risks

MDPI

Article

AI Risk Management: A Bibliometric Analysis

Adelaide Emma Bernardelli [1,2] **and Paolo Giudici** [2,*]

[1] School for Advanced Studies, IUSS Pavia, 27100 Pavia, Italy; adelaide.bernardelli@iusspavia.it
[2] Department of Economics and Management, University of Pavia, Via San Felice 5, 27100 Pavia, Italy
* Correspondence: paolo.giudici@unipv.it

Abstract

The growth of Artificial Intelligence applications requires the development of risk management models that can balance opportunities with risks. This paper contributes to the development of Artificial Intelligence risk management models by means of a thorough bibliometric analysis. The analysis highlights the need to develop a quantitative AI risk management framework.

Keywords: AI risk management; sustainability; responsible AI; Safe AI; human-centered AI

1. Introduction

The growth of Artificial Intelligence applications requires the development of risk management models that can balance opportunities with risks. An AI system is a sophisticated machine-based framework designed to perform tasks that involve learning from data and making informed predictions or decisions based on that data. These systems operate with either explicit objectives—clear, predefined goals that guide their processes—or implicit objectives, which may be less apparent but nonetheless drive the system's behavior. The hallmark of an AI system is its capability to analyze the input it receives and utilize that analysis to generate a range of outputs. These outputs may manifest as predictions, content creations, personalized recommendations, or strategic decisions tailored to specific contexts.

One of the defining features of AI systems is their ability to influence both physical and virtual environments. This means that the outcomes produced by AI can have tangible effects on the world, be it in the realm of digital interactions or in the physical domain where human activities take place. This distinguishes Artificial Intelligence from traditional software or standard machine-based systems. While conventional systems may process information and return outputs, AI systems possess a deeper level of sophistication that allows their outputs to actively shape and alter the settings in which they operate. The implications of this capability are profound. The influence exerted by AI can be beneficial, opening doors to new opportunities, enhancing efficiencies, and driving innovation across various sectors. For instance, AI can streamline operations, improve customer experiences, and aid in complex problem-solving, leading to positive transformations in industries such as healthcare, finance, and education.

It is crucial to recognize that this capacity for influence is a double-edged sword. Alongside the potential for positive impact, AI systems can also give rise to negative consequences. These adverse effects may include biases in decision-making processes, breaches of privacy, and unintentional harm resulting from automated actions. As AI continues to evolve and become more integrated into everyday life, it is essential to address these risks and develop risk management frameworks that ensure the responsible use of Artificial Intelligence, balancing its beneficial capabilities against the potential for harm.

check for updates

Academic Editors: Alejandro Balbás and Michel Dacorogna

Received: 9 May 2025
Revised: 10 June 2025
Accepted: 4 July 2025
Published: 7 July 2025

Citation: Bernardelli, Adelaide Emma, and Paolo Giudici. 2025. AI Risk Management: A Bibliometric Analysis. *Risks* 13: 131. https://doi.org/10.3390/risks13070131

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

In the light of the previous discussion, AI risk management can be defined as a framework that can measure and, therefore, manage and mitigate the risk that the application of AI can generate harms for individuals, organizations, or the environment. See, e.g., Giudici et al. (2024) and Novelli et al. (2024).

This paper contributes to the development of the Artificial Intelligence risk management framework by means of a thorough bibliometric analysis that aims to establish and organize the state-of-the-art research knowledge on AI risk management.

2. Bibliometric Analysis

2.1. Introduction and Process

In the context of Artificial Intelligence (AI), "risk" encompasses a broad range of potential negative outcomes associated with the development, deployment, and use of AI systems. These include operational risks (e.g., model failure), ethical risks (e.g., bias and discrimination), security risks (e.g., adversarial attacks), societal risks (e.g., job displacement or misinformation), and long-term existential threats (e.g., from advanced general-purpose AI). Given the pervasive nature of AI across domains and its potential to cause harm at individual, organizational, and systemic levels, AI risk has become a key concern for risk professionals. Our analysis contributes to this discussion by systematically mapping the research landscape and identifying the key themes, actors, and methodological approaches shaping the field of AI risk management.

Starting from the Scopus database, we conducted a bibliometric review using the software R, Version 4.4.2, Aria and Cuccurullo (2017). Scopus is a comprehensive and widely used platform that indexes peer-reviewed literature across a broad range of disciplines. The platform includes journals, conference proceedings, and books, making it particularly well-suited for capturing the interdisciplinary nature of AI risk research, which spans computer science, law, ethics, policy, and business. It also provides structured metadata such as keywords, author affiliations, and citation counts, enabling robust and replicable bibliometric studies.

Compared to the Web of Science, Scopus offers broader indexing of conference proceedings and technical publications, which are especially relevant in the fast-evolving field of AI. Although no database is exhaustive, Scopus' curated and wide-ranging coverage ensures the inclusion of key contributions to the field and supports systematic keyword analysis, co-authorship mapping, and thematic clustering.

We selected documents by searching for the term "*AI risk*" or the term "*Artificial Intelligence risk*" in both article titles and keywords, using the "OR" operator to maximize correspondences. The search was restricted to specific subject areas: Computer Science; Engineering; Social Sciences; Business, Management and Accounting; Mathematics; Decision Sciences; and Economics, Econometrics, and Finance. No time range was set for the search, but documents were restricted to those written in English. This choice reflects the dominance of English as the lingua franca of scientific communication, particularly in high-impact and policy-relevant fields. However, it may introduce linguistic bias by underrepresenting valuable research conducted and published in other languages. Notably, the presence of non-Anglophone authors (e.g., from China, Germany, or Italy) in our dataset partially mitigates this limitation, as many scholars choose to publish internationally in English. Still, future studies may benefit from the integration of multilingual sources or the use of databases with more inclusive language indexing.

AI risk is inherently multidisciplinary, as reflected in recent policy frameworks such as the OECD AI Principles, the EU AI Act, and the NIST AI Risk Management Framework, all of which advocate a horizontal, cross-sectoral approach. These documents emphasize that AI risks are not confined to technical failures but extend to ethical, legal, societal,

and governance challenges. As such, our study includes documents from a wide range of subject areas. While this breadth introduces conceptual heterogeneity, it is necessary to capture the full spectrum of how "AI risk" is being framed and addressed. We acknowledge this diversity in our interpretation and caution against overgeneralization across disciplinary boundaries.

We selected the 2015–2025 period to capture the emergence and evolution of AI risk as a topic of scholarly interest. The year 2015 marks the beginning of significant regulatory and academic engagement with AI governance, coinciding with early proposals for ethical AI guidelines and the inception of discussions surrounding the EU AI Act. While the early years in our time frame contain fewer publications, they are essential for capturing foundational contributions and shifts in thematic focus over time.

Overall, the bibliometric analysis was conducted following a systematic and replicable sequence of steps:

1. **Database selection**: We selected Scopus as our primary data source due to its comprehensive and curated coverage of interdisciplinary scientific literature.
2. **Search strategy**: We searched for documents containing the keywords "*AI risk*" or "*Artificial Intelligence risk*" in article titles or author keywords, using an "OR" Boolean operator to broaden the match.
3. **Inclusion criteria**: We limited the corpus to English-language documents and to relevant subject areas (Computer Science, Engineering, Business, Law, Economics, etc.).
4. **Data extraction**: Metadata, including author names, affiliations, keywords, publication sources, and citation counts, were exported from Scopus.
5. **Data processing**: The dataset was imported into `R` and processed using the `Bibliometrix` package. Pre-processing included keyword normalization and synonym merging where appropriate.
6. **Analysis**: We applied several analytical modules—keyword co-occurrence analysis, thematic mapping, country collaboration networks, and source analysis—to identify major clusters and trends.
7. **Interpretation**: Finally, the results were interpreted based on the outputs of clustering algorithms and visualized in strategic thematic maps to guide discussion of emerging themes and gaps in the literature.

This multi-step protocol ensures transparency, reproducibility, and methodological rigor, providing a framework that can be applied or adapted by future researchers in related bibliometric studies.

2.2. Documents and Sources

The final dataset comprises 129 documents, with an annual growth rate of 26.51%. The average document age is 1.99 years, and each document receives an average of 6.14 citations. The dataset includes contributions from 103 different sources (journals, books, etc.) and contains a total of 432 Author's Keywords. The collection consists mainly of conference papers (53) and journal articles (46), along with book chapters (12), reviews (7), books (5), and other minor sources (6).

The dataset reflects contributions from 354 authors, with 31 unique authors having contributed single-authored works. On average, each document has approximately 2.94 co-authors, and the percentage of international co-authorship stands at 27.13%. This metric provides insight into the degree of internationalization in research collaboration: a higher percentage typically indicates broader cross-country cooperation and knowledge exchange. In this case, the research appears to be moderately internationalized, suggesting a fair level of collaboration between institutions from different countries, although domestic or national collaborations still represent a significant portion of the overall scholarly production.

The number of publications has varied over time, with a significant increase observed in 2023. The dataset shows a steady rise from 2015 to 2022, with a jump from 9 documents in 2022 to 27 in 2023. A visual representation of this trend is provided in Figure 1.

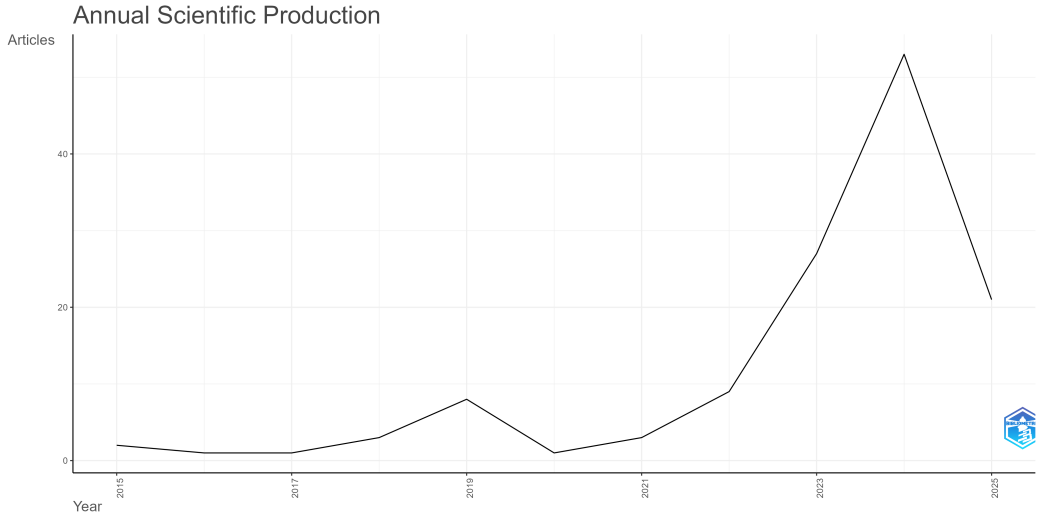


Figure 1. Annual scientific production of the documents that contain the term "AI risk" or "Artificial Intelligence risk" either in the title or in the keywords (Source: Scopus Database). Production peaked in 2024.

Figure 1 shows the highest productivity peak in 2024, with 53 publications. This trend highlights a growing academic interest in AI risk, particularly in recent years, suggesting that the topic has gained substantial relevance within the research community. Although the dataset includes papers from 2025, we acknowledge that the current year is incomplete, and the publication count for 2025 should not be interpreted as conclusive. Nonetheless, we chose to include it in order to gain a more comprehensive understanding of the thematic evolution surrounding the topic of "AI risk".

Examining the most relevant sources within the dataset, we observe a dominance of conference proceedings and specialized journals reflecting the interdisciplinary and evolving nature of research on AI risk. The ACM International Conference Proceeding Series emerges as the most prolific source (six documents), followed by the CEUR Workshop Proceedings (five publications), indicating the central role of technical conferences in presenting early findings and fostering discussion around algorithmic risks and responsible AI design. Among peer-reviewed journals, *AI and Society*, *Communications in Computer and Information Science*, and the *Computer Law and Security Review* (each with three publications) stand out as influential platforms. These journals reflect the multidisciplinary engagement with AI risk, ranging from social implications and regulatory concerns to technical communication. The presence of both legal and technical venues among the top sources further underscores the inherently cross-sectoral nature of discourse surrounding AI risk governance.

In order to better understand the topic of *AI risk* and the surrounding themes, we begin by examining the most globally cited documents, as shown in Figure 2.

Figure 2 highlights the 10 papers that, to date, appear most cited. Note that the journal venue in which they are published is mainly from computer science, as expected, but also from law and societal applications, emphasizing the interdisciplinary nature of AI risk management. The literature on AI risk presents a multifaceted landscape of perspectives, each contributing to a deeper understanding of the challenges and opportunities AI presents to organizations, policymakers, and society at large.

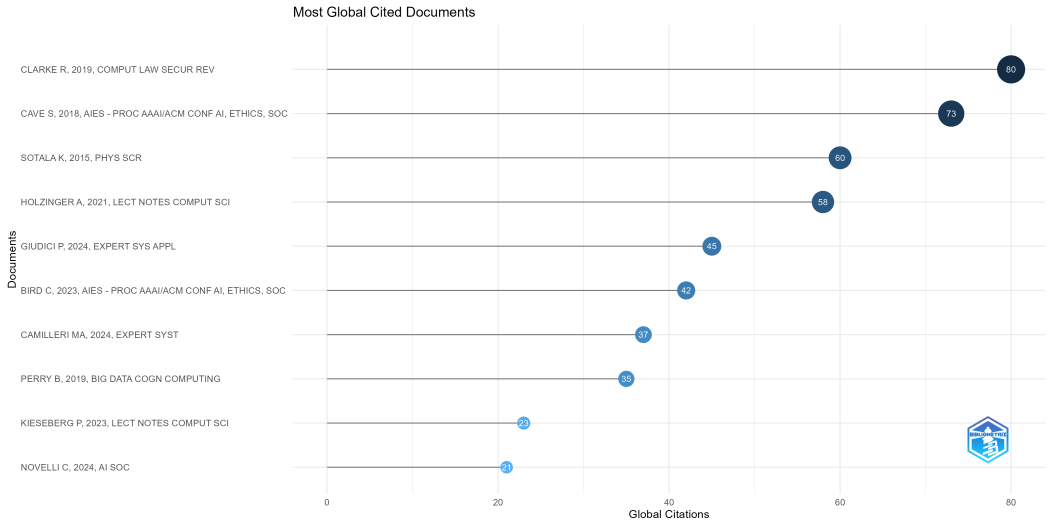


Figure 2. Most cited documents.

Clarke (2019) offers a foundational analysis of the integration of responsible AI practices within organizational frameworks. His work underscores the dual nature of AI, which, while offering significant benefits, simultaneously introduces complex risks. The author emphasizes the need for a comprehensive approach to managing these risks, including stakeholder engagement, organizational culture, and the adaptation of traditional risk assessment techniques. He proposes 50 Principles for Responsible AI, drawn from diverse sources, as a guide for organizations to align their practices with ethical considerations and stakeholder expectations. A key contribution of Clarke's work is his call for a multi-stakeholder approach to risk management, stressing the importance of considering the perspectives of users, affected communities, and policymakers to ensure AI technologies serve the broader societal good.

Cave and ÓhÉigeartaigh (2023) further build on this discourse by examining the risks associated with framing AI development as a competitive "race". Their paper highlights how such a narrative can incentivize shortcuts in safety and governance, potentially exacerbating geopolitical tensions and leading to conflicts. The authors categorize AI risks into three types: (i) those stemming from the race rhetoric, (ii) those associated with the actual race, and (iii) the dangers following an AI race victory, such as power concentration. They advocate for a shift from a race mentality to a collaborative approach, promoting international cooperation and responsible AI development.

The catastrophic risks associated with Artificial General Intelligence (AGI) are discussed in depth by Sotala and Yampolskiy (2014). This paper addresses the global-scale threats AGI could pose, emphasizing the importance of proactive risk mitigation strategies. Sotala categorizes potential responses to AGI risk into societal proposals, external constraints, and internal design modifications, including the development of Oracle AIs. The paper advocates for differential intellectual progress, ensuring that advancements in AI safety outpace AI capabilities, a concept that aligns with ethical considerations and proactive risk management in AI development.

In a more practical perspective, Giudici et al. (2024) introduce an integrated risk management model specifically tailored for AI applications. This paper distinguishes itself by focusing on AI-specific risks, such as "model" risks, and draws parallels to financial regulation, particularly the Basel frameworks. The authors propose a set of Key AI Risk Indicators (KAIRIs) that measure critical AI principles—Sustainability, Accuracy, Fairness, and Explainability. This model provides a structured approach for organizations, particularly in the financial sector, to manage AI risks in compliance with emerging regulatory standards.

Camilleri (2024) brings another crucial view to the literature by discussing AI governance, emphasizing the social and ethical responsibilities of all stakeholders involved in AI development. The paper stresses that the growing recognition of AI's potential risks calls for governance frameworks that balance the benefits of AI innovation with the necessity of risk mitigation. The key concerns highlighted include privacy and security, with a particular focus on preventing the spread of misinformation and bias. Camilleri's work aligns with Clarke's and Cave and ÓhÉigeartaigh's by underscoring the importance of ethical governance in managing AI risks, while also addressing the specific challenges posed by AI in relation to social responsibility.

Perry and Uuk (2019) shift the focus to AI governance from a policy perspective. The paper advocates for the establishment of a subfield dedicated to AI governance, emphasizing the importance of understanding the policymaking process itself in mitigating AI risks. By outlining the stages of policy formulation and identifying challenges such as a lack of institutional expertise, the authors highlight the complexities involved in regulating AI.

Kieseberg et al. (2023) introduce the concept of Controllable AI as an alternative to traditional notions of Trustworthy AI. This framework aims to manage complex AI systems without relying heavily on transparency or explainability, which are often difficult to achieve in advanced AI systems. The paper also examines security concerns related to AI, particularly in the context of adversarial threats and harmful decision-making by AI systems. Kieseberg's work complements the broader discourse by proposing new strategies for ensuring AI systems remain under control and ethically sound, adding another layer to the conversation on AI risk management.

Finally, Novelli et al. (2024) discuss the limitations of the EU Artificial Intelligence Act (AIA) in categorizing AI risks, particularly for general-purpose AI (GPAI), which can lead to ineffective enforcement and misestimation of risk magnitude. The paper proposes a new risk assessment model that integrates the AIA with frameworks from the Intergovernmental Panel on Climate Change (IPCC), emphasizing the need for scenario-based evaluations that consider multiple interacting risk factors. The authors argue that this approach will enhance regulatory measures, promote sustainable AI deployment, and better protect fundamental rights by providing a more nuanced understanding of AI risks, particularly in the context of large language models (LLMs).

In summary, the literature on AI risk spans a range of topics, from responsible organizational practices and geopolitical dynamics to the potentially catastrophic risks of AGI and the development of regulatory frameworks. These contributions collectively highlight the need for a proactive, multi-faceted approach to AI governance that incorporates ethical, security, and regulatory considerations to mitigate the risks associated with AI technologies.

2.3. Authors

The temporal distribution of publications by the most prolific authors is illustrated in Figure 3. It highlights patterns of sustained and emerging contributions to the field. In particular, Holzinger A. and Kieseberg P. stand out as key figures, with publications spanning multiple years and a relatively high citation impact (as indicated by the darker circles and larger size representing total citations per year). In contrast, other authors—including Bertaina S., Biganzoli I., and De Silva D.—appear to be more recent contributors, with publications concentrated in 2023 and 2024, indicating a growing scholarly interest in the topic. Overall, the data suggest that while the field is attracting new contributors, a core group of authors is consistently shaping the dialogue through ongoing research output.

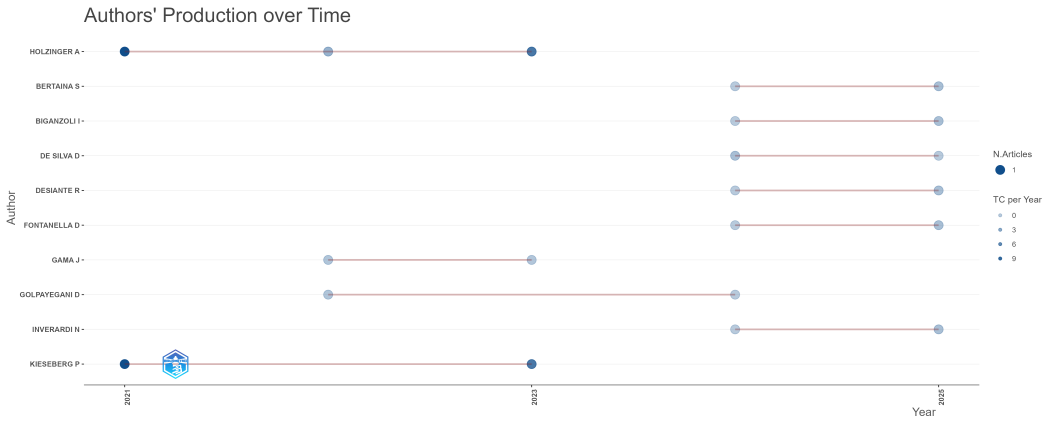


Figure 3. Top 10 most relevant authors' production over time.

Figure 4 shows the distribution of research contributions by country of origin. It reveals a strong international engagement in the literature related to AI risks. The United States leads in the total number of publications, primarily through Single Country Publications (SCPs), indicating a high volume of domestically conducted research. In contrast, countries such as China, Italy, Germany, and India also show a significant research output, but with a more balanced mix of SCPs and Multiple Country Publications (MCPs), suggesting a stronger orientation toward international collaboration. Countries like France, Canada, and Israel demonstrate a higher proportion of MCPs relative to their overall output, while countries like Belgium and Portugal appear exclusively in MCPs. This distribution highlights both the dominant national research hubs and the critical role of international cooperation in shaping the discourse on emerging risks and technological development.

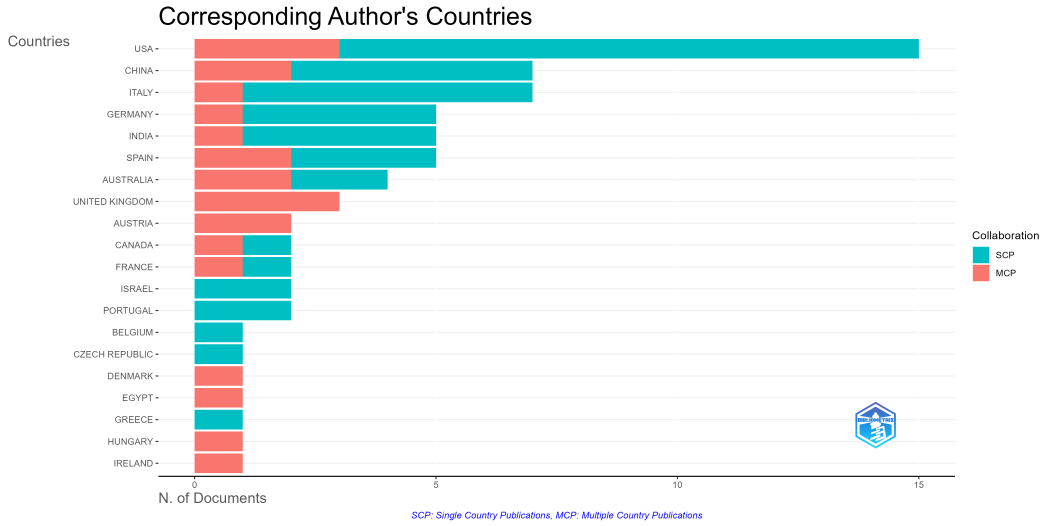


Figure 4. Distribution of corresponding authors by country, distinguishing between Single Country Publications (SCPs) and Multiple Country Publications (MCPs), highlighting the USA as the leading contributor and the prevalence of international collaboration in countries like China, Italy, and Germany.

To offer an alternative perspective, Figure 5 depicts the global collaboration network among the countries of the authors. In this network, nodes represent individual countries, while edges denote co-authored publications involving researchers from the connected countries. The size of each node is proportional to the total number of publications produced by authors from that country, and the thickness of each edge reflects the frequency of collaborations between the respective countries. Consistent with expectations, the United States and the United Kingdom appear as prominent central nodes, indicating their leading

roles in AI risk research. Additionally, several other clusters emerge, forming significant hubs of research activity characterized by both domestic and international collaborative efforts. This visualization effectively illustrates the geographic distribution and interconnected nature of the global AI risk research community.

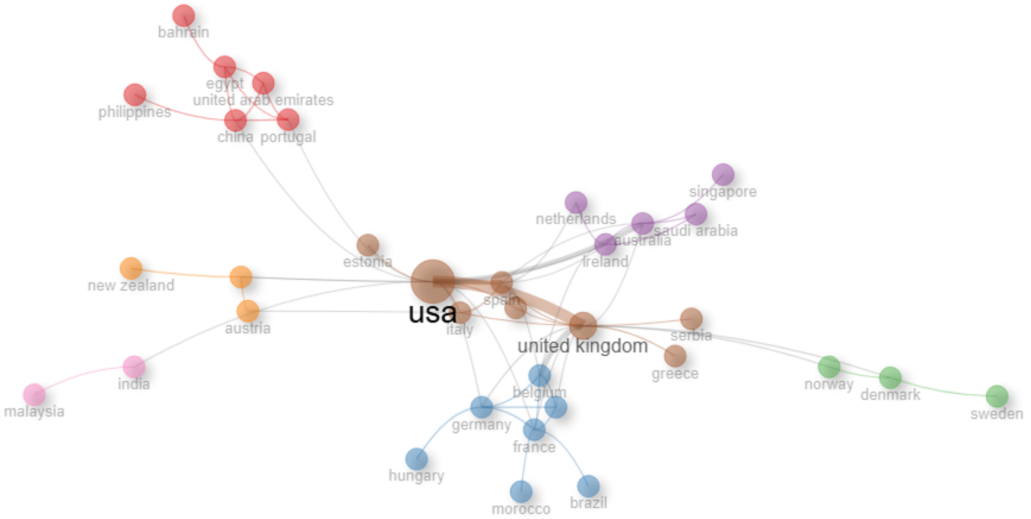


Figure 5. Country collaboration map.

2.4. Keywords

In order to gain a deeper understanding of thematic trends and their evolution over time, a detailed analysis of the authors' keywords was conducted. This approach offers valuable insight into the recurring concepts that shape the research agenda on AI risk and reveals how these focal points have shifted in recent years.

Figure 6 presents a word cloud that visually summarizes the most frequently occurring terms among authors' keywords. The visualization includes the top 50 terms by occurrence, after excluding a few general terms that were already used in the dataset construction (e.g., *AI*, *Artificial Intelligence*, *AI risk*, *AI risks*), as well as merging synonyms to group conceptually similar keywords (e.g., *AI Act*, *EU AI Act*, *Artificial Intelligence Act*; *generative AI*, *gen-AI*). However, we retained semantically distinct terms such as "*risk management*" and "*AI risk management*" due to their differing scope and frequency of use, also taking into account the fact that we encompassed many areas of study for the analysis of the term "*AI risk*". This choice was deliberate, as these terms, although conceptually adjacent, often reflect different levels of abstraction: "risk management" is a general concept spanning multiple domains, while "AI risk management" is a domain-specific application that frequently co-occurs with unique regulatory or technical keywords. A threshold of three co-occurrences was set to determine inclusion in the network, balancing granularity and interpretability. Multi-word terms were preserved as provided by the authors, and no stemming or lemmatization was applied.



Figure 6. Word cloud created starting from authors' keywords included in the documents. Prominent terms highlight ethical concerns (e.g., *AI ethics*, *trustworthy AI*), regulatory focus (*AI Act*, *AI regulation*), technical aspects (*explainable AI*, *large language models*), and both short- and long-term risks (*cybersecurity*, *existential risk*).

Figure 6 shows that, at the core of the thematic landscape are terms such as *AI ethics*, *trustworthy AI*, *responsible AI*, and *AI governance*, which underscore a widespread concern with the ethical and normative dimensions of Artificial Intelligence. Closely linked to these, we find some policy-oriented keywords like *AI regulation* and *AI Act*, which reflect an increasing scholarly engagement with the legal and institutional frameworks needed to ensure responsible AI development. Complementing these themes, keywords such as *AI risk management* and *risk assessment* indicate a strong focus on operational strategies to evaluate and mitigate potential harms posed by AI systems. The notable presence of *generative AI* reflects an emerging interest in the implications of AI-generated content, particularly in light of recent technological advances. From a technical standpoint, keywords like *explainable AI* and *large language models* highlight an ongoing effort to enhance AI transparency, interpretability, and accountability, features that are considered critical for fostering trust in AI technologies. Finally, terms such as *cybersecurity*, *risk perception*, and *existential risk* reveal a broader view of AI-related dangers, spanning both immediate safety concerns and more speculative, long-term threats with potentially catastrophic consequences.

A co-occurrence network of authors' keywords was created, and is presented in Figure 7. The same list of terms was removed, and the same synonym grouping was applied as in the previous wordcloud to ensure consistency in the analysis of the 50 most frequent nodes created on the basis of the Louvain clustering algorithm[1]. Node size here represents keyword frequency and edge thickness indicating co-occurrence strength. The relatively uniform node sizes reflect the low frequency counts across a small corpus (129 documents), which limits the visual contrast.

Figure 7 visualizes the relationships between keywords that frequently co-occur in the same publications. It shows that the keywords with stronger co-occurrence links are connected by thicker edges, and those that are conceptually related tend to form clusters. Larger nodes represent keywords that appear more frequently in the dataset. This analysis reveals several critical insights into the thematic structure of discussions on AI risk. We see that *AI ethics* emerges as the central node, which is closely linked to multiple clusters, suggesting its foundational role in the broader discourse. The blue cluster includes terms such as *AI governance*, *responsible AI*, *generative AI*, and *AI risk management*, indicating

a thematic focus on institutional frameworks and emerging challenges associated with generative technologies. The red cluster centers around *AI Act*, *trustworthy AI*, and *risk management*, reflecting a regulatory and normative perspective that overlaps with legal frameworks and risk-related governance. The brown cluster groups together keywords like *explainable AI*, *security*, *robustness*, and *safety*, which emphasize technical aspects and the need for transparent and resilient AI systems. A green cluster containing *machine learning*, *data analytics*, and *risk mitigation* reflects a more data-driven and application-oriented stream of research. Smaller, more isolated clusters are also present: the purple cluster links *existential risk* with *artificial general intelligence*, highlighting long-term, high-impact concerns, while the orange cluster featuring *impact assessment* and *risk assessment* suggests a niche but significant methodological focus. Again, we see that the network underscores the multidisciplinary and multi-level nature of AI risk research, spanning ethical, technical, legal, and strategic dimensions.

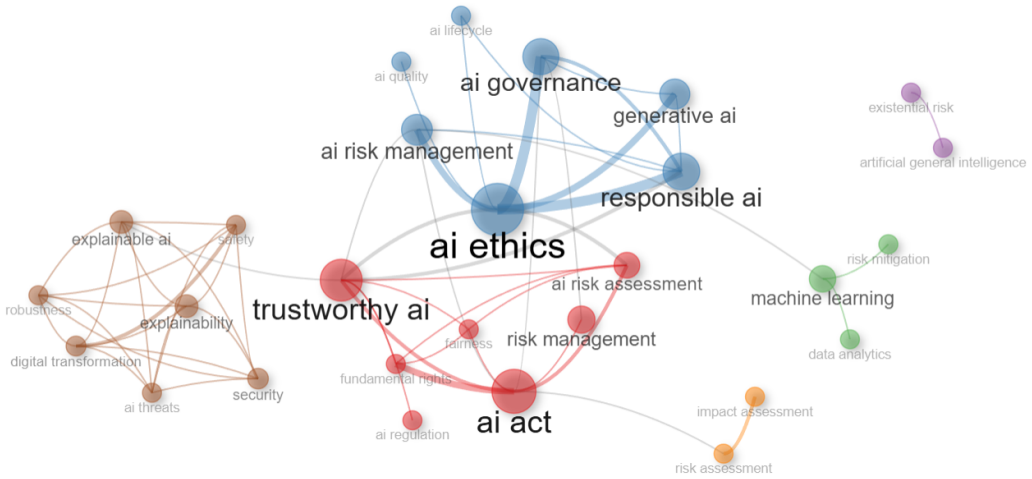


Figure 7. Word co-occurrence network. A list of synonyms was used to exclude similar words, like 'ai' and 'artificial intelligence', 'ai risk' and 'ai risks'.

The previous analysis enables us to outline three broader themes that look like a step-by-step guide to approaching AI risks:

- *Foundational Principles: Ethics and Governance*. These actions are directed towards establishing ethical guidelines, governance frameworks, and regulatory policies to ensure responsible AI development and deployment. Key aspects include AI ethics, trustworthy AI, AI governance, and compliance with regulatory frameworks (e.g., EU AI Act).
- *Risk Identification and Security Measures*. After the first step, it is critical to recognize and address AI-related risks, including large-scale threats, safety concerns, and security vulnerabilities. This refers to AI safety, risk perception, existential risks, and AI threats in cybersecurity and digital transformation.
- *Evaluation and Mitigation Strategies*. Once having understood the potential risks of AI applications, it is fundamental to create structured methodologies to systematically assess and mitigate AI risks. This involves impact assessment and risk assessment techniques to measure and manage AI-related uncertainties.

2.5. Thematic Map

To further explore the conceptual structure of the field, a thematic map was generated based on authors' keywords, and is shown in Figure 8.

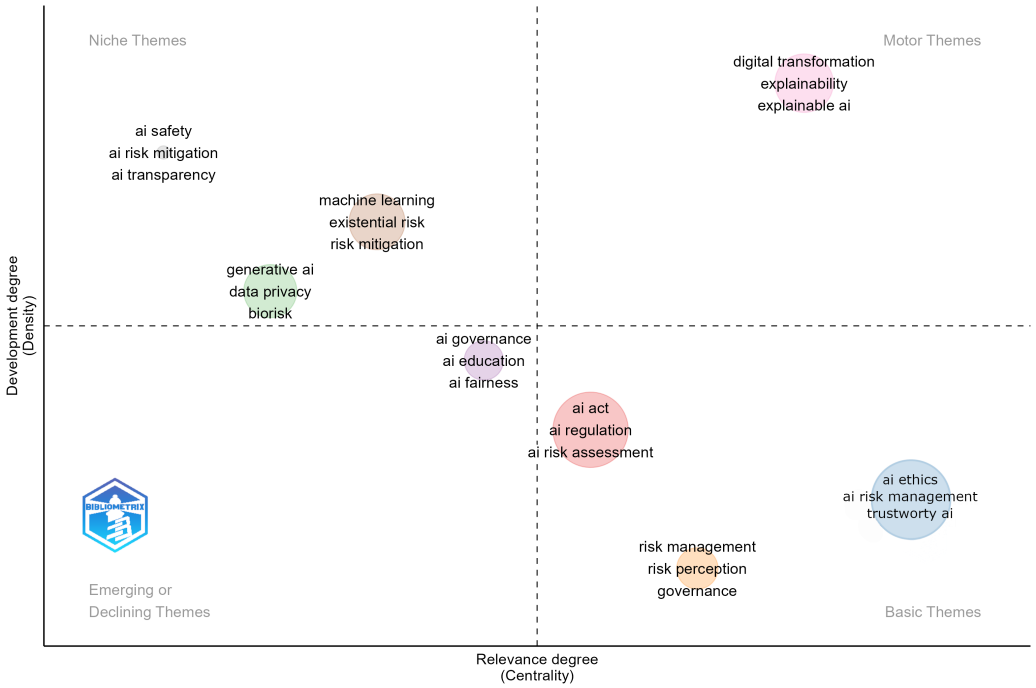


Figure 8. Thematic map of authors' keywords using Louvain clustering. Themes are positioned by centrality (importance) and density (development), identifying motor, niche, emerging/declining, and basic/transversal themes.

This map provides a strategic overview of the conceptual structure of AI risk research by organizing keyword clusters based on their interconnections and development stages. The analysis was conducted on the 250 most frequent author keywords extracted from the Scopus dataset, after applying a two-step pre-processing procedure coherently with the analysis presented in the section about keywords: (i) the removal of the general search terms used to construct the corpus; (ii) synonymous or closely related terms were grouped manually to ensure semantic consistency across variations, keeping related but distinct terms as separate entities when they appeared with different co-occurrence patterns or thematic associations.

To ensure thematic significance and exclude sparse or overly idiosyncratic topics, we applied a minimum cluster frequency threshold of 10 per 1000 documents. Given our corpus size (129 documents), this translates to a minimum of roughly 1–2 occurrences, which helps retain diversity while focusing on recurring concepts. The keyword co-occurrence network was constructed by creating a symmetrical matrix, where each cell M_{ij} represents the number of documents in which both keywords i and j appear together. From this network, we applied the Louvain community detection algorithm, a modularity-based method that clusters keywords into coherent thematic groups by maximizing intra-cluster connections and minimizing inter-cluster ones. The resulting thematic clusters were then plotted on a two-dimensional Cartesian plane based on two computed metrics: centrality and density Cobo et al. (2011). *Centrality* measures the degree of interaction a given cluster has with other clusters. It is computed as the sum of the normalized co-occurrence strengths among all keywords in one cluster and all keywords in the remaining clusters. High centrality values indicate that the cluster functions as a thematic hub, bridging multiple research areas, and suggesting that the concepts it contains are widely referenced across the literature. *Density*, in contrast, measures the internal cohesion of a cluster. It is calculated as the average co-occurrence strength among all pairs of keywords within the same cluster, normalized by the number of possible keyword connections. A high density indicates that the cluster is internally well-integrated and thematically mature.

Based on these two axes, the thematic map is divided into four quadrants, each conveying a distinct strategic positioning.

Figure 8 leads to the following findings. The upper right quadrant (*Motor themes*) includes themes characterized by both high centrality and high density, indicating that they are well-developed, conceptually robust, and play a pivotal role in structuring the field. They exhibit strong internal coherence while also serving as hubs that connect with other thematic areas. In our map, a prominent cluster located in this quadrant revolves around *digital transformation* and *explainability*. This suggests that discussions around the interpretability of AI systems and their transformative impact on organizational and societal levels are both mature and central to the discourse on AI risk. Themes in the upper-left quadrant (*Niche themes*) exhibit high density but low centrality. They are internally well-developed and conceptually coherent but remain marginal to the overarching structure of the field. Such themes often reflect specialized or highly technical areas of inquiry. In our analysis, three distinct clusters occupy this space: the first pertains to *AI safety*, the second to *machine learning* and *existential risk*, and the third to specific risk typologies such as *biorisk* and *data privacy*. These findings suggest that while these areas are thematically rich, they remain relatively isolated within the broader network of research on AI risk. The quadrant in the lower-left corner (*Emerging or declining themes*) is characterized by low density and low centrality, typically associated with themes that are either in the early stages of development or in decline. They are weakly connected to other areas and lack internal cohesion. In our map, themes related to *AI education* and *AI fairness* are located in this quadrant. Given the growing policy and societal interest in these topics, it is plausible that these represent emerging rather than declining themes. Moreover, the cluster is positioned close to the origin of the coordinate plane, probably indicating a transitional status that may evolve as these topics gain traction in the literature. Themes in the lower-right quadrant (*Basic and transversal themes*) are characterized by high centrality and low density. They are conceptually important for the field due to their high level of connectivity, yet they are not internally well-developed. These themes are typically foundational or transversal, supporting broader debates without constituting standalone areas of inquiry. In our case, three main clusters were identified in this quadrant. The first pertains to *AI risk assessment* and *AI regulation*, the second to *AI risk management* and *AI ethics*, and the third to *risk perception*. These themes, due to their high centrality, appear to function as conceptual anchors within the field and could benefit from further refinement and empirical exploration, especially in relation to their application across more specific or emerging subtopics.

This thematic configuration reveals several important implications. First, the coexistence of robust technical themes (e.g., explainability, digital transformation) with specialized but more isolated areas (e.g., AI safety, existential risk) highlights the fragmented nature of the field. Despite shared concerns, disciplinary boundaries and conceptual inconsistencies hinder the emergence of a unified risk management approach. Second, the marginal position of fairness and education within the current literature—despite their visibility in policy—signals a gap between societal concerns and academic treatment. Addressing this disconnect will be essential for inclusive and equitable AI governance. The presence of foundational but loosely structured clusters (e.g., AI ethics, regulation, risk perception) suggests a need for deeper theoretical integration and methodological refinement. Researchers should focus on developing standardized frameworks and indicators—such as Key AI Risk Indicators (KAIRIs)—that can bridge normative principles with measurable risk outcomes. For policymakers, the map offers guidance on where regulatory attention is already well-supported by academic knowledge (e.g., explainability) and where it remains conceptually underpowered (e.g., risk perception or fairness metrics). For practitioners, it

highlights which themes are most actionable, and which require caution due to their lack of methodological maturity.

Ultimately, this analysis illustrates that AI risk management is still in a formative phase. However, it is also a field rich with opportunity, ready for interdisciplinary collaboration, methodological innovation, and strategic policy alignment. By clarifying the conceptual landscape, this study provides a roadmap for researchers and stakeholders seeking to develop robust, responsible, and human-centered approaches to AI governance.

2.6. Interpretation of the Bibliometric Findings

The bibliometric analysis conducted insofar reveals that, to balance opportunities with risks, a quantitative—albeit practical—approach to AI risk management is necessary. Our findings suggest that while ethical and regulatory discussions around AI risk have gained significant traction, there is a pressing need for frameworks that not only articulate principles but also provide measurable and actionable risk indicators. Such an approach could be grounded in the four main principles commonly found in established risk management frameworks: Security, Accuracy, Fairness, and Explainability.

Security refers to the requirement for AI systems to achieve an adequate level of robustness and cybersecurity. In this context, the failure to ensure security is closely linked to well-known operational and cyber risks in traditional risk management.

Accuracy is equally critical: AI systems must produce reliable and valid outputs, with the risk of inaccuracies corresponding to the established notion of model risk. These two dimensions—security and accuracy—reflect the technical foundation of AI risk, and their prominence in the literature underscores the continuing influence of computer science and engineering perspectives.

Fairness, however, introduces a more socially-oriented dimension. It emphasizes the importance of data representativeness and non-discrimination in AI outcomes, particularly for groups or individuals affected by automated decisions. This type of risk differs from conventional organizational risk, in that the loss or harm is often externalized onto users or society at large. Our analysis indicates that while fairness appears as a recurrent keyword, its operationalization remains inconsistent across disciplines, calling for more structured, cross-sectoral methodologies for assessment and mitigation.

Explainability completes the framework, highlighting the need for AI systems to be understandable and auditable by humans. This dimension aligns with governance risk, reinforcing the importance of transparency and oversight in complex algorithmic systems. Although terms like “explainable AI” and “trustworthy AI” feature prominently in our keyword analysis, their clustering suggests a fragmentation of the debate, with little convergence on shared definitions or evaluation metrics.

Taken together, these four dimensions constitute a baseline for evaluating AI systems in a way that is both rigorous and adaptable to different sectors. A valid risk management framework should, therefore, be capable of quantifying the risk of an AI system failing along any of these axes. By managing and mitigating such risks, AI systems can evolve toward being not only technically sound but also ethically responsible and human-centered.

Beyond identifying major clusters and trends, our analysis also reveals several tensions and knowledge gaps in the current literature. A key insight is the coexistence of highly normative debates—such as those focused on ethics, values, and governance—with more recent but still underdeveloped technical concerns, particularly around robustness, explainability, and scalability. There is a notable misalignment between the rapid pace of AI innovation and the comparatively slower development of empirical tools and standards for risk measurement. Moreover, the increasing visibility of keywords like “generative AI”, “existential risk”, and “catastrophic AI” suggests a growing academic interest in long-term

and global-scale implications. However, these are not yet consistently integrated into the core methodological literature, which continues to prioritize regulatory compliance and system-level auditing.

Foundational concepts such as "risk perception", "AI impact assessment", and "trustworthiness", although present, appear with low frequency or in peripheral thematic clusters. This indicates that some critical areas remain underdeveloped or clustered within specialized communities. These findings reinforce the necessity of interdisciplinary collaboration, both to fill empirical gaps and to harmonize conceptual frameworks across domains.

Policymakers and practitioners can use this bibliometric mapping to identify underexplored but strategically important themes, prioritize research funding toward areas of convergence (such as explainability and regulation), and foster the development of integrated risk assessment frameworks. Ultimately, bridging these conceptual and methodological divides is essential for building AI systems that are not only efficient and innovative, but also safe, equitable, and aligned with societal values.

3. Conclusions

This study presents the first bibliometric analysis explicitly focused on the emerging field of AI risk management, offering a structured and replicable overview of its evolving research landscape. Our findings show that interest in AI risk has grown significantly in recent years, yet the field remains conceptually fragmented, with limited integration across technical, ethical, legal, and policy dimensions. Through co-occurrence analysis, thematic mapping, and collaboration network visualizations, we identified key topics—such as fairness, explainability, cybersecurity, governance, and regulation—that form the backbone of current debates. However, these foundational themes often remain unevenly developed and disconnected, indicating the need for more cohesive theoretical frameworks and applied methodologies.

From a practical standpoint, the literature highlights the urgency of adopting a proactive, multi-dimensional approach to AI governance. Addressing AI risk requires the combined efforts of computer scientists, ethicists, legal scholars, and policymakers to develop systems that are not only efficient and innovative, but also safe, fair, and accountable. In this context, our work emphasizes the importance of advancing quantitative tools for risk assessment, such as the proposed Key AI Risk Indicators (KAIRIs), which can help standardize how risks are measured and managed across sectors.

Nevertheless, some limitations should be considered. First, the reliance on the Scopus database, while methodologically sound, may exclude relevant studies indexed in other platforms. Second, restricting the search to English-language publications introduces a potential linguistic bias, despite the global reach of the topic. Third, the relatively small sample size (129 documents) reflects both the niche status of "AI risk" as a bibliographic keyword and the infancy of the field. Fourth, subjective decisions in keyword standardization and clustering may have shaped the structure of the thematic networks. Finally, the interpretability of co-word and thematic analyses depends heavily on the quality and consistency of author-assigned metadata, as well as the clustering algorithm chosen for the bibliometric analysis.

Despite these limitations, this paper provides a robust foundation for future research and policy efforts. It maps the contours of an emerging field, identifies underexplored areas, and offers a roadmap for further academic inquiry and regulatory design. By doing so, it contributes to the development of a more responsible, sustainable, and risk-aware ecosystem for Artificial Intelligence.

The key takeaway for this paper is that research on AI risk management is very relevant and necessary, particularly in terms of measurement models. This paper summarizes the available research efforts and contributes to identifying which research activities are in place and which require more attention.

Author Contributions: Conceptualization, A.E.B. and P.G.; methodology, P.G.; software, A.E.B.; validation, P.G.; formal analysis, A.E.B.; investigation, A.E.B.; resources, P.G.; data curation, A.E.B.; writing—original draft preparation, A.E.B.; writing—review and editing, P.G.; visualization, A.E.B.; supervision, P.G.; project administration, P.G.; funding acquisition, P.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data is available upon request to the authors.

Acknowledgments: The work of A.E.B. has been conducted during and with the support of the Italian national inter-university PhD course in Sustainable Development and Climate change: www.phd-sdc.it, accessed 6 July 2025, Cycle XL, with the support of a scholarship co-financed under Ministerial Decree no. 630 of 28 March 2024, (Investment 3.3—Innovative PhDs), based on the National Recovery and Resilience Plan (NRRP)—funded by the European Union—NextGenerationEU—Mission 4 "Education and Research", Component 2 "From Research to Business", and by Noto Sondaggi Srl.

Conflicts of Interest: The authors declare no conflicts of interest.

Note

[1] This algorithm is widely used for community detection in large networks. It is a modularity-based method that optimizes the modularity score—a measure of the density of links inside communities compared to links between communities. The Louvain method operates in two phases: (i) nodes are assigned to communities in a way that locally maximizes modularity, and (ii) the resulting communities are aggregated into a new network where each community becomes a node, and the process is repeated. The result is an efficient partition of the network into thematically coherent clusters Blondel et al. (2008).

References

Aria, Massimo and Corrado Cuccurullo. 2017. bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics* 11: 959–75. [CrossRef]

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008: P10008. [CrossRef]

Camilleri, Mark Anthony. 2024. Artificial intelligence governance: Ethical considerations and implications for social responsibility. *Expert Systems* 41: e13406. [CrossRef]

Cave, Stephen, and Seán S. ÓhÉigeartaigh. 2023. Limitations of the EU Artificial Intelligence Act in Categorizing AI Risks: A New Risk Assessment Model for General-Purpose AI. Paper presented at 2023 AI Risk Assessment Conference, Chengdu, China, November 24–26. pp. 1–5.

Clarke, Roger. 2019. Principles and business processes for responsible AI. *Computer Law & Security Review* 35: 410–22.

Cobo López-Herrera, Ernesto Herrera-Viedma, and Francisco Herrera. 2011. An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. *Journal of Informetrics* 5: 146–66. [CrossRef]

Giudici, Paolo, Mattia Centurelli, and Stefano Turchetta. 2024. Artificial Intelligence risk measurement. *Expert Systems with Applications* 235: 121220. [CrossRef]

Kieseberg, Peter, Edgar Weippl, A. Min Tjoa, Federico Cabitza, Andrea Campagner, and Andreas Holzinger. 2023. Controllable AI—An alternative to trustworthiness in complex AI systems? Paper presented at International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Benevento, Italy, August 29–September 1. pp. 1–12.

Novelli, Claudio, Federico Casolari, Antonino Rotolo, Mariarosaria Taddeo, and Luciano Floridi. 2024. Taking AI risks seriously: A new assessment model for the AI Act. *AI & Society* 39: 2493–97.

Perry, Brandon, and Risto Uuk. 2019. AI governance and the policymaking process: Key considerations for reducing AI risk. *Big Data and Cognitive Computing* 3: 26. [CrossRef]

Sotala, Kaj, and Roman V. Yampolskiy. 2014. Responses to catastrophic AGI risk: A survey. *Physica Scripta* 90: 018001. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.