Available online at
https://mijuitm.com.my

e-ISSN: 2735-0703

Mathematical Sciences and Informatics Journal 6(2) 2025, 132 – 151.

Mathematical Sciences and Informatics Journal

Global Trends and Governance Insights in AI Safety: A Bibliometric Analysis

Nor Nashrah Azmi[1,2*], Fiza Abdul Rahim[1], Noor Hafizah Hassan[1]

[1]*Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Kuala Lumpur, Wilayah Persekutuan Kuala Lumpur, Malaysia*
[2]*College of Computing and Informatics, Universiti Tenaga Nasional, Kajang, Selangor, Malaysia*

ARTICLE INFO

Article history:
Received 30 August 2025
Revised 16 September 2025
Accepted 30 September 2025
Online first
Published 31 October 2025

Keywords:
AI Safety
Safe AI
Artificial Intelligence
Bibliometric
Scopus
VOSViewer

DOI:
10.24191/mij.v6i2.9041

ABSTRACT

This study provides a comprehensive bibliometric analysis of AI safety through a Scopus-indexed literature analysis between 1995 and June 2025, with VOSviewer applied to generate bibliometric maps and network visualizations of co-authorship, keyword co-occurrence, and citation metrics. AI safety research experienced rapid growth starting from 2020, with 81.58% of all publications emerging over the last five years. The research field of AI safety primarily relies on computer science, engineering, and mathematics, with the United States and the United Kingdom serving as its primary contributors. Research clusters in the field encompass technical areas such as reinforcement learning and adversarial robustness, alongside ethical governance and long-term risk. The study reveals disciplinary and geographic gaps, underscoring the need for global participation in this field. Quantitative analysis shows 3,971 citations, resulting in an h-index of 32 and a g-index of 46. Collaboration analysis indicates an average of 3.55 authors per paper, with the strongest co-authorship networks linking the United States, the United Kingdom, and Germany. The research provides valuable insights into current trends and suggests additional areas for investigation. The research establishes a vital, data-driven framework that supports researchers, policymakers, and funding agencies in advancing AI safety studies while creating comprehensive, responsible AI deployment strategies. The study provides empirical evidence to inform future interdisciplinary studies and help establish AI governance approaches and promoting the development of responsible AI safety worldwide.

[1,2*] Corresponding author. *E-mail address*: nashrah@uniten.edu.my
https://doi.org/10.24191/mij.v6i2.9041

1. INTRODUCTION

The 21st century has seen Artificial Intelligence (AI) become one of the most revolutionary technologies, promising to enhance human existence while transforming various sectors. The rapid development of AI technologies creates substantial ethical concerns, along with system breakdowns and risks to society. AI system safety and human value alignment have become an essential worldwide concern for researchers, policymakers, and practitioners. The field of AI safety research has garnered substantial attention as an interdisciplinary area of research due to these concerns. The research field investigates both the technical aspects of robustness and reliability, as well as governance frameworks and ethical principles that guide the responsible deployment of AI technology. The evolution of AI safety research necessitates analysis, as it reveals current knowledge gaps, collaborative patterns, and emerging concepts that define the field.

This research employs bibliometric analysis to examine AI safety publications listed in the Scopus database between 1995 and 2025, providing a quantitative account of research activity and impact. Bibliometrics can be used to determine patterns of publication, citation, authors, and institutions to determine research trends, popular works, and co-authoring networks. The approach is essential for comprehending the development and success of new areas, such as AI safety, and for identifying knowledge gaps and future research perspectives. The paper will also explore potentially new areas and knowledge gaps using keyword and thematic analysis (Haruna et al., 2024). Visualization of intellectual structures can be facilitated using bibliometric tools, enabling researchers, policymakers, and funding organizations to make informed, evidence-based decisions. It offers a complete and objective view at academic achievements (Russell, 2023). The study utilizes VOSviewer to visualize publication patterns, co-authorship networks, and keyword clusters, thereby facilitating an understanding of the intellectual structure of the field.

Previous bibliometric studies on artificial intelligence have focused mainly on specific subfields such as neuroscience (Tekin & Dener, 2025), cybersecurity (Chadha et al., 2024), and healthcare applications (Haruna et al., 2024). However, very few have systematically analysed the dedicated domain of AI safety. As a result, the interdisciplinary and governance dimensions of AI safety remain underexplored in bibliometric literature. This study is significant because it extends bibliometric analysis beyond technical subfields to explicitly cover AI safety as a stand-alone research area, highlighting its growth, collaboration patterns, and intellectual structure. By doing so, it addresses the gap in understanding how AI safety research has developed differently from adjacent AI domains and why global, cross-disciplinary engagement is urgently needed. The research also aims to inform future investigations while fostering international partnerships to develop ethical AI safety frameworks. The synthesis of research findings will deliver essential insights to stakeholders seeking identify new research areas and make evidence-based decisions about AI safety policy and funding. Our review examines the development patterns of AI safety and current research areas and predicts future research directions. The paper uses this format to fulfil its objectives. The second part of this paper reviews the existing literature on AI safety. The research methodology section describes the data sources used in the study. The bibliometric results in Section 4 include publication patterns, co-authorship networks, keyword analysis, and citation metrics. The final section presents the conclusions.

2. LITERATURE REVIEW

AI safety refers to the processes and safeguards that aim to ensure that artificial intelligence systems operate as intended and pose minimal risks to people or the planet. The area addresses both implementation issues and infrastructure issues, including robustness (systems operate reliably under various, possibly adversarial, conditions), assurance (human operators can analyse and understand system behaviour), and specification (system behaviour aligns with the designers' intent). These fundamental ideas drive research aimed at

https://doi.org/10.24191/mij.v6i2.9041

avoiding unexpected outcomes and emergent behaviours in AI systems that are becoming increasingly autonomous and complex. The cornerstones of safe and reliable deployment of AI are robustness, assurance, and specification (Tamascelli et al., 2024).

2.1 Historical Development of AI Safety

The evolution of AI safety as a field of inquiry has a strong connection to the history of artificial intelligence technologies and the growing awareness of the risks and challenges associated with their introduction. Early-stage AI safety was not a developed field, but an assortment of concerns within the broader body of computer science and engineering research. With the increasing adoption of machine learning systems and AI systems in the late 20th and early 21st centuries, more researchers began to design explicitly around concerns such as system robustness, reliability, and the avoidance of unintended consequences (Gou et al., 2022). The capacity of machine learning techniques to capture uncertainty, describe soft associations among variables, and discover hidden structures in the data makes them invaluable to safety and reliability communities, but also exposes them to new risks and weaknesses that need to be addressed systematically (Bautista-Bernal et al., 2024).

The study of AI safety has rapidly expanded over the past few years, coinciding with an increase in the scope of AI applications and sector penetration, including healthcare, transportation, and security. Early efforts focused on addressing technical issues, including ensuring that AI systems could operate reliably under diverse conditions and withstand adversarial attacks. With time, the context of AI safety has developed beyond ethical, legal, and societal concerns, especially as AI started intervening in sensitive areas of human activity (Kunal & Patkar, 2023). Emerging methodologies and frameworks also paralleled this widening of the scope to evaluate and advance the safety of AI, such as robust classification algorithms, adversarial robustness strategies, and techniques to improve interpretability and domain generalization (Chadha et al., 2024).

Today, AI safety is recognised as an emergent field that encompasses the skills of computer science, engineering, ethics, law, and the social sciences. The historical trend points from more reactive measures against technical failures to more proactive and science-based methods of developing robust, transparent, and accountable AI systems. The global conversation about AI safety has gained rapid momentum through recent advancements. The European Union passed the AI Act in 2024 to create formal regulations that focus on managing high-risk AI system risks through enhanced transparency measures, human oversight, and accountability standards (Smuha, 2025). The research community has devoted more attention to Explainable AI (XAI) and AI alignment, as these topics ensure that large-scale generative models align their goals with human values (Amodei et al., 2016). The current trends demonstrate the urgent requirement to track AI safety research development across technical, ethical, and governance aspects.

2.2 Previous Bibliometric Studies in AI and Related Fields

Bibliometric analysis was chosen in this study to trace the history and track the research topics in several subdisciplines of artificial intelligence, which has proven to be a rich source of information about the growth, development, cooperation patterns, and intellectual organization of these fast-moving fields (Tekin & Dener, 2025). The analysis underscored the innovative contributions of the United States, China, and the United Kingdom, and the ubiquity of cross-border cooperation (Luka et al., 2024). Other issues, challenges, and opportunities related to AI research, such as ethical concerns, data privacy concerns, and model interpretability, were also highlighted. On the same note, a bibliometric study of AI and security research in publications between 2020 and 2024 found a sustained increase in interest surrounding deep learning, machine learning, and security frameworks. These results underlined the importance of the facilitation of strong security standards to reduce potential risks of AI integration into key systems like healthcare and finance (Haruna et al., 2024).

https://doi.org/10.24191/mij.v6i2.9041

Bibliometric studies have been crucial in monitoring the spread of AI technology and its implications on medical practice in the field of healthcare. Researchers have recorded the fast development of work in artificial intelligence, the leading role of peer-reviewed journals and conference proceedings as publications, and the growing contribution of global research teams (Tekin & Dener, 2025). The interdisciplinary character of AI research, encompassing computer science, medicine, and engineering, is also reflected in these analyses, as well as the need to consider ethical and regulatory issues continually. Bibliometric analysis of AI in the Internet of Medical Things (IoMT), as one example, plotted the terrain of the research in terms of trending topics, high-output researchers, impactful institutions, and networks of collaborative authors (Herman, 2023).

2.3 Identified Gap and Contribution of the Present Study

Previous bibliometrics focus on technical innovations and collaborations but ignore global governance issues or superintelligent philosophical debates (Tekin & Dener, 2025). Similarly, cybersecurity bibliometrics are concerned with adversary attacks and intrusion detection without regard to topics such as accountability, transparency, or global policy regimes (Luka et al., 2024). These gaps indicate that most of the intellectual and practical roots of AI safety, which are social and technical, remain unmapped. This is particularly important because AI safety, unlike other areas in AI, necessarily requires interdisciplinary cooperation. Technical scientists need to work alongside ethicists, jurists, policymakers, and social scientists so that AI systems function not only as intended but also for the good of humankind in a safe, transparent, and equitable manner (Egghe, 2006; Hirsch, 2005). Unless the field is mapped bibliometrically, it becomes difficult to determine if and where such interdisciplinarity takes place, and if so, were, in which institutions, or by which authors. Additionally, earlier bibliometric research lacks an open understanding of spatial imbalances. It is now apparent that most AI research today present times is concentrated in the United States, the United Kingdom, and other Western nations, but no earlier bibliometric research has quantified the contributions of Africa, Southeast Asia, and Latin America to AI safety research. This is a critical flaw, since global cooperation is required in creating inclusive and culture-sensitive governance structures.

In trying to overcome these shortcomings, the present study introduces a valuable contribution. It maps out AI safety research as an independent subject of research, and not a subservient subcomponent of applied AI research. It not only isolates the technical clusters, such as adversarial robustness and reinforcement learning, but also the governance and ethical clusters, and in doing so, illustrates how these clusters overlap and shift over time (Haruna et al., 2024). In doing this, the study explains the intellectual landscape of AI safety, quantifies its global scope, and proposes the interdisciplinary networks that constitute it. This systematic interdisciplinary mapping is part of the larger bibliometric literature. It provides researchers, policymakers, and practitioners with an evidence-based overview of the development of AI safety to date, where gaps in current work are located, and where opportunities exist for increased diversity and cross-disciplinary progress.

3. MATERIALS AND METHODS

3.1 Data Collection and Cleaning

This paper examines the trends and productivity of research on AI safety using bibliometric analysis based on published documents indexed in the Elsevier Scopus database from 1995 to June 20, 2025. Some of the bibliometric indicators and network visualization are presented in this paper. The search topic was “AI Safety” in the publication section with no restriction on publication years of the article. We accessed bibliographic data used in this study from the Scopus database because it is the largest multidisciplinary database of peer-reviewed literature in social science research. Scopus is also widely recognized and frequently accessed for quantitative analyses (Durán-Sánchez et al., 2019). The published documents on

https://doi.org/10.24191/mij.v6i2.9041

AI safety are identified by executing a search query (TITLE-ABS-KEY (("AI Safety" OR "safe AI")) with no limit on publication years, based on the keywords in the paper title. The 608 most cited publications were selected, and the records of published documents were retrieved in a Comma Separated Values (CSV) format for screening. The cleaning process involves removing non-conventional and unrelated documents from the search. Data cleaning also standardised author names, institutional affiliations, and keyword synonyms to ensure accurate network mapping.

3.2 Bibliometric Parameters and Tools

The following bibliometric parameters of each article were analysed: publication title, citation count, citation density (the average number of citations per annum), publication year, authorship, country and institution of origin, topic of interest, and keywords. We also utilised VOSViewer (version 1.6.2) to construct collaborative networks among authors and identify frequently occurring keywords among authors. VOSviewer utilises two standardized weights, including the number and total strength of the links, to visualize the nodal network graphically. The size of the nodes and the interlinking lines connecting the nodes denote the relevance and strength of the links.

To improve the transparency of the analysis, the study recorded all the thresholds and parameter values in VOSviewer directly because such choices have a significant impact on the robustness and replicability of bibliometric results. The study employed a minimum threshold of five keyword occurrences for keyword co-occurrence analysis. This threshold was chosen because it is a compromise between inclusiveness and specificity. With respect to author co-authorship, two or more of the authors' works had to be present in the network. This cut-off was selected to emphasize prolific scholarly writers instead of infrequent writers. When examining co-authorship at the national level, the break point was established at three papers per country. This filter ensured that the map showed only nations with a significant contribution to AI safety research, excluding those that contributed nothing or only by chance.

The second methodological decision was to employ fractional and full counting. Full counting assigns the total weight of a paper or citation to all authors, institutions, or nations involved, which may overly amplify the impact of highly collaborative work. By contrast, fractional counting distributes credit proportionally between contributors, which evens the score in favour of larger teams or more productive institutions. In this study, fractional counting was used for co-authorship analysis, while full counting was used for keyword co-occurrence. Then, we present productivity and impact in the form of an h-index and a g-index (Egghe, 2006; Hirsch, 2005; Tsay, 2009). Using fractional counting, the study ensures that smaller institutions, less productive authors, and less influential areas are more accurately represented and thus provides a more even representation of the interdisciplinarity and worldwide character of the research. Through balanced utilization of these thresholds and fractional counting, bibliometric mapping seeks to achieve a balance between inclusiveness, openness, and fairness (Haruna et al., 2024; Tekin & Dener, 2025).

3.3 Structured Steps for Reproducibility

To ensure the greatest reproducibility, the study was carried out using a systematic stepwise procedure that other researchers could easily replicate. Fig. 1 illustrates the workflow diagram, providing a visual overview of the process. Instead of a PRISMA diagram, which typically emphasizes exclusions, a tailored seven-step workflow diagram was developed to summarize the bibliometric procedure. This visual enhances clarity and aligns directly with the study design. The following are the six steps used:

Step 1: Scope definition. The investigation scope was determined using the Scopus database because it is multidisciplinary and contains an optimum index of peer-reviewed articles. The time frame was specified as 1995–2025, and the words "AI safety" and "Safe AI" were used as search terms in the field "title, abstract, and keywords".

https://doi.org/10.24191/mij.v6i2.9041

Step 2: Exclusion and inclusion criteria. Duplicates were also deleted manually and automatically via Scopus checks. The exclusion process solely selected peer-reviewed and original work on the subject, resulting in the final dataset.

Step 3: Data cleaning and export. The cleaned data was exported in CSV format. Data cleaning was applied to rectify inconsistencies, such as institutional affiliation ("MIT" vs "Massachusetts Institute of Technology"), and keywords ("AI safety" vs "Artificial Intelligence safety").

Step 4: Bibliometric Mapping and Analysis. Pre-cleaned data were tabulated for analysis and imported into VOSviewer. Parameters of analysis were clearly documented, and these cut-offs sacrifice inclusivity for analytical accuracy so that results can be compared with other comparable bibliometric studies.

Step 5: Citation analysis. The calculations of citation-based indices like total citations, mean citations per paper, h-index, and g-index were performed with Publish or Perish software.

Step 6: Documentation. For maximum transparency, search terms, filtering options, threshold levels, and counting methods were all documented. This allows other researchers to reproduce the study under similar conditions or alter parameters for comparative research.

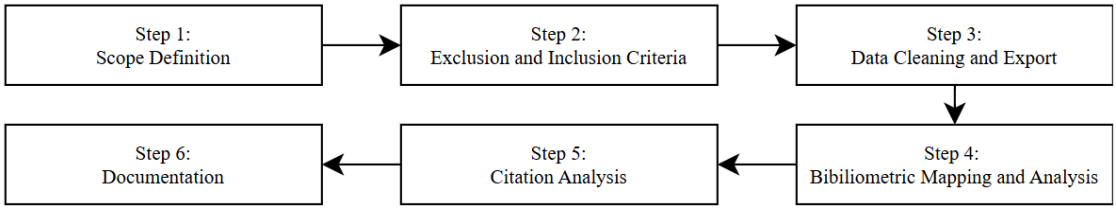

Fig. 1. Workflow diagram

4. RESULTS AND DISCUSSION

The results section below provides an in-depth discussion of Scopus-indexed works relating to AI safety. The key indicators include annual trends, types of documents, subject areas, and worldwide distribution. This section outlines the explosive rise of research, institutions, and authors who have had a significant impact, as well as collaboration and thematic networks, all in a data-driven overview of the changing state of AI safety.

4.1 Year of Publications and Annual Trends

Table 1 displays the distribution of 608 annual publications, illustrating the trends over time. The statistics indicate a sharp increase in the number of publications over the past few years, with 2024 and 2025 accounting for 32.57% and 19.74% of the total, respectively. The 2025 publications are fewer than those in 2024, as the data extraction for this study was conducted in June 2025, i.e., half of the year. However, the results show that the number of publications in 2025 has already reached more than half of the 2024 total, even after just half a year. The drastic rise indicates an interest or focuses on the topic, which may be explained by new trends, technology, or better funding of research may explain. Notably, the years 2023, 2022, and 2021 also account for a substantial share of publications, with 14.14%, 7.40%, and 7.73% of the total share, respectively, suggesting continued academic and industry interest.

Conversely, the number of previous years (2011-2020) is comparatively low each year, contributing less than 6.25 percent. The lowest number of publications was recorded in 2011, 2013, 2014, and 2016, with fewer than three publications each. The fact that one of the publications dates to as early as 1995 may indicate a long-standing interest, but it remained relatively niche in nature until now. More than 80 percent of publications were created within the past six years (2020-2025), suggesting a significant increase in

https://doi.org/10.24191/mij.v6i2.9041

research activity. This is indicative of larger changes in academic priorities, the availability of new methodologies, or improvements in access to publishing platforms. In general, the figures indicate a rapidly developing area with a new surge of research activity.

Table 1. Annual growth trend of publications in Scopus (1995–2025)

Year	Number of Publications	Percentage (%) (N=608)	Cumulative Frequency (%)
2025	120	19.74	19.74
2024	198	32.57	52.30
2023	86	14.14	66.45
2022	45	7.40	73.85
2021	47	7.73	81.58
2020	38	6.25	87.83
2019	31	5.10	92.93
2018	19	3.13	96.05
2017	10	1.64	97.70
2016	2	0.33	98.03
2015	3	0.49	98.52
2014	2	0.33	98.85
2013	2	0.33	99.18
2012	3	0.49	99.67
2011	1	0.16	99.84
1995	1	0.16	100.00
Total	608	100.00	

4.2 Document and Source Types

The data in Table 2 clearly show that conference papers represent the most common type of document published, constituting 52.14% of all publications, followed by articles, which account for 31.25% of all publications. A less salient yet worthwhile contribution comprises other documentation forms, including book chapters (6.58%) and reviews (4.44%), with contributions to books, editorials, and letters being minimal. Regarding the types of sources, conference proceedings (42.93%) and journals (37.99%) are the most common publication domains, indicating an academic focus on peer-reviewed conferences and journals. These findings suggest that published documents on AI safety are adequately disseminated in various forms, enhancing the communication of scientific developments and the research impact of the topic. Multiple studies have emphasized that unhindered access to research findings on scientific topics increases readership, citations, and the application of such findings, which subsequently enhances research impact (Niyazov et al., 2016).

Table 2. Types of documents published

Document Type	Number of Publications	Percentage (%) (N=608)
Conference paper	317	52.14
Article	190	31.25
Book Chapter	40	6.58
Review	27	4.44
Conference Review	15	2.47
Note	8	1.32
Book	6	0.99
Editorial	3	0.49
Erratum	1	0.16
Letter	1	0.16
Total	608	100.00

In Table 3, book series (13.65%) and books (5.26%) are fewer in number, which implies that although books are being used to contribute to the field, they are not primarily being used to disseminate knowledge. The insignificant number of trade journals (0.16%) suggests that industry-specific publications are scarce in this dataset. Generally, the results indicate a heavy reliance on conferences and journals as primary

https://doi.org/10.24191/mij.v6i2.9041

knowledge-sharing mediums, with books and other formats used in secondary roles. Such dissemination patterns follow common academic publishing trends of rapid dissemination and peer review.

Table 3. Source type classification of publications

Source Type	Number of Publications	Percentage (%) (N=608)
Conference Proceeding	261	42.93
Journal	231	37.99
Book Series	83	13.65
Book	32	5.26
Trade Journal	1	0.16
Total	608	100.00

4.3 Subject Area

Table 4 shows the publications categorized by subject areas, indicating a significant emphasis on Computer Science (40.30%), as this subject area dominates the research landscape. Engineering (14.44%) and Mathematics (12.49%) come next, which strongly suggests an intense focus on technical and quantitative subjects. Arts and Humanities (6.82%) and Social Sciences (6.82%) are also represented, which indicates some interdisciplinary interaction. Others include Medicine (3.54%), Decision Sciences (2.92%), and Business (2.13%), among others, which occur less frequently but are also relevant. Energy, Materials Science, Physics, and Economics contribute less than 2 percent each.

Table 4. Subject area distribution with a minimum of ten publications

Subject Area	Number of Publications	Percentage (%)
Computer Science	455	40.30
Engineering	163	14.44
Mathematics	141	12.49
Arts and Humanities	77	6.82
Social Sciences	77	6.82
Medicine	40	3.54
Decision Sciences	33	2.92
Business, Management and Accounting	24	2.13
Energy	18	1.59
Materials Science	17	1.51
Physics and Astronomy	15	1.33
Economics, Econometrics and Finance	10	0.89

*The publications are classified based on the source title categorisation. Some documents are categorized as more than one subject area.

A more detailed examination of the disciplinary distribution yields clearer evidence for this disparity. Specifically, approximately 67% of the papers in the dataset were concentrated in STEM-related fields such as computer science (40.30%), engineering (14.44%), and mathematics (12.49%). In contrast, only about 13.6% of the publications were associated with the social sciences (6.82%) and arts and humanities (6.82%), while law and ethics appeared only as a marginal component within these categories. This quantification confirms that AI safety research is heavily skewed toward technical domains, with limited engagement from governance, ethical, and policy-related scholarship. Such an imbalance indicates that, although AI safety is inherently interdisciplinary, current contributions remain dominated by technical approaches like algorithmic resilience, adversarial robustness, and machine learning safety. To build a more holistic foundation, greater cross-disciplinary collaboration is required, particularly from ethicists, legal scholars, sociologists, and political scientists, to ensure that AI safety research integrates both technical and governance perspectives.

https://doi.org/10.24191/mij.v6i2.9041

4.4 Most Active Source Titles

Table 5 provides an overview of the most active source titles of publications, where Lecture Notes in Computer Science (8.22%) emerges as the most frequently used publication venue, followed by Central Europe Workshop Proceedings (CEUR-WS) (5.10%) and Advances in Neural Information Processing Systems (2.80%). The field heavily relies on conferences because it requires the rapid dissemination of findings to address the rapidly evolving AI technologies and safety concerns.

Table 5. The most active source title with more than three publications.

Source Title	Publisher	Source Types	Number of Publications	Percentage (%)
Lecture Notes in Computer Science including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics	Springer Science and Business Media	Conference Paper, Conference Review	50	8.22
Central Europe - Workshop Proceedings (CEUR-WS)	Central Europe - Workshop Proceedings (CEUR-WS)	Conference Paper	31	5.10
Advances in Neural Information Processing Systems	Neural Information Processing Systems Foundation	Conference Paper	17	2.80
Philosophical Studies	Springer Science and Business Media	Article	12	1.97
Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)	International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS)	Conference Paper	11	1.81
AI and Society	Springer Science and Business Media	Article	10	1.64
Proceedings of Machine Learning Research	ML Research Press	Conference Paper	8	1.32
Communications in Computer and Information Science	Springer Science and Business Media	Conference Paper	7	1.15
Institute of Electrical and Electronics Engineers (IEEE) Access	Institute of Electrical and Electronics Engineers (IEEE) Incorporated	Review, Article	7	1.15
Lecture Notes in Networks and Systems	Springer Science and Business Media	Conference Paper, Conference Review	6	0.99
2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)	Association For Computational Linguistics (ACL)	Conference Paper	5	0.82
Ethics of Artificial Intelligence	Oxford University Press	Book Chapter	5	0.82
Philosophies	Multidisciplinary Digital Publishing Institute (MDPI)	Article	5	0.82
Proceedings of the 2023 Advancement of Artificial Intelligence/Association for Computing Machinery (AAAI/ACM) Conference on AI Ethics and Society (AIES)	Association for Computing Machinery (ACM) Incorporated	Conference Paper	4	0.66
Big Data and Cognitive Computing	Multidisciplinary Digital Publishing Institute (MDPI)	Article	4	0.66
International Conference on Acoustics, Speech and Signal Processing (ICASSP) Proceedings	Institute of Electrical and Electronics Engineers (IEEE) Incorporated	Conference Paper	4	0.66
Journal of Artificial Intelligence and Consciousness	World Scientific	Article	4	0.66

Springer is a leading publishing company that contributes to various top-ranking sources, including Philosophical Studies (1.97%) and AI and Society (1.64%), among others, which focus on journal articles. Additional popular outlets, including Institute of Electrical and Electronics Engineers (IEEE) Access (1.15%) and Proceedings of the Machine Learning Research (1.32%), also underscore the importance of

https://doi.org/10.24191/mij.v6i2.9041

peer-reviewed conferences and journals in advancing research. The existence of specialized publications such as Ethics of Artificial Intelligence (0.82%) and Journal of Artificial Intelligence and Consciousness (0.66%) indicates niche but increasingly popular fields of interest. The variety of publishers, including Springer, IEEE, Multidisciplinary Digital Publishing Institute (MDPI), and Association for Computing Machinery (ACM), indicates a broad academic interest.

4.5 Distribution of Publications by Countries

A closer look at Table 6 confirms a clear geographic imbalance in AI safety research. The United States alone contributed 37.34% of all documents, followed by the United Kingdom with 18.91% and Germany with 9.05%. Together, these three countries account for over 65% of the total research output, underscoring the heavy concentration of scholarship in Western contexts. China (8.06%) and India (4.44%) combined produced around 12.5% of the publications, signalling a growing Asian presence, though contributions from Southeast Asia were negligible. The European nations, such as the Netherlands, Spain, France, and Switzerland, collectively contribute a significant share (12.67%), underscoring the European contribution to research development. In the meantime, representatives from the Asian continent, including South Korea, Singapore, and Japan, make a slight contribution, ranging between 2.63% and 2.80%, with Hong Kong and Israel also in the top 20. The larger consistent contributors are Poland (1.83-2.05%), Austria, Norway, and Finland (1.15-2.14%).

Table 6. Top 20 contributing countries by publication

Country	Number of Publications	Percentage (%) (N=608)
United States	227	37.34
United Kingdom	115	18.91
Germany	55	9.05
China	49	8.06
Australia	30	4.93
Canada	28	4.61
India	27	4.44
Netherlands	22	3.62
Spain	22	3.62
France	18	2.96
Singapore	17	2.80
South Korea	16	2.63
Italy	15	2.47
Switzerland	15	2.47
Austria	13	2.14
Hong Kong	11	1.81
Israel	11	1.81
Japan	10	1.64
Norway	8	1.32
Finland	7	1.15

The evidence suggests a Western-centric tendency, where the United States and Europe are leading, and Asia, particularly China, is exhibiting increased prominence. However, African countries accounted for only about 2% and Latin American countries for roughly 1% of the dataset, highlighting severe underrepresentation of the Global South. This imbalance shows that Western and select Asian nations largely shape AI safety debates, while voices from regions most affected by AI's social and economic consequences remain almost absent. Such limited geographic representation also raises concerns about inclusivity and equity in AI governance, as policies informed by a narrow set of regions may fail to capture diverse cultural, societal, and economic contexts. Thus, this highlights the importance of conducting further research on AI safety.

https://doi.org/10.24191/mij.v6i2.9041

4.6 Most Influential Institutions

Table 7 reveals the most productive institutions in the research field, with the University of Louisville (4.11%) and the University of Oxford (4.11%) producing the highest publication rates. Top-tier American universities, such as Carnegie Mellon University (2.80%), the University of California (2.63%), and Stanford (1.81%), follow closely, reflecting a high academic presence in America. A few other institutions, such as the James Breckenridge (J.B.) Speed School of Engineering (2.63%) also makes significant contributions.

Table 7. Most influential institutions with a minimum of seven publications

Institution	Country	Number of Publications	Percentage (%)
University of Louisville	United States	25	4.11
University of Oxford	United Kingdom	25	4.11
Carnegie Mellon University	United States	17	2.80
University of California, Berkeley	United States	16	2.63
James Breckenridge (J.B.) Speed School of Engineering	United States	16	2.63
Universiteit Utrecht	Netherlands	11	1.81
Stanford University	United States	11	1.81
Imperial College London	United Kingdom	11	1.81
Netherlands Organisation for Applied Scientific Research	Netherlands	11	1.81
University of Cambridge	United Kingdom	11	1.81
Nanyang Technological University	Singapore	10	1.64
Harvard University	United States	9	1.48
University of York	United Kingdom	9	1.48
University of Liverpool	United Kingdom	9	1.48
Technical University of Munich	Germany	8	1.32
Massachusetts Institute of Technology	United States	8	1.32
The University of Edinburgh	United Kingdom	8	1.32
Google Limited Liability Company (LLC)	United States	7	1.15
Peking University	China	7	1.15
DeepMind Technologies Limited	United Kingdom	7	1.15

The significant roles of the United Kingdom and the Netherlands are evident in European institutions, such as Universiteit Utrecht (1.81%), Imperial College London (1.81%), and the University of Cambridge (1.81%). Nanyang Technological University (1.64%) is an indicator of Singaporean strength, whilst Peking University (1.15%) is the only Chinese university in the top 20. The contributions of the private sector to research are evident, with notable examples including industry giants such as Google (1.15%) and DeepMind (1.15%). These Anglophone countries (the United States and the United Kingdom) and Western Europe's domination align with wider scholarly tendencies, yet the influence of Asia is on the rise. This allocation raises institutional prestige, unequal funds, and collaborative webs that determine how global research is produced.

4.7 Authorship Analysis

Table 8 displays that the most widespread types are single-authored and dual-authored works, with 19.24% and 19.24%, respectively, suggesting a balanced trend towards individual and small-group research. Articles by three authors occupy second place (17.93%), followed by articles with more authors, with decreasing percentages (4+ authors, 10.20%, and five authors, 11.35%). There are significantly fewer contributions with six authors, 6.25%, and then even fewer still with seven or more authors, 5.59% altogether. Exceptions to the commonality of authorship are also notable, with 17 posts (2.80%) lacking named authors (such as conference reviews). The 0 number of the author represents the conference review type, for which the author's information was omitted during the initial evaluation phase to maintain an unbiased assessment. The distribution suggests that large-team studies play a significant role in the field, although small-team research is more prevalent.

https://doi.org/10.24191/mij.v6i2.9041

Table 8. Authorship patterns based on the number of authors per publication

Number of Author(s)	Number of Publications	Percentage (%) (N=608)
1	117	19.24
2	117	19.24
3	109	17.93
4	62	10.20
5	69	11.35
6	38	6.25
7	28	4.61
8	19	3.13
9	6	0.99
10	5	0.82
11	8	1.32
12	4	0.66
13	2	0.33
14	1	0.16
15	1	0.16
17	1	0.16
20	1	0.16
24	1	0.16
0*	17	2.80
Total	608	100.00

*Conference review document. No author is listed for this type of document.

Table 9 summarizes the authors with the most significant impact in the field, where Yampolskiy, R.V., has the most impact with 25 publications (4.11%), representing substantial scholarly productivity. Notable other contributors are Aliman, N.M. (11 publications, 1.81%) and Kester, L. (10 publications, 1.64%), indicating their active position in research. Other authors, such as Huang, X., and Zhao, X., have equal publications of 9 (1.48 percent), indicating equal contributions. There are also several authors whose publications range between 4 and 7 (0.66 to 1.15 percent), indicating an extensive and selective representation of major researchers who influence the field. Although citation statistics are not provided here, the number of publications attests to the influence of these authors, with both collaborative and individual efforts facilitating improvements. Such a divide shows a combination of both well-established and new directions in the field.

Table 9. Most influential authors with a minimum of four publications

Author	Number of Publications	Total Citations
Yampolskiy, R.V.	25	4.11
Aliman, N.M.	11	1.81
Kester, L.	10	1.64
Huang, X.	9	1.48
Zhao, X.	9	1.48
Lam, K.Y.	7	1.15
Huang, W.	6	0.99
Waser, M.R.	6	0.99
Herrera, F.	5	0.82
Liu, Z.	5	0.82
Haider, T.	4	0.66
Hay, N.J.	4	0.66
Hernández-Orallo, J.	4	0.66
Homan, C.M.	4	0.66
Jiang, Y.	4	0.66
Kasirzadeh, A.	4	0.66
Sarma, G.P.	4	0.66
Ziesche, S.	4	0.66
De Witt, C.S.	4	0.66

https://doi.org/10.24191/mij.v6i2.9041

Fig. 2 displays the network visualization map of co-authorship by country, illustrating the patterns of international co-authorship in research collaborations. As shown in Fig. 2, there are robust clusters of concentrated contributions throughout the United States, the United Kingdom, Germany, and China, indicating their preeminent global research stewardship. The smaller yet busy networks feature India, Brazil, Italy, and South Korea, which shows new areas of collaboration hubs. It also demonstrates that major collaborations are primarily driven by Western nations and some Asian countries, with Europe serving as a bridge. The visualizations also highlight how the global academic cooperation is interconnected and unequal.

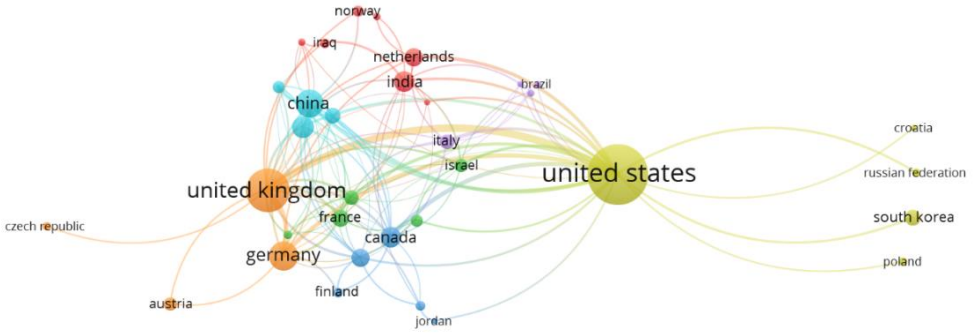

Fig. 2. Network visualization map of co-authorship by countries

Fig. 3 displays the network visualization map of the co-authorship by authors, illustrating the distribution of collaboration among researchers in the field. According to Fig. 3, Zhao Xingyu and Huang Xiaowei are central figures, indicating they are engaged in active collaborations. Smaller clusters centre around Flynn, David, Avin, Shahar, and Manheim, David, implying the presence of concentrated research teams. All in all, the maps reveal a choice of early-stage and mature researchers, with interdisciplinary relationships observable through the loosely connected clusters. The visualizations portray intimate teams, as well as the more general but intermittent groupings, more characteristic of academic networks. Beyond disciplinary dominance, the analysis also reveals meaningful but limited interdisciplinary interactions within research. Co-authorship patterns show that collaborations between computer scientists and ethicists/legal scholars accounted for less than 7% of the total co-authored papers, suggesting that while such partnerships exist, they remain rare compared to technical-only collaborations.

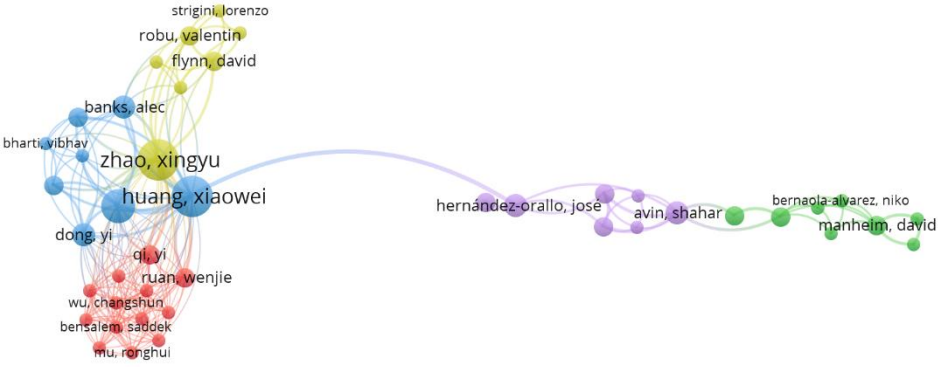

Fig. 3. Network visualization map of the co-authorship by authors

https://doi.org/10.24191/mij.v6i2.9041

4.8 Keywords Analysis

Figure 4 demonstrates that keyword co-occurrence networks reveal the primary topics in AI studies. AI and ChatGPT are the key nodes featured in both maps, as they are current topics of discussion. The cluster on "Ethics", "AI Governance", and "Responsible AI" is illustrated in Fig. 5. There is also an increased interest in ethical frameworks. In contrast, issues such as "Cybersecurity", "Adversarial Attacks", and "Robustness" express technical safety concerns. The highly weighted connection between "Existential Risk", "Superintelligence", and "Value Alignment" is illustrated in Fig. 5, which focuses on the intense debates surrounding long-term AI risks. Several new terms are embedded in each, including Large Language Models and Generative AI, which are trending. This citation network analysis also illustrates this divide: technical works are heavily cited within technical clusters, while governance-oriented studies form relatively isolated clusters, with less than 10% cross-referencing.

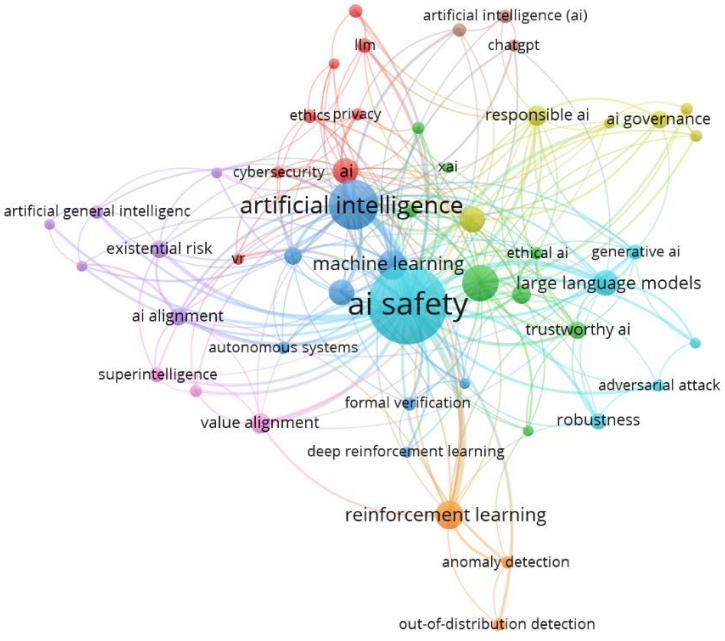

Fig. 4. Co-occurrence network showing thematic clustering of keywords related to AI safety

The most common author keywords are summarized in Table 10, with the most popular ones being "AI Safety" (36.68) and "Artificial Intelligence" (27.30), which is understandable given the current interest in the safe and ethical development of AI. Technical approaches, such as "Reinforcement Learning" (10.86%), "Deep Learning" (9.87%), and "Language Models" (9.05%), are characterized by the salience of machine learning approaches. High-priority keywords related to risk mitigation include "Safety Engineering" (8.55%), "Safe AI" (8.22%), and "Adversarial Machine Learning" (6.58%), whereas the ethical perspective can be seen in "AI Ethics" (4.11%) and "Risk Assessment" (4.61%). Other terms, such as Large Language Model (5.26%) and Computational Linguistics (4.44%), represent research specialties. The frequencies of the categories "Human" (5.26) and "Decision Making" (4.61) indicate interdisciplinary activity involving the human factor. This word distribution reflects the twofold emphasis of the field on both technical innovation and positive societal impact, with AI safety serving as the thread that ties them together.

https://doi.org/10.24191/mij.v6i2.9041

Table 10. Most frequently used author keywords

Author Keywords	Frequency	Percentage (%)
AI Safety	223	36.68
Artificial Intelligence	166	27.30
Reinforcement Learning	66	10.86
Deep Learning	60	9.87
Language Model	55	9.05
Safety Engineering	52	8.55
Safe AI	50	8.22
Machine Learning	47	7.73
AI Systems	42	6.91
Learning Systems	41	6.74
Adversarial Machine Learning	40	6.58
Reinforcement Learnings	38	6.25
Human	32	5.26
Large Language Model	32	5.26
Machine-learning	32	5.26
Decision Making	28	4.61
Risk Assessment	28	4.61
Computational Linguistics	27	4.44
Article	26	4.28
AI Ethics	25	4.11

4.9 Citation Analysis

Table 11 presents the key citation performance of the dataset, spanning the period from 1995 to 2025. This citation metric was generated by Publish or Perish software, which imports a Research Information Systems (RIS)-formatted file from the Scopus database to present the raw citation metrics. The average citations per paper (6.53) and the citations per year (132.37) of 6,071 citations in 608 papers show consistent academic impact. Prolonged influence is demonstrated by the h-index (32) and g-index (46), and via a selection of the most frequently cited outputs (such as 38 papers cited 10 or more times). Co-authorship is reflected in the 3.55 co-authors per paper and 1667.83 citations per author. It is worth mentioning that 286 papers (47%) are cited at least once, yet only 12 (2%) have received at least 20 citations, which fits a long-tail distribution. The annual rate (0.63) and normalized h-index (19) frameworks provide insight into productivity over time. Viewed together, these metrics paint a picture of a relatively influential research body, which balances general collaborations with a few high value works, indicative of a developing interdisciplinary field such as AI.

Table 11. Descriptive citation metrics of publications

Metrics	Data
Reference Date and Time	20/06/2025, 11:00 AM
Publication Years	1995-2025
Citation Years	30
Papers	608
Citations	3971
Citations/Year	132.37
Citations/Paper	6.53
Citations/Author	1667.83
Papers/Author	257.10
Authors/Paper	3.55
Hirsch (h-index)	32
Egghe (g-index)	46
Individual h-index (hI, norm)	19
Annualized h-index (hI, annual)	0.63
Papers with Annual Citation Count (ACC) >= 1,2,5,10,20	286, 180, 93, 38, 12

https://doi.org/10.24191/mij.v6i2.9041

Table 12 presents the 20 most frequently cited articles that represent critical areas in AI research, including safety, ethics, and governance. Topping the list is "Governing AI safety through independent audits" (2021, 104 citations), a study by a group of scholars published in Nature Machine Intelligence, which includes a call for regulatory frameworks due to the increased focus on accountability. Second is a paper in British Medical Journal (BMJ) Global Health, which addresses threats to human health and existence by AI (2023, 97 citations). In contrast, "Generalized out-of-distribution detection: a survey" (2024, 95 citations) examines the technical robustness of machine learning.

Table 12. Top 20 most cited articles in the dataset

No	Title	Authors	Source	Year	Cites	Cites per Year
1	Governing AI Safety through Independent Audits	Falco, Gregory; Shneiderman, Ben; Badger, Julia; Carrier, Ryan; Dahbura, Anton ; Danks, David; Eling, Martin ; Goodloe, Alwyn; Gupta, Jerry; Hart, Christopher; Jirotka, Marina; Johnson, Henric; Lapointe, Cara; Llorens, Ashley J.; Mackworth, Alan K.; Maple, Carsten; Pálsson, Sigurður Emil; Pasquale, Frank; Winfield, Alan; Yeong, Zee Kin	Nature Machine Intelligence	2021	104	20.80
2	Threats by Artificial Intelligence to Human Health and Human Existence	Federspiel, Frederik; Mitchell, Ruth; Asokan, Asha; Umana, Carlos; Mccoy, David	British Medical Journal (BMJ) Global Health	2023	97	32.33
3	Generalized Out-of-Distribution Detection: A Survey	Yang, Jingkang; Zhou, Kaiyang; Li, Yixuan; Liu, Ziwei	International Journal of Computer Vision	2024	95	47.50
4	AI Chatbots Not Yet Ready for Clinical Use	Au Yeung, Joshua; Kraljevic, Zeljko; Luintel, Akish; Balston, Alfred; Idowu, Esther; Dobson, Richard J.; Teo, James T.	Frontiers in Digital Health	2023	92	30.67
5	An AI Race for Strategic Advantage: Rhetoric and Risks	Cave, Stephen; Óhéigeartaigh, Seán S.	Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society	2018	80	10.00
6	A Multimodality Fusion Deep Neural Network and Safety Test Strategy for Intelligent Vehicles	Nie, Jian; Yan, Jun; Yin, Huilin; Ren, Lei; Meng, Qian	IEEE Transactions on Intelligent Vehicles	2021	76	15.20
7	Evaluating Attribution for Graph Neural Networks	Sanchez-Lengeling, Benjamin; Wei, Jennifer; Lee, Brian; Reif, Emily; Wang, Peter Y.; Qian, Wesley Wei; Mccloskey, Kevin; Colwell, Lucy; Wiltschko, Alexander	Advances in Neural Information Processing Systems	2020	76	12.67
8	Classification of Global Catastrophic Risks Connected with Artificial Intelligence	Turchin, Alexey; Denkenberger, David	AI and Society	2020	70	11.67
9	The Situation Awareness Framework for Explainable AI (SAFE-AI) and Human Factors Considerations for XAI Systems	Sanneman, Lindsay; Shah, Julie A.	International Journal of Human-Computer Interaction	2022	60	15.00
10	Artificial Superintelligence: A Futuristic Approachroman	Yampolskiy, Roman V.	Artificial Superintelligence: A Futuristic Approachman	2015	60	5.45
11	Typology of Risks of Generative Text-to-Image Models	Bird, Charlotte; Ungless, Eddie; Kasirzadeh, Atoosa	Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society	2023	54	18.00

https://doi.org/10.24191/mij.v6i2.9041

12	When to Make Exceptions: Exploring Language Models as Accounts of Human Moral Judgment	Jin, Zhijing; Levine, Sydney; Gonzalez, Fernando; Kamal, Ojasv; Sap, Maarten; Sachan, Mrinmaya; Mihalcea, Rada; Tenenbaum, Joshua; Schölkopf, Bernhard	Advances in Neural Information Processing Systems	2022	54	13.50
13	Hard Choices in Artificial Intelligence	Dobbe, Roel; Krendl Gilbert, Thomas; Mintz, Yonatan	Artificial Intelligence	2021	52	10.40
14	Ethics And Governance of Artificial Intelligence: Evidence from a Survey of Machine Learning Researchers	Zhang, Baobao; Anderljung, Markus; Kahn, Lauren; Dreksler, Noemi; Horowitz, Michael C.; Dafoe, Allan	Journal of Artificial Intelligence Research	2021	51	10.20
15	Comprehensive Review of Battery State Estimation Strategies using Machine Learning for Battery Management Systems of Aircraft Propulsion Batteries	Raoofi, Tahmineh; Yildiz, Melih	Journal of Energy Storage	2023	50	16.67
16	Moral Uncanny Valley: A Robot's Appearance Moderates How Its Decisions Are Judged	Laakasuo, Michael; Palomäki, Jussi; Köbis, Nils	International Journal of Social Robotics	2021	50	10.00
17	Predicting Future AI Failures from Historic Examples	Yampolskiy, Roman V.	Foresight	2019	50	7.14
18	Artificial Intelligence and Administrative Evil	Young, Matthew M.; Himmelreich, Johannes; Bullock, Justin B.; Kim, Kyoung-Cheol	Perspectives on Public Management and Governance	2019	48	6.86
19	Safety Engineering for Artificial General Intelligence	Yampolskiy, Roman; Fox, Joshua	Topoi	2013	48	3.69
20	Trafficgen: Learning to Generate Diverse and Realistic Traffic Scenarios	Feng, Lan; Li, Quanyi; Peng, Zhenghao; Tan, Shuhan; Zhou, Bolei	Institute of Electrical and Electronics Engineers (IEEE) International Conference on Robotics and Automation	2023	46	15.33

Indicatively, Yampolskiy, Roman V., is featured thrice, working on superintelligence (2015), failures of AI (2019), and AI safety engineering (2013), thereby establishing his expertise in AI risk research. Frequent citation within a given year of current papers (such as 47.50/year in the case of the 2024 survey) reflects the increasing rate of interest in AI issues. The ethical and social issues are the driving theme, and papers such as "Hard choices in AI" (2021) and "Ethics and governance of AI" (2021) address ethical conflicts and regulatory deficits. Applied research has an impact on technical contributions, such as TrafficGen (2023), in the field of autonomous vehicles.

The list has a dual nature, with a double focus that presents both the future AI potentials (such as language models and robotics) and threats (such as adversarial attacks and existential threats). The fact that AAAI/ACM conferences and interdisciplinary journals (Nature, BMJ) are at the forefront emphasizes the collaborative nature of the field. In general, such citations outline the direction of AI studies towards responsible innovation, combining innovations with crucial precautions.

5. CONCLUSIONS

This study aims to examine the trend of research about AI safety using bibliometric analysis. By adopting this bibliometric analysis, it can evaluate the performance of a research area (Gu, 2004), explain aspects that support the involvement of studies in a research area and help researchers in directing their efforts toward making impactful studies. Thus, a bibliometric analysis of Scopus-indexed papers on AI safety was conducted to promote the development of research in AI safety. The study on AI safety was initiated by Rodd (1995) and has been cited by 10 papers. This report highlights an area of study that has undergone

https://doi.org/10.24191/mij.v6i2.9041

significant expansion over the past few years, evolving into a diverse and multidisciplinary field of research. More than 80 percent of publications emerged after 2020, indicating an increased academic and practical interest, which is attributed to the development of generative AI and growing concerns about ethical and existential risks. The findings reveal an enduring technical underpinning, with computer science and engineering dominating, but increasingly supplemented by the social sciences and humanities.

This study also makes several contributions to the field of AI safety. First, it studies publication patterns by analysing document and source types, yearly publications, subject areas, countries, author contributions, institutional contributions, and abstracts. Second, this study recognizes the leading studies and authors by mapping citations. Lastly, this study identifies the knowledge-able structure by recognizing the most knowledgeable structure using citation analyses. The added value of this bibliometric study lies in its in-depth description of the AI safety space, based on the literature indexed by Scopus, and thus provides a data-informed evaluation of growth, crossover with other fields, and topic-related trends. The study provides actionable findings for researchers, policymakers, and funding agencies by identifying peaks in research production, the most active institutions, authors, and countries, and revealing the shifting proportion between technical and ethical research. It identifies interdisciplinary strategies and international cooperation, as well as knowledge gaps and regions that remain underserved. It serves as a baseline guide for future research priorities and the context required in the evidence base for the rapidly growing realm of AI safety. This can help other researchers study the topic further.

Certain limitations and constraints are associated with the present bibliometric analysis. First, the results only occurred for the specific keyword, “AI safety,” based on the title, keyword, and abstract of the documents. At times, the breadth of this topic led to ambiguity regarding the application of inclusion/exclusion criteria during the search phase of the review. Therefore, future research can likely be expanded by filtering and cleaning data before analysis can be conducted. A second limitation arises from our reliance on the Scopus database as the primary source of documents, which narrows the study's scope to specific journals and types of documents, potentially overlooking influential sources. Although Scopus is among the most extensive databases that index all scholarly works (Sweileh et al., 2017), it does not naturally encompass all available sources. Therefore, other available databases can be utilized in future research, such as Web of Science, ScienceDirect, and Google Scholar. By combining these three databases, it might contribute to more interesting and valuable results. Third, the authors' self-citations are included in the analysis. However, authors' self-citations are sometimes appropriate because those may be linked to the continuation of an author’s or research group’s previous work (Liu, 2025). Lastly, individual contributions or ground-breaking, small-volume research may be missed by bibliometrics.

Despite these limitations, this study has contributed to the knowledge and research field by presenting the current trends in research on AI safety. Thus, this study has clarified the future development direction of research on AI safety by systematically and comprehensively understanding the current state of research and its trends. First, the interdisciplinary work and sociotechnical lines of research on AI safety are promising areas for future exploration, encompassing both technical robustness and social governance of AI. Second, it is also necessary to widen the scope of multilingual, non-Western voices and create standardized measures to evaluate them. Lastly, to continue making progress on responsible and inclusive AI development, an investigation of long-term existential risks and short-term safety problems is necessary.

6. ACKNOWLEDGEMENTS

The authors are grateful to the anonymous reviewers for their helpful, thorough, and constructive recommendations. All opinions presented here are the author's own.

https://doi.org/10.24191/mij.v6i2.9041

7. CONFLICT OF INTEREST STATEMENT

The authors declared that there is no conflict of interest in this paper.

8. AUTHORS' CONTRIBUTIONS

All authors contributed equally to this work. All authors read and approved of the final manuscript.

REFERENCES

Bautista-Bernal, I., Quintana-García, C., & Marchante-Lara, M. (2024). Safety culture, safety performance and financial performance. A longitudinal study. *Safety Science*, *172*, 106409. https://doi.org/10.1016/j.ssci.2023.106409

Chadha, P., Gera, R., Sharma, Y., & Dixit, S. (2024). Artificial Intelligence in Cyber Security: A Bibliometric Analysis. *Applications of Artificial Intelligence in Business and Finance 5.0*, 231–255. https://doi.org/10.1201/9781003535133-12

Durán-Sánchez, A., Del Río-Rama, M. de la C., Álvarez-García, J., & García-Vélez, D. F. (2019). Mapping of scientific coverage on education for Entrepreneurship in Higher Education. *Journal of Enterprising Communities: People and Places in the Global Economy*, *13*(1/2), 84–104. https://doi.org/10.1108/JEC-10-2018-0072

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, *69*(1), 131–152. https://doi.org/10.1007/s11192-006-0144-7

Gou, X., Liu, H., Qiang, Y., Lang, Z., Wang, H., Ye, D., Wang, Z., & Wang, H. (2022). In-depth analysis on safety and security research based on system dynamics: A bibliometric mapping approach-based study. *Safety Science*, *147*, 105617. https://doi.org/10.1016/j.ssci.2021.105617

Gu, Y. (2004). Global knowledge management research: A bibliometric analysis. *Scientometrics*, *61*(2), 171–190. https://doi.org/10.1023/B:SCIE.0000041647.01086.F4/METRICS

Haruna, C., Hashem, I. A. T., & Maray, M. (2024). Bibliometric analysis for artificial intelligence in the internet of medical things: mapping and performance analysis. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1347815

Herman, L. (2023). Use of Artificial Intelligence in Public Services: A Bibliometric Analysis and Visualization. *TEM Journal*, *12*(2), 798–807.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102

Kunal, P., & Patkar, P. (2023). *Bibliometric Analysis of global Research Trends on AI and ML Used to Resolve Emerging Vehicle Technology Problems using Scopus Database*.

Liu, Y. (2025). Self-Citation Versus External Citation in Academic Publishing: A Critical Analysis of Citation Reliability, Publication Biases, And Scientific Quality Assessment. *Publication Biases, And Scientific Quality Assessment (August 14, 2025)*.

Luka, I., Aleksandar, Š., Bratislav, P., Dejan, V., & Darjan, K. (2024). Research Trends in Artificial Intelligence and Security—Bibliometric Analysis. *Electronics*, *13*(12), 2288. https://doi.org/10.3390/electronics13122288

https://doi.org/10.24191/mij.v6i2.9041

Niyazov, Y., Vogel, C., Price, R., Lund, B., Judd, D., Akil, A., Mortonson, M., Schwartzman, J., & Shron, M. (2016). Open Access Meets Discoverability: Citations to Articles Posted to Academia.edu. *PLOS ONE*, *11*(2), e0148257. https://doi.org/10.1371/journal.pone.0148257

Norris, M., & Oppenheim, C. (2007). Comparing alternatives to the Web of Science for coverage of the social sciences' literature. *Journal of Informetrics*, *1*(2), 161–169. https://doi.org/10.1016/j.joi.2006.12.001

Rodd, M. G. (1995). Safe AI—is this possible? *Engineering Applications of Artificial Intelligence*, *8*(3), 243–250. https://doi.org/10.1016/0952-1976(95)00010-X

Russell, S. (2022). Human-Compatible Artificial Intelligence. *Human-like machine intelligence*, *1*, 3-22.

Sweileh, W. M., Al-Jabi, S. W., AbuTaha, A. S., Zyoud, S. H., Anayah, F. M. A., & Sawalha, A. F. (2017). Bibliometric analysis of worldwide scientific literature in mobile - health: 2006-2016. *BMC Medical Informatics and Decision Making*, *17*(1), 1–12. https://doi.org/10.1186/S12911-017-0476-7/TABLES/9

Tamascelli, N., Campari, A., Parhizkar, T., & Paltrinieri, N. (2024). Artificial Intelligence for safety and reliability: A descriptive, bibliometric and interpretative review on machine learning. *Journal of Loss Prevention in the Process Industries*, *90*, 105343. https://doi.org/10.1016/j.jlp.2024.105343

Tekin, U., & Dener, M. (2025). A bibliometric analysis of studies on artificial intelligence in neuroscience. *Frontiers in Neurology*, *16*. https://doi.org/10.3389/fneur.2025.1474484

Tsay, M.-Y. (2009). Citation analysis of Ted Nelson's works and his influence on hypertext concept. *Scientometrics*, *79*(3), 451–472. https://doi.org/10.1007/s11192-008-1641-7

https://doi.org/10.24191/mij.v6i2.9041