AI-Generated Text Detection Tools through a PRISMA-based systematic review combined with bibliometric analysis to ensure methodological transparency and evidence-based insights. The PRISMA process refined a large corpus of literature into 73 high-quality studies, emphasizing the need for reliability, fairness, and accountability in AI detection research.

The findings reveal that while recent detection models—such as BERT-based and graph neural network approaches—demonstrate strong technical accuracy, their performance remains inconsistent across languages and contexts. This highlights that the reliability of AI detectors cannot rely solely on algorithmic performance, but must also integrate ethical considerations, including transparency, explainability, and research integrity.

The bibliometric mapping further underscores the interdisciplinary nature of this field, connecting artificial intelligence, ChatGPT, academic integrity, and AI ethics as central research clusters. This convergence shows an increasing focus on ensuring that AI detection tools are not only technologically robust but also ethically responsible in mitigating risks such as bias, plagiarism, and misinformation.

Overall, the synthesis of systematic and bibliometric evidence affirms that the development of trustworthy AI-generated text detectors requires a balanced integration of technical validation and ethical governance. Future research should prioritize open, interpretable, and ethically aligned detection systems to strengthen public confidence and academic integrity in the era of generative AI.

Acknowledgments

The authors would like to thank Suan Sunandha Rajabhat University, Bangkok, Thailand to provide funding support to attend the dissemination of research on this and thank family, friends, colleagues, students in the field of Educational Innovation and Technology, Faculty of Education and also The Office of General Education and Innovative e-Learning for cooperation and provide the dataset in research, all of you.

References

Abbas, H. M. (2025). A novel approach to automated detection of AI-generated text. مجلة القادسية لعلوم الحاسبات والرياضيات, 17(1). https://doi.org/10.29304/jqcsm.2025.17.11958

An Overview of Artificial Intelligence Ethics: Issues and Solution for Challenges in Different Fields. (2023). Journal of Artificial Intelligence and Capsule Networks, 5(1), 69–86. https://doi.org/10.36548/jaicn.2023.1.006

Bayan, F. M. H. (2024). The ethics of AI: Navigating the moral dilemmas of artificial intelligence. https://doi.org/10.36571/ajsp661

Bharti, I., Chauhan, K., & Aggarwal, P. (2024). Generative AI. Advances in Linguistics and Communication Studies, 1–36. https://doi.org/10.4018/979-8-3693-9246-1.ch001

Chaiwchan, P., Kaewrattanapat, N., & Nookhong, J. (2025). Students' perceptions of educational public relations using artificial intelligence conversation agent technology. In Proceedings of the ACE-2A25: Actual Economy – Afro-Asian Solutions to Global Challenges 2025 (pp. 148–152). The Eurasia Proceedings of the Eurasian Conferences on Educational Research.

Chakraborty, S., Bedi, A. S., Zhu, S., An, B., Manocha, D., & Huang, F. (2023). On the possibilities of AI-generated text detection. arXiv.org, abs/2304.04736. https://doi.org/10.48550/arXiv.2304.04736

Chaudhary, J., Parmar, N., & Mehta, A. (2024). Artificial intelligence and expert systems. International Journal of Advanced Research in Science, Communication and Technology, 535–546. https://doi.org/10.48175/ijarsct-15988

Chow, W. (2024). Generative AI. ASCILITE Conference Proceedings, 330–335. https://doi.org/10.14742/apubs.2024.730

Gao, D. K., Haverly, A., Mittal, S., Wu, J., & Chen, J. (2024). AI ethics. International Journal of Business Analytics. https://doi.org/10.4018/ijban.338367

Kazim, E., & Koshiyama, A. (2020). A high-level overview of AI ethics. Social Science Research Network. https://doi.org/10.2139/SSRN.3609292

Kennedy, H., & Wanless, L. (2022). Artificial intelligence. In M. L. Naraine, T. Hayduk III, & J. P. Doyle (Eds.), The Routledge Handbook of Digital Sport Management (pp. 333–345). Routledge. https://doi.org/10.4324/9781003088899-29

Morandín-Ahuerma, F. (2022). What is artificial intelligence? International Journal of Research Publication and Reviews, 03(12), 1947–1951. https://doi.org/10.55248/gengpi.2022.31261

Sarzaeim, P., Doshi, A., & Mahmoud, Q. H. (2023). A framework for detecting AI-generated text in research publications. Proceedings of the International Conference on Advanced Technologies. https://doi.org/10.58190/icat.2023.28

Sowmiya, B. (2025). Generative AI. Advances in Computational Intelligence and Robotics Book Series, 135–156. https://doi.org/10.4018/979-8-3693-5623-4.ch006

Suresh, S., & Sasidharan, C. (2025). Artificial intelligence. International Scientific Journal of Engineering and Management, 04(03), 1–7. https://doi.org/10.55041/isjem02419

Taşabat, S. E. (2025). The revolution of generative AI. Advances in Computational Intelligence and Robotics Book Series, 149–178. https://doi.org/10.4018/979-8-3373-0735-0.ch004

Wang, H., Li, J., & Li, Z. (2024). AI-generated text detection and classification based on BERT deep learning algorithm. Theoretical and Natural Science, 39(1). https://doi.org/10.54254/2753-8818/39/20240625