Analyzing the Ecosystem Contexts in the AI Literature Using Latent Dirichlet Allocation and Exploratory Factor Analysis

KRESHNIK VUKATANA[1], ELIRA HOXHA[1], ARBEN ASLLANI[2]
[1]Department of Statistics and Applied Informatics
University of Tirana
Nënë Tereza Square, 4
ALBANIA
[2]Department of Data Analytics
University of Tennessee at Chattanooga
601 McCallie Avenue
USA

Abstract: This study aims to explore the major topics in the recent artificial intelligence (AI) ecosystem literature and identify and categorize those topics into categories of AI ecosystems. The study analyzed 149 publications from Google Scholar using two text mining techniques: latent Dirichlet allocation (LDA) and exploratory factor analysis (EFA). The LDA identified 12 major topics, while the EFA grouped them into six common factors: (a) human resources-driven, (b) technology and algorithm-based, (c) business and entrepreneurial-driven, (d) legal, ethical, privacy, and regulatory framework, (e) innovation-based, and (f) government-supported. The goal is to suggest various AI ecosystems and their best fit for a country or region based on its characteristics and resources. Understanding these types of AI ecosystems can provide valuable insights for government agencies, policymakers, businesses, educational institutions, and other stakeholders to align strategies with resources for developing successful AI-driven ecosystems.

Key-Words: AI ecosystem, LDA, EFA, AI Literature.

Received: June 22, 2024. Revised: January 9, 2025. Accepted: April 3, 2025. Published: May 19, 2025.

1 Introduction

Artificial intelligence, the capability of a computational system to imitate intelligent behavior [1], has the potential to revolutionize many aspects of everyday citizens and generate new business models for creating new products and services. Traditionally, the development of AI is focused on specific applications without analyzing the system of actors involved, [2], [3]. As such, most advancements in AI are driven by a small group of firms with vast computational and data resources, [4]. However, prior research also emphasizes the importance of analyzing AI as part of an entire ecosystem of interconnected firms and institutions, [5]. Creating an effective ecosystem for the development and implementation of AI technologies could be the best approach to promote regional growth, address unemployment challenges, and ensure that the benefits of AI are distributed fairly and equitably across regions and their citizens, [6]. Government agencies and policymakers are expediting their efforts to understand the fundamental nature and components of such systems.

Previous literature has identified two major aspects of an ecosystem in general: innovation ecosystem and entrepreneurial ecosystem. An innovation ecosystem is defined as IT-enabled collaborative arrangements through which firms combine their individual offerings into coherent, customer-facing solutions, [7]. On the other hand, an entrepreneurial ecosystem is a market-driven collaboration between the private sector and social actors to create wealth in a symbiotic relationship, [8]. The two sides of the ecosystem contribute to each other's growth. The entrepreneurial ecosystem allows for the creation of a strong network effect, which can facilitate the sharing of information and resources and attract new firms and talent to the cluster. This, in turn, leads to further growth, innovation, and development of new products and services, [9].

One can infer the key actors of an AI ecosystem by applying the principles of ecosystem analysis more broadly. An AI ecosystem has both sides of the ecosystem in general as well: one side represents co-creation, knowledge sharing, and open innovation aspects of the AI ecosystem, and the other side represents a territorial approach to the AI ecosystem, a regional cluster composed of universities and research institutes, tacit knowledge, social capital,

agglomeration spillovers, and anchoring, [10]. However, while much research has been devoted to understanding the technological impact of AI, the role of actors within the AI ecosystem is still under-explored, and there is a need for continued research into the business models and power dynamics that govern the AI ecosystem, as well as the strategies employed by smaller firms and startups to compete in this rapidly evolving landscape, [5].

This paper uses two text-mining techniques to synthesize existing literature on AI ecosystems, providing a comprehensive overview of the various types of AI ecosystems and their distinguishing characteristics. Text mining is previously used to extract patterns from collections of text documents [11], or used in theme analysis to identify patterns and quantify emerging keywords, which in turn can provide insights into the structure, context, and trends of social media or other text documents, [12], [13]. Also, the studies [14], [15], [16], demonstrate how to extract themes from published articles in selected fields using exploratory factor analysis on the frequency of keywords in those articles. Most recently, a text-mining approach to introduce and define the umbrella concept of the "AI entrepreneurial ecosystem" has been utilized, [17]. Their method relies on identifying the most frequent keywords within the "AI ecosystem" corpus.

This paper enhances the prior methodology by combining LDA and EFA. As explained later, the LDA approach enhances the accuracy of the analysis since it relies on the most salient keywords instead of the most frequent ones. As such, the LDA makes it easier to analyze large corpora and extract relevant insights without being overwhelmed by noise from less significant but frequent words, [18]. The paper answers the following research questions:

- *What are the major topics discussed in the AI ecosystems literature, and what is the frequency distribution of the most salient keywords for each topic?*

- *How do the most likely topics correlate with each other in terms of latent factors, and how do these latent factors inform the structure of AI ecosystems?*

The following section provides a brief description on the advantages of combining LDA and EFA for text mining in general and for exploring AI ecosystems in particular. Following that, the paper provides a detailed description of the methodology and presents the results and analysis in the subsequent two sections. Lastly, the paper concludes with final remarks and suggestions for future research.

2 Materials and Methods

2.1 LDA, EFA, and the Advantages of Combining them

The LDA methodology assumes that any given document in a corpus is a mixture of latent topics, and each topic is a statistical distribution of words. LDA can be used to identify the most likely topics in a document and to identify the most salient words and their distribution in each topic. The model assigns a prior distribution of topics and words and, through an unsupervised learning process, identifies patterns of word co-occurrence to reveal latent structures. LDA has been applied in tasks ranging from document classification to recommendation systems, [18], [19], [20].

EFA is a statistical method used to reduce the dimensionality of a dataset by identifying the most important principal variables (principal components analysis) or latent factors that account for most of the variance in a dataset. EFA is mostly applied to structured data sets; however, it has been widely used in text-mining tasks such as document clustering, feature selection, and information retrieval, [21]. One of the early studies on using EFA for text mining was conducted by [22], who proposed a method for document classification using EFA and singular value decomposition. They demonstrated the effectiveness of the proposed method in a large collection of Reuters news articles and showed that it outperformed traditional methods such as Naive Bayes and k-nearest neighbors. EFA has been used in text classification for feature selection, on a movie reviews dataset, [23]. The study compared EFA, mutual information, and information gain as selection methods and EFA had the highest accuracy in terms of classification.

Related to information retrieval tasks in a research papers dataset, EFA improved the system performance related to the precision and recall features, [24]. In a web search engine EFA was used for query expansion and improved its performance regarding the mean average precision, [25]. Anyway, EFA has difficulties in finding complex relationships when working with high-dimensional text data, therefore this paper proposes using also the LDA method to first reduce the dimensionality and assist EFA in the process.

LDA and EFA are similar in their aim in finding and extracting important features from a collection of documents, but not in the way they do so. LDA finds the main topics of the documents whereas EFA identifies the main patterns or features in the whole dataset. Using both of these methods helps improving the accuracy and effectiveness of the analyses on the extracted information from large

datasets of text documents. In previous studies [26], EFA has been used before LDA to reduce the dimensionality of the document-term matrix. In a dataset of patent abstracts the performance of LDA was improved related to coherence and topic quality when applying EFA first. Other studies [27], [28], show that the combination of LDA and EFA helps in document classification, topic modelling, and information retrieval, as part of text-mining tasks. The methodology used in this paper follows the steps below:

(1) *select the journal papers, where text mining will be applied*

(2) *prepare the created corpus for analysis*

(3) *apply LDA to extract the main topics and the most important keywords*

(4) *apply EFA to further reduce the dimensionality of the data and identify the latent factors*

The Python's utilities genism [29], [30], and pyLDA, and its visualization libraries [31], [32], are used to implement the LDA. While the IBM's SPSS package is used for the factor analysis. The detailed description of the four steps is given in the following sections.

2.2 Select journal articles for text mining

Journal articles published during the 2012-2022 timeline on artificial intelligence ecosystems were downloaded from Google Scholar using a query that combines search words "artificial intelligence" OR "AI" with the words "ecosystem" OR "cluster" OR "hub" OR "region". Google Scholar's search engine allows for the sorting of results by relevance, and the most relevant 50 articles were selected for the study. A preliminary scanning of the article keywords and abstracts was conducted to filter out those articles that clearly do not discuss any topics related to the AI ecosystems. A total of 149 articles were used for the LDA analysis. As shown in Table 1 (Appendix), 67 articles (45%) are peer-reviewed papers in journals published by publishers such as Elsevier, Springer, IEEE, Sage, and MDPI, with an average of 123 citations per paper. There are 42 papers published in other peer-reviewed journals, 12 non-peer-reviewed journals, five working papers, five conference publications, and three dissertations. The study also included 15 nonacademic publications, such as technical reports or posts.

Table 2 (Appendix) shows the average citations per publication and the number of publications each year. Almost 95% of publications in a 5-year period (2018-2022) and there are an average of 161 citations, according to the "cite" tool from the Google Scholar website.

2.3 Prepare corpus for analysis

The 149 pdf files are further processed to prepare the corpus for the LDA analysis. These steps include removing stop words, punctuation, numbers, or other characters. In addition, lemmatization techniques are applied, and bigrams and trigrams are identified. Finally, the final list of words with the highest level of originality was filtered and included in the LDA analysis. Below is a more detailed description of the data-cleaning steps:

2.3.1 Remove stop words, punctuations, numbers, and other characters

Python's Natural Language Toolkit (nlkt) contains a total of 179 stop words [33], such as 'these', 'of', 'at', 'by', 'for', 'with', and 'about'. The stop word list was extended to include words usually associated with journal articles or website documents but do not carry any meaning to the AI ecosystem discussion. For example, worlds like 'et al.', 'doi:', 'http', '©', 'journal', 'page', and 'figure' were removed. Further, string.punctuation function was used to remove all the punctuation.

2.3.2 Lemmatization

Lemmatization is a technique used in NLP that normalizes the text by replacing any word with its base root mode, [34]. Lemmatization allows for keeping the root or similar meaning of the word, which is derived from the inflection of the words. For example, the words 'employee', 'employed', 'employee', 'employees', 'employer', 'employers', and 'employment' are represented as 'employ', or words like 'innovation', 'innovative', 'innovativeness', 'innovating', 'innovator', and 'innovators' are represented by 'innovat'. The library WorldNetLemmatizer from the nltk module is used for lemmatization.

2.3.3 Identify bigrams and trigrams

Another step in preparing the corpus for analysis is identifying in the text pairs or trios of words that are often associated with each other. During the modelling process these associations are treated together as one single word by using bigram or trigram functions, [35]. Examples of bigrams are 'artificial intelligence' or 'information technology', whereas examples of trigrams are 'artificial intelligence application' or 'AI entrepreneurial ecosystem'.

2.3.4 Creating the final list of words

As the last step of the data-cleaning process of the corpus, the Term Frequency, and Inverse Document Frequency (TF-IDF) method will be used to indicate the originality of each word. The TF is the ratio

of word appearances in each document with the total number of words in the same document, while TF-IDF is the multiplication of TF with IDF. IDF, a measure of how rare a word is across all documents in the dataset, can be calculated as the logarithm of the total number of documents divided by the number of documents that contain the word. Since TF is always a number between 0 and 1, and since the number of documents divided by the number of documents that contain the word can be very large, the log function is applied to balance the impact of the second term. TF-IDF increases when the word has a high frequency in each document and when the word appears in fewer documents. Therefore, words with higher TF-IDF scores are more informative and discriminative, and those with lower TF-IDF scores are less exploratory and uninformative words, [18]. The final list of words, also known as the bag of words (BoW), used in this study, contains only words with TF-IDF greater than 0.03.

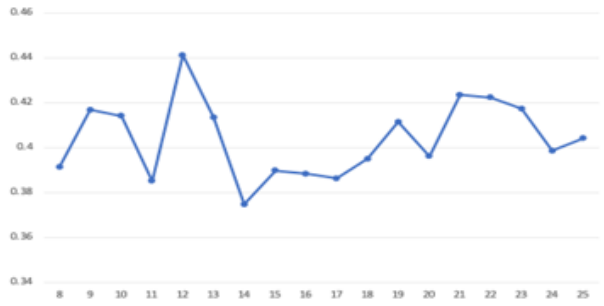

Figure 1: Coherence Level for Various Number of Topics.

2.4 Apply LDA to extract main keywords from AI Ecosystem topics

As mentioned earlier, gensim libraries are used to implement the LDA algorithm, visualize the output, and extract the topics from the documents. A series of LDA models, with various numbers of topics, were executed to find the optimal number of topics that have the highest average coherence score, as shown in Figure 1.

The coherence score measures the quality of a topic. It calculates the semantic similarity between words in a topic, [36]. For a topic t that is characterized by a set of word frequencies W_t={w_1, w_2, ...}, the coherence level is calculated as:

$$c(t, W_t) = \sum_{w_1, w_2 \in W_t}^{\infty} log\left(\frac{d(w_1, w_2) + \varepsilon}{d(w_1)}\right)$$

Where:

$d\ (w_1)$ = the number of documents that contain the word w_1

$d\ (w_1, w_2)$ = the number of documents that contain words w_1 AND w_2

ε = smoothing constant set to 1 to avoid taking the algorithm of zero.

2.5 Apply EFA to identify patterns and factors

As shown in Figure 1, the LDA model with 12 topics achieves the highest average coherence score. The LDA approach also identifies the most prominent keywords in those topics. At this stage, the EFA is used to further reduce the dimensionality of the model by creating keyword clusters through linear transformations. Initially, for this stage, we selected 12 topics and keywords with the TF-IDF greater than 0.03 for each topic, resulting in a total of 70 keywords. The initial run of the model, with 70 features, resulted in a KMO value of 0.53. The Kaiser-Meyer-Olkin (KMO) measure is used to evaluate the sampling adequacy. The KMO indicates the proportion of variance among variables that might be exploratory variance. In [37], is suggested that KMO values closer to 1.0 are considered ideal while values less than 0.5 are considered unacceptable. An iterative and manual process was conducted to improve sampling adequacy, and 34 keywords, as shown in Table 3 (Appendix), were selected. This was an iterative process, and removing the keywords also resulted in removing four topics from our initial list of 12 topics, as well as the articles where these topics were prominent. As such, the EFA analysis continued with 130 articles. Using a MapReduce Python code, the frequency of the 34 keywords in the corpus of 130 articles is calculated. The new model with the frequency of the 34 remaining keywords generated an acceptable level of KMO = 0.670 and passed Bartlett's test of sphericity (significant <0.001) (Table 4, Appendix).

3 Findings and Analysis

As explained earlier, there are two sets of results: the latent topics generated by the LDA, the keyword distributions in those topics, and the latent factors generated by the EFA and their suggested major AI ecosystems.

3.1 Results of the LDA modeling

Figure 2 shows the distribution of the original 12 topics in the 2-dimensional space. The bubble sizes do not vary significantly, which indicates that the topics are represented almost with the same frequency among the documents. The figure 2 also indicates

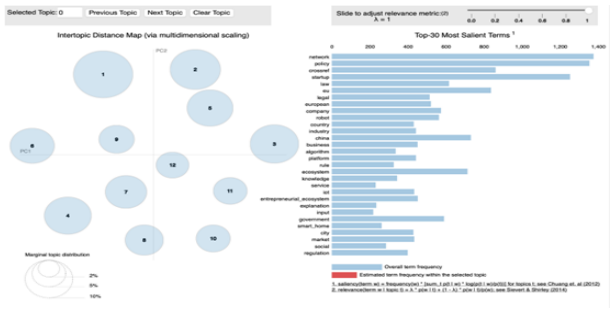

Figure 2: Inter-topic Distance Mapping Among Major Topics.

that the 12 topics are distanced from each other and do not share many exploratory words. The horizontal bar graph on the right shows the articles' top 30 salient (most prominent) words. The saliency measure ranks the more discriminative terms higher, enabling faster assessment and comparison of topics, [38]. As mentioned earlier, we kept eight topics as described below.

3.1.1 Topic 1-Startup and Entrepreneurship

Topic 1 (Figure 3) is the most exploratory in most articles. Based on the keywords like "startup", "company", "platform", "market," and "ecosystem", this topic very likely covers the entrepreneurial aspects of AI. This topic illustrates the relevance of entrepreneurship in the AI discourse and can demonstrate a strong connection between AI and entrepreneurship. Entrepreneurs could start businesses that focus on developing and selling AI products or services, such as software tools for data analysis or machine learning as a service, [39]. Also, an entrepreneur might use AI to analyze market trends and identify untapped niches for new products or services, [40], [41], [42].

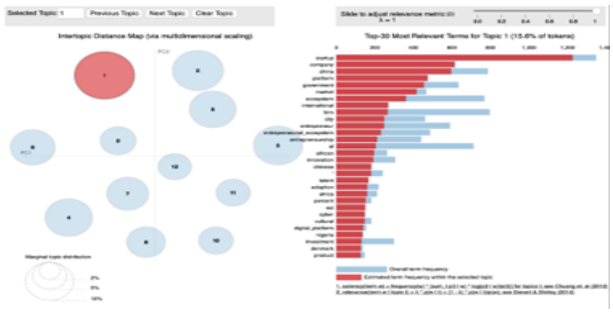

Figure 3: Topic 1-Startup and Entrepreneurship.
Top Keywords: 0.038*"startup"; 0.019*"company"; 0.018*"China"; 0.015*"platform"; 0.014*"government"; 0.013*"market"; 0.011*"ecosystem"; 0.008*"international"; 0.008*"firm".

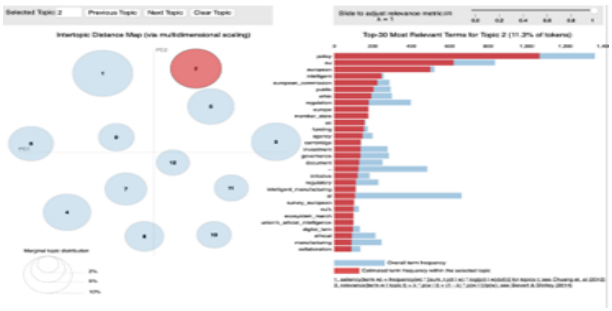

Figure 4: Topic 2 - European Policy and Regulations.
Top Keywords: 0.049*"policy"; 0.029*"eu"; 0.023*"european"; 0.011*"intelligent"; 0.010*"european_commission"; 0.009*"public"; 0.009*"ethic"; 0.008*"regulation"; 0.008*"member_state".

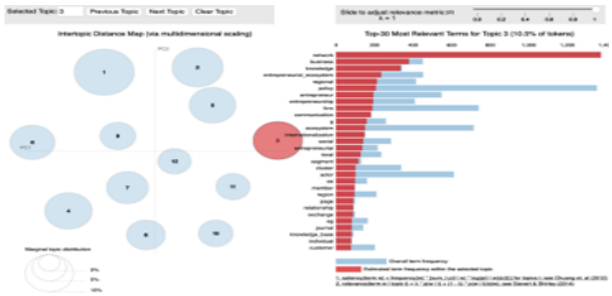

Figure 5: Topic 3- Entrepreneurial Ecosystem.
Top Keywords: 0.069*"network"; 0.019*"business"; 0.017*"knowledge"; 0.012*"entrepreneurial_ecosystem; 0.011*"regional"; 0.010*"policy"; 0.010*"firm"; 0.010*"entrepreneurship"; 0.009*"communication".

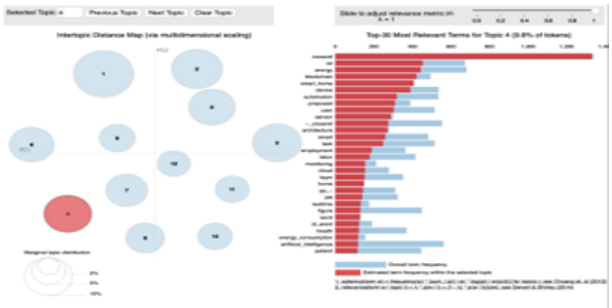

Figure 6: Topic 4-Technology Development.
Top Keywords: 0.016*"iot"; 0.015*"energy"; 0.014*"smart_home"; 0.014*"blockchain"; 0.013*"device"; 0.011*"automation"; 0.010*"user"; 0.010*"sensor"; 0.009*"task".

3.1.2 Topic 2-European Policy and Regulations

The second topic (Figure 4) is the most exploratory topic in most articles. Based on the keywords like ”policy”, ”European”, ”EU”, and ”regulations,” this topic is named ”European Policy and Regulations”. The dominance of this topic in the AI ecosystem discourse illustrates the role of the European Union (EU) in regulating the emergence of AI applications in the region. The EU regulations are intended to ensure that AI is developed and used in a way that is ethical, transparent, and fair and does not pose a risk to individuals' safety, security, or fundamental rights. One key regulation is the EU's ”Ethics Guidelines for Trustworthy AI,” which were released in 2019 by the European Commission's High-Level Expert Group on Artificial Intelligence (AIHLEG), [43].

3.1.3 Topic 3-Entrepreneurial Ecosystem

Topic 3 (Figure 5) is also an entrepreneurial topic with a slight focus on networking and regional policies. The “Entrepreneurial Ecosystem” topic emphasizes the entrepreneurial aspect of AI. Being the third listed topic in the discourse, the result emphasizes the significant role that the entrepreneurial-based AI ecosystem plays in the growth of AI in a specific region.

3.1.4 Topic 4-Technology Development

Topic 4 (Figure 6) is named “Technology Development” due to keywords like “IoT”, “smart home”, “blockchain”, and “sensor”. The AI ecosystem discourse recognizes that several of these technologies have contributed to the recent growth of AI. At the same time, AI leads to the development of several technologies that can think and act like humans. At the same time, AI leads to the development of several technologies that can think and act like humans such as machine learning (ML), NLP, and robotics. These technologies help in several dimensions, such as data analysis, decision-making, and problem-solving [44], by improving the overall performance of the computer system.

3.1.5 Topic 5-Industry and Regional Development

Topic 5 (Figure 7) is named ”Industry and Regional Development” and includes keywords such as ”industry”, ”country”, ”region”, ”actor”, etc., representing the regional AI clusters. AI helps the regional development and may create new business or job opportunities in data science, machine learning, etc., , [45], [46]. Anyway, it is important to ensure the ethical and transparent use of AI in order to build trust.

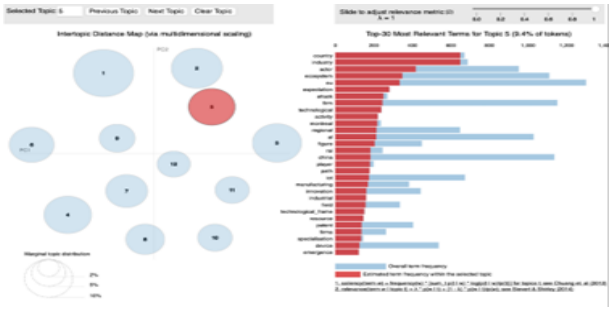

Figure 7: Topic 5-Industry and Regional Development.
Top Keywords: 0.023*”industry”; 0.023*”country”; 0.015*”actor”; 0.012*”eu”; 0.012*”ecosystem”; 0.009*”technological”; 0.009*”firm”; 0.009*”attack”; 0.008*”regional”.

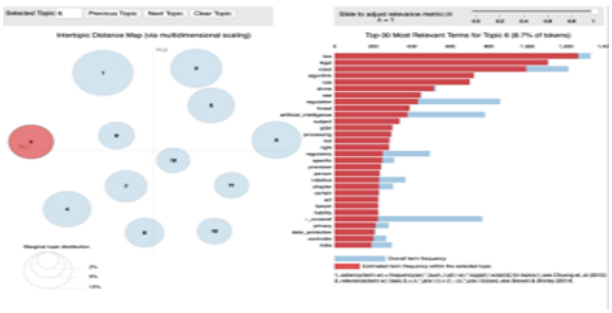

Figure 8: Topic 6-Laws and Regulations.
Top Keywords: 0.035*”law”; 0.031*”legal”; 0.028*”robot”; 0.020*”algorithm”; 0.019*”rule”; 0.014*”drone”; 0.012*”regulation”; 0.011*”forest”.

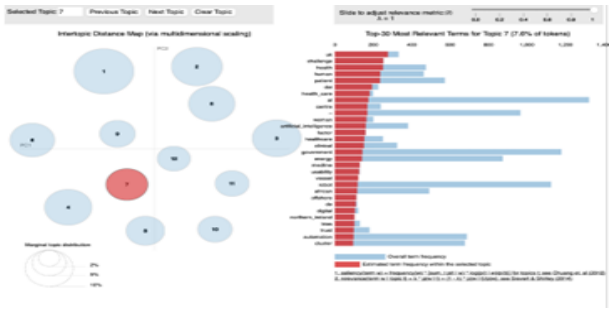

Figure 9: Topic 7-Healthcare.
Top Keywords: 0.010*”uk”; 0.009*”health”; 0.009*”challenge”; 0.008*”patient”; 0.008*”human”; 0.006*”woman”; 0.006*”health_care”; 0.006*”factor”; 0.006*”centre”.

3.1.6 Topic 6-Laws and Regulations

Topic 6 (Figure 8) named ”Laws and Regulations” shows that with the widespread and use of AI in various sectors, there is also an imminent need to develop standard laws and regulations. They should ensure the ethical, transparent, and fair use and development of AI so that the safety, security, and fundamental rights of individuals are not at risk, [47].

3.1.7 Topic 7-Healthcare

The AI articles have also focused their research on the applications of AI in healthcare. This is illustrated by the 7th topic in the LDA model. This topic 7 (Figure 9) is represented by keywords like ”health”, ”patient”, ”human”, ”healthcare” and so on. AI has the potential to revolutionize healthcare by improving the accuracy and efficiency of diagnoses, identifying patterns and trends in patient data, and automating routine tasks. AI technologies such as machine learning and NLP can be used to analyze medical records, imaging studies, and other data to identify patterns and predict outcomes, [48], [49], [50]. AI can also be used to develop new drugs and treatments by analyzing large amounts of research data and identifying potential avenues for further investigation [51].

Topic 7 (Figure 9) represented by keywords like ”health”, ”patient”, ”human”, and ”healthcare” focuses on applications of AI in healthcare, used to improve the accuracy of diagnoses, identify patterns in patient data, or automate routine tasks. ML and NLP are used to analyse medical records, including images, in order to identify patterns or predict outcomes, [48], [49], [50]. Another use of AI is in the direction of developing new drugs and treatments by analyzing large amounts of research data, [51].

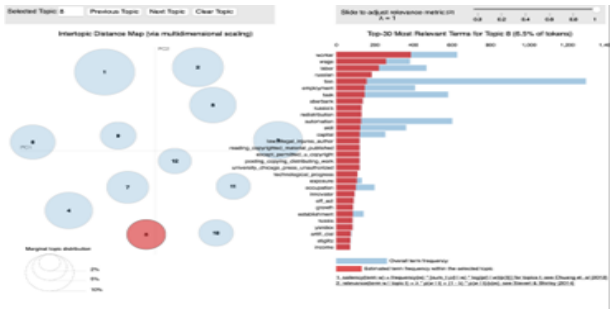

Figure 10: Topic 8-Human Resources.
Top Keywords: 0.018*”worker”; 0.012*”wage”; 0.010*”labor”; 0.008*”russian”; 0.007*”task”; 0.007*”firm”; 0.007*”employment”; 0.006*”skill”; 0.006*”redistribution”

3.1.8 Topic 8-Human Resources

Topic 8 (Figure 10) includes keywords that describe the human factor of AI such as “worker”, “wage”, “labor”, “task”, “employment”, and “skill”. The relationship between AI and human factors can be explored in different directions. AI technologies can be used to improve the efficiency and effectiveness of human tasks, such as by automating routine tasks or providing decision support, [52], [53], [54]. However, the development and use of AI can also raise concerns related to job displacement and the potential for AI to make biased or unfair decisions, [55].

3.2 Results of the EFA modeling

The frequency of the most significant keywords from the eight topics mentioned above is used as input for the EFA, and the principal axis factoring method is used to extract the factors. The scree plot (Figure 11) indicated that the 34 variables could be reduced to five latent factors. Since the factors cannot be assumed to be orthogonal, the Promax is selected as a rotation method. The pattern matrix (Table 5, Appendix) shows the five factors with different loadings of the keywords. Keywords with a loading below 0.3 are removed. Additionally, if a keyword loads on multiple factors, the one with the lowest loading is discarded, ensuring that each keyword is associated with only one factor. Several keywords loaded into the second factor have a negative sign. A negative sign of loading indicates that some variables are related in the opposite direction from the factor, [56]. As such, the keywords with negative loadings are grouped into separate factors. Therefore, six factors in the discourse of AI ecosystems are suggested and explained below.

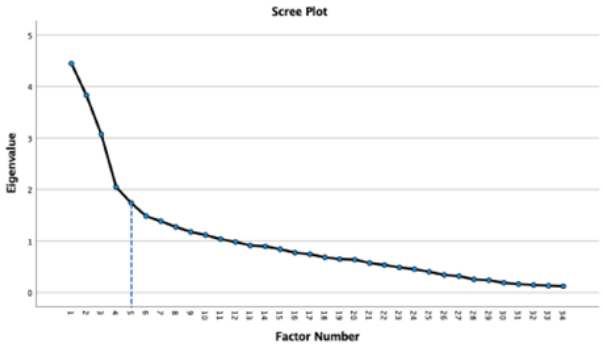

Figure 11: Scree plot indicating the number of factors to be extracted.

3.2.1 Factor 1: Human Resources Driven Ecosystems

The keywords most associated with the first factor in Table 5 (Appendix) relate to concerns about human resources. These associations highlight the critical importance of skilled human capital, the

impact of AI on employment levels and wages due to automation, and the growing demand for new skills in the AI-driven economy. Universities play a significant role in this type of AI ecosystem by developing talent pipelines through AI-specific education and training programs, ensuring that the workforce is prepared for the demands of the evolving AI ecosystem, [57]. Higher education institutions may align their curricula with the skills needed in AI research and development, as well as AI application across different sectors, [58], [59]. By supporting AI talent development, universities directly contribute to strengthening the human capital aspect of AI ecosystems, ensuring that businesses and governments have access to the expertise required to navigate the complex landscape of AI, [60].

3.2.2 Factor 2: Technology and Algorithms-Based Ecosystems

The second factor revolves around key technical components such as AI algorithms, which include neural networks, deep learning, algorithms, and machine learning (Table 5, Appendix). These keywords represent various AI systems and play a significant role in the successful deployment of AI systems. Efficient AI ecosystems require not just advanced algorithms but also stable computational resources and frameworks for handling large-scale data and training complex models, [61]. Integrating specialized hardware, such as GPUs and TPUs, further accelerates deep learning and neural network computations, presenting the importance of technical resources in AI development, [62]. A technology-driven ecosystem requires that universities and industry partners engage in cutting-edge AI techniques while preparing a workforce that tackles the technical demands of AI projects, [63], [64].

3.2.3 Factor 3: Business and Entrepreneurship-Driven Ecosystems

The third factor focuses on AI business and entrepreneurial aspects, captured by keywords such as regional, industry, entrepreneurship, firm, company, and knowledge. These keywords highlight the critical role of businesses in shaping AI ecosystems, alongside the necessity of focusing such efforts on a specific region, [65]. The U.S. stands out in this regard, where the private sector—led by Big Tech companies like Google, Amazon, and Microsoft—drives both AI research and practical applications. Anyway, the AI development is associated with high costs which often are not affordable for small companies. This issue is emphasized by factor 3, [5].

3.2.4 Factor 4: Legal, Regulatory, Ethical, and Privacy Driven Ecosystems

The fourth factor (Table 5, Appendix) relates to the need of creating and using an ethical and legal framework anytime we build an AI ecosystem and includes keywords laws, ethics, liability, regulation, and privacy. EU prioritize data privacy and ethical AI development and has invested large amounts of money in this direction, but despite this the U.S. and China surpasses it in terms of the scale and speed of AI adoption, [66]. One of the reasons is the fact that EU puts a lot of efforts on ethics and sustainability to ensure a responsible AI growth, [5]. Another reason is the EU fragmented regulatory environment that interfere with the innovation of multimodal AI models, [67]. To ensure that Europe remains competitive, there is an urgent need for regulatory harmonization and clarity. Without decisive action, Europe may miss out on the technological advancements and economic gains that regions like the US, China, and India are poised to capture.

3.2.5 Factor 5: Innovation Based Ecosystems

The fifth factor highlights the strong connection between keywords such as innovation, ecosystem, and market (Table 5, Appendix), emphasizing the crucial role of innovative ecosystems within the broader AI landscape. As previously discussed in section 4.2.3, AI innovation ecosystems are essential to fostering technological advancement and entrepreneurship. In this context, the U.S. exemplifies a highly dynamic AI innovation ecosystem, driven largely by the private sector, where major technology companies such as Google, Amazon, and Microsoft lead both upstream research and downstream applications, [5].

3.2.6 Factor 6: Government Supported Ecosystems

The final factor addresses the vital role of government in fostering AI development (Table 5, Appendix). Governments can offer support through research funding, regulatory frameworks, and incentives for businesses and research centers. These incentives may include tax breaks, grants, and subsidies, crucial for encouraging innovation and making AI ecosystems more competitive. China's AI development, for example, benefits from a strong, government-led initiative to establish global leadership in AI by 2030. Chinese tech giants like Alibaba, Baidu, and Tencent receive extensive government support, which accelerates the commercialization of AI technologies across industries. The cooperation between these three pillars related to government, businesses and

education helps in the development and adoption of AI, [68]. On the other hand, the U.S. up to now has based the AI ecosystem development more on the private sector than on the government, even though for specific sectors such as the national security and defence there are a lot of government funding for AI research, [5]. The different national policies related to AI in the U.S., China, and the EU shows the different approaches to regulatory strategies, priorities, and fundings, [66], [69]. This divergence highlights how varying funding strategies and regulatory frameworks shape the dynamics of AI entrepreneurship across regions.

4 Conclusions

This paper uses recent publications on AI ecosystems and regional development and performs topic modeling and factor analysis to unveil various features of AI ecosystems. The results indicate that the most prominent topics in the AI ecosystem discourse during the last decade are startup and entrepreneurship ecosystems, policy and regulations (especially in the European Union), the role of technology, the impact of AI in the economic developments of geographic regions, and the importance of laws and regulations around the AI ecosystems.Also, special attention is placed on the applications of AI in healthcare and the human aspects of implementing AI applications. Further exploration of the keywords associated with the topics reveals six latent factors that characterize different models of AI ecosystems: human resources-driven, technology and algorithm-based, business and entrepreneurial-driven, legal, ethical, privacy, and regulatory framework, innovation-based, and government-supported ecosystems.

The findings of this study contribute to the existing body of literature as they organize the current research discourse on the AI ecosystem around specific topics and factors. The study also guides scholars, publishers as well as practitioners about the specific areas of future research in the field of AI ecosystems. In addition, government institutions can use these findings to create legal and ethical frameworks and policies for supporting startups, businesses, and universities to support the creation of AI clusters in their region.

5 Limitations and future research

Applying factor analysis to text documents is associated with several challenges. For example, due to the high-dimensional nature of text data, factor analysis can easily lead to overfitting of the data. Also, with text data, it is not always easy to determine which variables should be included in the analysis. The paper addresses these challenges by combining the LDA with EFA: the LDA approach is used to reduce dimensionality, extract the main discourse topics from the article corpus, and identify the most salient keywords of these topics. This process allows us to achieve an acceptable level of sampling adequacy and ensure a proper exploratory factor analysis.

While EFA examines the relationships between keyword frequencies, it's crucial to follow up with confirmatory factor analysis and empirical studies to assess the significance of these relationships. Additionally, the study would benefit from a broader corpus on AI, particularly focusing on the articles and white papers published after 2022, when there was a surge in AI literature largely driven by news about ChatGPT.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work the authors used ChatGPT, in order to improve readability and language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References:

[1] Choudhury, P., Starr, E., & Agarwal, R. *Machine learning and human capital complementarities: Experimental evidence on bias mitigation*, Strategic Management Journal, Vol. 41, No. 8, 2020, pp. 1381-1411.

[2] Agrawal, A., Gans, J., & Goldfarb, A. *The economics of artificial intelligence*, Innovation Policy and the Economy, Vol. 19, No. 1, 2019, pp. 111-127.

[3] Aghion, P., Jones, B. F., & Jones, C. I. *Artificial intelligence and economic growth*, The economics of artificial intelligence: An agenda, University of Chicago Press, 2019b, pp. 237-282.

[4] Brynjolfsson, E., Rock, D., & Syverson, C. *Artificial intelligence and the modern productivity paradox*, The Economics of Artificial Intelligence: An Agenda, Vol. 23, 2019, pp. 23-57.

[5] Jacobides, M. G., Brusoni, S., & Candelon, F. *The evolutionary dynamics of the artificial intelligence ecosystem*, Strategy Science, Vol. 6, No. 4, 2021, pp. 412-435.

[6] Aghion, P., Antonin, C., & Bunel, S. *Artificial intelligence, growth and employment: The role*

of policy, Economie et Statistique, Vol. 510, No. 1, 2019, pp. 149-164.

[7] Adner, R. *Match your innovation strategy to your innovation ecosystem*, Harvard Business Review, Vol. 84, No. 4, 2006, p. 98.

[8] Prahalad, C. K., Prahalad, C. K., Fruehauf, H. C., & Prahalad, K. *The Fortune at the Bottom of the Pyramid*, Wharton School Publishing, 2005.

[9] Porter, M. E. *Clusters and the new economics of competition*, Harvard Business Review, Vol. 76, No. 6, 1998, pp. 77-90.

[10] Scaringella, L., & Radziwon, A. *Innovation, entrepreneurial, knowledge, and business ecosystems: Old wine in new bottles?* Technological Forecasting and Social Change, Vol. 136, 2018, pp. 59-87. ISSN 0040-1625, https://doi.org/10.1016/j.techfore.2017.09.023.

[11] Feldman, R., & Sanger, J. *The text mining handbook: Advanced approaches in analyzing unstructured data*, Cambridge University Press, 2007.

[12] Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. *A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press*, Discourse & Society, Vol. 19, No. 3, 2008, pp. 273-306.

[13] Morley, J., & Bayley, P. (Eds.). *Corpus-assisted discourse studies on the Iraq conflict: Wording the war*, Vol. 10, Routledge, 2011.

[14] Roundy, P. T., & Asllani, A. *The factors of entrepreneurship discourse: A data analytics approach*, Journal of Entrepreneurship, Management and Innovation, Vol. 14, No. 3, 2018, pp. 127-158.

[15] Roundy, P. T., & Asllani, A. *Understanding the language of entrepreneurship: An exploratory analysis of entrepreneurial discourse*, Journal of Economic and Administrative Sciences, 2018.

[16] Aghakhani, N., & Asllani, A. *A text-mining approach to evaluate the importance of Information Systems research factors*, Communications of the IIMA, Vol. 18, No. 1, 2020, p. 3.

[17] Roundy, P. T., & Asllani, A. *Understanding AI innovation contexts: A review and content analysis of artificial intelligence and entrepreneurial ecosystems research*, Industrial Management & Data Systems, 2024. doi:10.1108/IMDS-08-2023-0551.

[18] Blei, D. M., Ng, A. Y., & Jordan, M. I. *Latent Dirichlet allocation*, Journal of Machine Learning Research, Vol. 3, 2003, pp. 993-1022.

[19] Griffiths, T. L., & Steyvers, M. *Finding scientific topics*, Proceedings of the National Academy of Sciences, Vol. 101, Suppl. 1, 2004, pp. 5228-5235.

[20] Hoffman, M. D., Blei, D. M., & Bach, F. R. *Online learning for latent Dirichlet allocation*, Advances in Neural Information Processing Systems, 2010, pp. 856-864.

[21] Jolliffe, I. *Principal component analysis*, 2nd ed., Springer, New York, 2002.

[22] Dhillon, I. S., & Modha, D. S. *Concept decompositions for large sparse text data using clustering*, Machine Learning, Vol. 42, No. 1, 2001, pp. 143-175.

[23] Zou, H., Hastie, T., & Tibshirani, R. *Sparse principal component analysis*, Journal of Computational and Graphical Statistics, Vol. 15, No. 2, 2004, pp. 265-286.

[24] Chaudhary, A., Hanmandlu, M., & Kaur, M. *Performance analysis of feature selection techniques for text classification*, International Journal of Computer Applications, Vol. 78, No. 1, 2013.

[25] Zhang, Y., Zhang, H., & Zhang, J. *Query expansion based on EFA in web search engines*, Chinese Journal of Electronics, Vol. 23, No. 4, 2014, pp. 724-727.

[26] Chen, Y., Liu, B., & Lai, S. *Dimensionality reduction for latent Dirichlet allocation*, Journal of Machine Learning Research, Vol. 13, 2012, pp. 667-699.

[27] Chen, Y., Chen, Y., & Zhou, L. *A joint model of text classification and topic modeling using Latent Dirichlet Allocation*, Proceedings of the 23rd International Conference on Computational Linguistics, 2010, pp. 951-959.

[28] Liu, Y., & Zhang, D. *A unified approach to text classification and topic modeling using Latent Dirichlet Allocation*, Proceedings of the 23rd International Conference on Computational Linguistics, 2010, pp. 947-950.

[29] Řehůřek, R., & Sojka, P. *Software Framework for Topic Modelling with Large Corpora*, Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 2010, pp. 45-50.

[30] Srinivasa-Desikan, B. *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*, Packt Publishing Ltd, 2018.

[31] Gardner, M. J., Lutes, J., Lund, J., & Hansen, J. *The topic browser: An interactive tool for browsing topic models*, NIPS Workshop on Challenges of Data Visualization, 2010.

[32] Gretarsson, B., O'Donovan, J., Bostandjiev, S., Höllerer, T., Asuncion, A., Newman, D., & Smyth, P. *TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling*, ACM Transactions on Intelligent Systems and Technology, Vol. 3, No. 2, 2012. doi:10.1145/2089094.2089099.

[33] Perkins, J. *Python text processing with NLTK 2.0 cookbook*, Packt Publishing, 2010.

[34] Korenius, T., Laurikkala, J., Järvelin, K., & Juhola, M. *Stemming and lemmatization in the clustering of Finnish text documents*, Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, 2004, pp. 625-633.

[35] Prabhakaran, S. *Topic Modeling with Gensim (Python)*, Machine Learning Plus, 2018. Available at: https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/

[36] Preet, P. *Standard Metrics for LDA Model Comparison*, GitHub, 2019. Available at: https://pahulpreet86.github.io/standard-metrics-for-lda-model-comparison/

[37] Kaiser, H. F., & Rice, J. *Little Jiffy, Mark IV*, Educational and Psychological Measurement, Vol. 34, 1974, pp. 111-117. doi:10.1177/001316447403400115.

[38] Chuang, J., Manning, C. D., & Heer, J. *Termite: Visualization techniques for assessing textual topic models*, Proceedings of the International Working Conference on Advanced Visual Interfaces, 2012, pp. 74-77.

[39] Giuggioli, G., & Pellegrini, M. *Artificial intelligence as an enabler for entrepreneurs: A systematic literature review and an agenda for future research*, International Journal of Entrepreneurial Behaviour & Research, Vol. 29, 2023, pp. 816-837. doi:10.1108/IJEBR-05-2021-0426.

[40] Yadav, M., Mittal, A., & Jayarathne, P. A. *Exploring untapped market niches with deep learning models*, Empowering Entrepreneurial Mindsets With AI, IGI Global, 2024, pp. 119-138.

[41] Singh, U., Rout, R., Dutta, G., & Patel, M. *Role of Technology Innovations in Providing Cutting Edge-Entrepreneurial Opportunities in India*, Journal of Informatics Education and Research, Vol. 4, No. 1, 2024.

[42] Gandía, J. A. G., González-Tejero, C. B., & Miguel, Á. J. Á. *New Lines of Business Development: Artificial Intelligence in Business*, Artificial Intelligence and Business Transformation: Impact in HR Management, Innovation and Technology Challenges, Springer, 2024, pp. 3-17.

[43] Smuha, N. A. *The EU approach to ethics guidelines for trustworthy artificial intelligence*, Computer Law Review International, Vol. 20, No. 4, 2019, pp. 97-106.

[44] Sarker, I. H. *AI-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems*, SN Computer Science, Vol. 3, No. 2, 2022, p. 158.

[45] Fonte, K. *The intersection of AI and emerging markets: Opportunities and challenges*, Samuel Curtis Johnson Graduate School of Management, 2024.

[46] Gonzales, J. T. *Implications of AI innovation on economic growth: A panel data study*, Journal of Economic Structures, Vol. 12, No. 1, 2023, p. 13.

[47] Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., de Prado, M. L., Herrera-Viedma, E., & Herrera, F. *Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation*, Information Fusion, Vol. 99, 2023, Article 101896.

[48] Kaswan, K. S., Gaur, L., Dhatterwal, J. S., & Kumar, R. *AI-based natural language processing for the generation of meaningful information from electronic health record (EHR) data*, Advanced AI Techniques and Applications in Bioinformatics, CRC Press, 2021, pp. 41-86.

[49] Lin, W. C., Chen, J. S., Chiang, M. F., & Hribar, M. R. *Applications of artificial intelligence to electronic health record data in ophthalmology*, Translational Vision Science & Technology, Vol. 9, No. 2, 2020, pp. 13-13.

[50] LIN, Y., Hsu, P.S., Cai, B.J., Hong, T.P. and Chen, R.C. *Text Mining Strategies: RoBERTa Optimization for Efficient Pain Assessment in Hospice Care*, International Journal of Applied Sciences & Development, 2024, Vol. 3, pp.166-170.

[51] Tiwari, P. C., Pal, R., Chaudhary, M. J., & Nath, R. *Artificial intelligence revolutionizing drug development: Exploring opportunities and challenges*, Drug Development Research, Vol. 84, No. 8, 2023, pp. 1652-1663.

[52] Schemmer, M., Kühl, N., & Satzger, G. *Intelligent decision assistance versus automated decision-making: Enhancing knowledge work through explainable artificial intelligence*, arXiv preprint arXiv:2109.13827, 2021.

[53] Jarrahi, M. H. *Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision-making*, Business Horizons, Vol. 61, No. 4, 2018, pp. 577-586.

[54] JIANG, X., *Human Resource Intelligent Recommendation Method based on Improved Decision Tree Algorithm*, WSEAS TRANSACTIONS on COMPUTER RESEARCH, 2024, Vol. 12, pp. 537-544.

[55] Ferrara, E. *Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies*, Sci, Vol. 6, No. 1, 2024, p. 3.

[56] De Vaus, D., & De Vaus, D. *Surveys in Social Research*, Routledge, 2013.

[57] Bresnahan, T. *Artificial intelligence technologies and aggregate growth prospects*, Prospects for Economic Growth in the United States, 2019, pp. 132-172.

[58] Ćudić, B., Alešnik, P., & Hazemali, D. *Factors impacting university–industry collaboration in European countries*, Journal of Innovation and Entrepreneurship, Vol. 11, No. 1, 2022, p. 33.

[59] O'Dwyer, M., Filieri, R., & O'Malley, L. *Establishing successful university–industry collaborations: Barriers and enablers deconstructed*, The Journal of Technology Transfer, Vol. 48, No. 3, 2023, pp. 900-931.

[60] OECD. *Adopted Updated OECD Principles on Artificial Intelligence*, 2024. Available at: https://digitalpolicyalert.org/event/19612-adopted-amendments-to-the-oecd-principles-on-artificial-intelligence-ai (Accessed: 20 September 2024).

[61] LeCun, Y., Bengio, Y., & Hinton, G. *Deep learning*, Nature, Vol. 521, No. 7553, 2015, pp. 436-444.

[62] Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., ... & Yoon, D. H. *In-datacenter performance analysis of a tensor processing unit*, Proceedings of the 44th Annual International Symposium on Computer Architecture, 2017, pp. 1-12.

[63] Wang, Y. *Synergy in Silicon: The Evolution and Potential of Academia-Industry Collaboration in AI and Software Engineering*, Authorea Preprints, 2023.

[64] George, B., & Wooden, O. *Managing the strategic transformation of higher education through artificial intelligence*, Administrative Sciences, Vol. 13, No. 9, 2023, Article 196.

[65] Kalogiannidis, S., Patitsa, C. and Chalaris, M. *The Integration of Artificial Intelligence in Business Communication Channels: Opportunities and Challenges*, WSEAS Transactions on Business and Economics, 2024, Vol. 21, pp.1922-1944.

[66] European Commission. *White Paper on Artificial Intelligence: A European Approach to Excellence and Trust*, Brussels: European Commission, 2020.

[67] Open Letter. *Europe Needs Regulatory Certainty on AI*, 2024. Available at: https://euneedsai.com/ (Accessed: 19 September 2024).

[68] Lee, K. F. *AI Superpowers: China, Silicon Valley, and the New World Order*, Houghton Mifflin, 2018.

[69] Mou, Y. *China's AI Ambition and Strategy*, AI & Society, Vol. 34, No. 1, 2019, pp. 27-42.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

APPENDIX

Table 1: Corpus of publications by category.

Category	Peer-reviewed High Impact Journal Papers	Other Peer-reviewed Journal Papers	Not Peer-reviewed or Working Papers	Conference Papers	Doctoral or Master Dissertations	White Papers Technical Reports or Posts	Total
No. of publications	67 (45%)	42 (29%)	17 (11%)	5 (3%)	3 (2%)	15 (10%)	149 (100%)

Table 2: Corpus of publications by year

Year of Publication	2022	2021	2020	2019	2018	=<2017	2012-2022
Average Citations per Publication	12	37	90	196	353	797	161
Number of Publications	21 (14%)	46(31%)	33(23%)	23(15%)	18(12%)	8(5%)	149

Table 3: Final Features of Exploratory Factor Analysis Model

Communalities			Communalities		
	Initial	Extraction		Initial	Extraction
algorithm	.346	.167	machine	.659	.609
capital	.511	.198	market	.623	.551
deep_learning	.499	.440	network	.262	.101
ecosystem	.717	.686	policy	.241	.135
employment	.786	.782	privacy	.245	.175
entrepreneurship	.381	.178	public	.688	.123
ethics	.357	.235	regional	.392	.225
firm	.310	.092	regulations	.507	.483
funding	.281	.075	risk	.436	.277
government	.708	.229	company	.195	.075
human	.328	.107	automation	.586	.337
industry	.371	.144	neural_network	.281	.254
innovation	.758	.738	skill	.743	.602
knowledge	.315	.149	liability	.571	.383
labor	.818	.885	worker	.530	.368
law	.664	.461	wage	.426	.363
learning	.619	.530	health	.214	.078

Table 4: KMO and Bartlett's Test of Sphericity

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.670
Barlett's Test of Sphericity	Approx. Chi-Square	1626.010
	df	561
	Sig.	<.001

Table 5: Loading of Keywords into Factors

Pattern Matrix[a]						
	Latent Factors					
	1	2	3	4	5	6
labor	.952					
employment	.883					
skill	.782					
wage	.609					
worker	.607					
automation	.587					
capital	.326					
machine		-.700				
learning		-.582				
deep_learning		-.549				
regional			.541			
industry			.424			
entrepreneurship			.384			
neural_network		-.379				
knowledge			.346			
firm			.332			
algorithm		-.319				
network			.316			
company			.306			
regulations				.711		
law				.700		
liability				.657		
risk				.486		
ethics				.451		
privacy				.398		
innovation					.894	
ecosystem					.860	
market					.668	
public						.974
government						.796
funding						.355
Extraction Method: Principal Axis Factoring. Rotation Method: Promax with Kaiser Normalization [a]Rotation converged in 8 iterations.						