Telecommunications Policy 44 (2020) 101943

Contents lists available at ScienceDirect

Telecommunications Policy

journal homepage: http://www.elsevier.com/locate/telpol

The AI techno-economic complex System: Worldwide landscape, thematic subdomains and technological collaborations

Riccardo Righi, Sofia Samoili *, Montserrat López Cobo, Miguel Vázquez-Prada Baillet, Melisande Cardona, Giuditta De Prato

European Commission, Joint Research Centre (JRC), Seville, Spain

ARTICLE INFO

Keywords:
Artificial intelligence
Complex systems
Agent-artifact space
Natural language processing
Semantic analysis
Network analysis

ABSTRACT

Artificial intelligence (AI) is playing a major role in the new paradigm shift occurring across the technological landscape. After a series of alternate seasons starting in the 60s, AI is now experiencing a new spring. Nevertheless, although it is spreading throughout our economies and societies in multiple ways, the absence of standardised classifications prevents us from obtaining a measure of its pervasiveness. In addition, AI cannot be identified as part of a specific sector, but rather as a transversal technology because the fields in which it is applied do not have precise boundaries. In this work, we address the need for a deeper understanding of this complex phenomenon by investigating economic agents' involvement in industrial activities aimed to supply AI-related goods and services, and AI-related R&D processes in the form of patents and publications. In order to conduct this extensive analysis, we use a complex systems approach through the *agent-artifact space* model, which identifies the core dimensions that should be considered. Therefore, by considering the geographic location of the involved agents and their organisation types (i.e., firms, governmental institutions, and research institutes), we (i) provide an overview of the worldwide presence of agents, (ii) investigate the patterns in which AI technological subdomains subsist and scatter in different parts of the system, and (iii) reveal the size, composition, and topology of the AI R&D collaboration network. Based on a unique data collection of multiple micro-based data sources and supported by a methodological framework for the analysis of techno-economic segments (TES), we capture the state of AI in the worldwide landscape in the period 2009–2018. As expected, we find that major roles are played by the US, China, and the EU28. Nevertheless, by measuring the system, we unveil elements that provide new, crucial information to support more conscious discussions in the process of policy design and implementation.

Disclaimer: The views expressed are purely those of the authors and may not in any circumstances be regarded as stating an official position of the European Commission. The authors R. Righi and S. Samoili contributed equally to this manuscript.

* Corresponding author.

E-mail addresses: riccardo.righi@ec.europa.eu (R. Righi), sofia.samoili@ec.europa.eu (S. Samoili), montserrat.lopez-cobo@ec.europa.eu (M. López Cobo), miguel.vazquez-prada-baillet@ec.europa.eu (M. Vázquez-Prada Baillet), melisande.cardona@ec.europa.eu (M. Cardona), giuditta.de-prato@ec.europa.eu (G. De Prato).

https://doi.org/10.1016/j.telpol.2020.101943
Received 30 April 2019; Received in revised form 19 February 2020; Accepted 9 March 2020
Available online 29 March 2020

1. Introduction

We experience the impact of technological disruptions every day. Nevertheless, investigations of how these changes are thriving in the economic system are affected by a lack of standard classification systems for emerging technologies that would allow a better detection of their pervasiveness in industry and research.[1] In order to provide indicators to support policy makers' decisions regarding the impact of emerging technologies, we propose an analysis that describes the complex system related to the activities of economic agents that are involved in a specific technology. In this study, we address Artificial Intelligence (AI) in the last decade (2009–2018). This work is developed in the context of a broader project that is focused on measuring the digital economy and assessing the emergence of transversal technological domains that can rapidly impact economic growth.

The aim of this study is to delineate the AI Techno-Economic Segment (TES) in a timely manner. The AI TES is a complex system consisting of agents that have taken part in relevant AI activities in the last decade. We consider relevant activities to include (i) industrial activities (i.e., production, services, and trade by firms) and R&D activities in the form of (ii) patent applications related to the considered technology and (iii) publications in related research domains. We consider agents to be the firms, research centres, academic institutions, and governmental institutions involved in the aforementioned economic and knowledge-related activities. We investigate this system in order to capture: (i) the location and organisation type of involved agents, (ii) the presence of AI thematic subdomains and the degree to which the agents and their countries are involved in each of them, and (iii) the structure of collaborative R&D technological relationships among agents.

The measurement of the fundamental dimensions of a TES makes this work original. More specifically, (i) we describe the presence of technological agents over the considered geographic areas[2] based on micro-level data (firms, research institutes, etc.) collected from different sources, (ii) we identify the thematic profile of countries and their involvement in technological subdomains, which allows the assessment of their technological competences, and (iii) we investigate emerging features in the structure of agents' interactions within and between areas. Another advantage of our study is that we include heterogeneous sources that address the industry and the research. The sources' variety enables us to cover different geographic regions and different stages of firms' lifecycles (from start-ups to established agents), as well as different types of R&D activities.

The considered technology, AI, has been selected because of its disruptive and transformative potential. The recent increase in data availability, processing power, and data storage, as well as the development of new algorithms, have triggered a new "spring" in AI (Russell & Norvig, 2016; Nilsson, 2014; Schwab, 2017; National Academies of Sciences Engineering and Medicine and others, 2017). This makes it highly relevant for policy (Craglia et al., 2018) in view of the attracted funds, expected impacts on industrial processes and labour market, and political-economic interests. However, while several studies describe AI applications and subdomains (Bughin et al., 2017, pp. 1–80) (see Appendix E), few measure the magnitude and structure of its pervasiveness in the economy (China Institute for Science and Technology Policy at Tsinghua University, 2019; WIPO, 2019). Moreover, there is no mutually agreed-upon definition of the AI technological domain, and it is not covered by a standard classification, such as the International Classification for Standards (ICS) (see Footnote 1). However, it is spreading in multiple economic sectors and its presence is embedded into different categories of products (De Prato et al., 2019). The representation and understanding of such a complex system are useful for policy (European Commission, 2018a, p. 237; 2018b, p. 795). For this reason, insights regarding the comprehension of agents' dynamics and specialisations are relevant for the design and implementation of policy interventions, which might be less effective and efficient if not supported by this knowledge.

This paper is structured as follows. In Section 2, we present the related literature, the background, and the research questions. Section 3 describes the proposed methodology. In Section 4, the process of data collection is described. The results are discussed in Section 5. In Section 6, the conclusions of the study and the perspectives for its advancement are presented.

2. Background and research questions

Technology plays an essential role in the field of economics. Neoclassic economic theories consider the economy as a system converging towards an equilibrium state that is strictly dependent on the available technology (Solow, 1957). Even if the role of technology is fundamental in this approach, it fails to give a description of the process that leads to the formation of that specific technological availability. In contrast to these approaches, subsequent studies in the fields of (i) macro-economic out-of-equilibrium theories, (ii) national and local innovation systems, and (iii) evolutionary economics, respectively, have made the following claims. First, technology is not an exogenous variable, but is instead the result of an endogenous process (Hicks, 1973; Georgescu-Roegen, 1971; Amendola & Gaffard, 1988, 1998).Second, interactive dynamics among involved agents foster technologically innovative

[1] The lack of a mutually agreed-upon definition and classification system, such as the International Classification for Standards (ICS), for an emerging technology, such as AI, is evident in a comparison of the significant variance of definitions and classifications in research, market, policy, and institutional publications (see 23 relevant research publications, 29 AI policy and institutional reports, and 3 market reports that can be found in Appendix E). In particular, regarding the lack of standard classifications for AI, the International Organisation for Standardization (ISO) is working to identify potential standard specifications, characteristics, and subcategories of AI that could be used internationally to classify the AI technological domain. The ISO is developing 10 AI standards for ISO/IEC (joint technical committee of the International Organisation for Standardization and the International Electrotechnical Commission). As of January 2020, three standards have been published with different objectives (big data overview, vocabulary, reference architecture, etc.); hence, AI definitions and classifications are not included.

[2] In this work, we refer to geographic areas, but the analysis can be developed at the level of regions, subregions, or cities.

bottom-up processes (Lundvall, Dosi, & Freeman, 1988, pp. 349–369; C.; Freeman, 1991; Nelson, 1993; Saxenian, 1994; Breschi & Malerba, 1997).Third, ontological recursive elements characterising technological evolution and innovation exist and can be identified (Dosi, 1982; Lane & Maxfield, 2005; Arthur, 2007, 2009).More recently, the study of emerging technologies has moved in the direction of patent analysis, especially because of the large availability of data. Such analyses are based on network indicators with patent citations (Breitzman & Thomas, 2015; Cho & Shih, 2011; Verhoeven, Bakker, & Veugelers, 2016), analysis of patent textual content (Gerken & Moehrle, 2012; Lee, Kang, & Shin, 2015), and machine learning processes to identify emerging technologies at early stages (C. Lee, Kwon, Kim, & Kwon, 2018). While these studies explore different methodologies to evaluate emerging technologies, they tend to not to associate the disruption caused by new technologies with the agents that generated such a disruption.

In order to consider the connection between economic agents and the technologies in which they are involved, we employ complex systems theory. Complex systems are defined as a set of interdependent agents whose joint behaviours reveal the emergence of a non-random structure (Newman, 2011). This approach has been proved pertinent to the representation and examination of economic phenomena, such as the development and spread of new technological domains in economic processes (Arthur, 1999; Arthur, Durlauf, & Lane, 1997). In particular, our work uses the concepts of the *agent-artifact space* theoretical model, which was developed for the study of emerging technologies (Lane & Maxfield, 1997, 2005), in order to explore the main dimensions of the complex system that we address. In this model, *agents* are the fundamental units of the system and the behaving entities by whom all activities are initiated. Depending on the type of system, *agents* can be individuals in a society, employees in a firm, or enterprises in local industrial districts. *Artifacts* are the technological objects with which the *agents* work in order to carry out their economic activities and, occasionally, to develop new functionalities, i.e., theoretical and empirical advancements allowing for changes in the technological landscape (Lane, Pumain, van der Leeuw, & West, 2009). Finally, the crucial element of the *agent-artifact space* theoretical model is that the process of technological evolution is triggered by *agents*' interactive dynamics (Lane, 2011, 2016).

The very first artifacts, such as stone tools and metal objects, allowed humans to relate with the environment in new ways (Lane et al., 2009; van der Leeuw, 2008; van der Leeuw & McGlade, 1997). The increase of knowledge and capability in handling matter, energy, and information led to the development of AI, which was ultimately enabled by the ICT revolution (Lane et al., 2009; Lane, van der Leeuw, Sigaloff, & Addarii, 2011; Young et al., 2006). Over time, social and economic interactions of increasing complexity have sustained radical technological changes, which in turn generated feedback loops affecting the structure of our societies and economies. In this work, we use this evolutionary and adaptive approach and consider all economic institutions worldwide that have been involved in the AI domain in the last decade as *agents*, i.e., firms, research centres, and governmental institutions. AI's *artifacts*, such as machine learning algorithms, devices using computer vision and speech recognition, and connected and automated vehicles, are investigated via the activities that the *agents* perform in order to produce, trade, and develop them. Hence, as activities, we consider the economic processes directed to (i) the supply of AI-related goods and services (industrial activities) and R&D activities in the form of (ii) patent applications with AI technological developments and (iii) AI-related academic research. These types of activities are selected in order to cover the main and most relevant agents' initiatives that concern AI's *artifacts*. Some of these are "individual" activities, i.e., the supply of goods and services does not imply any collaboration between agents; its design and achievement are carried out individually by the *agent*. On the other hand, there are "shared activities"; a patent or a publication can be developed by a single agent or by several agents in collaboration. Therefore, the work aims to answer the following research questions: *(i) Who are the agents involved in the supply and evolution of AI worldwide? (ii) Which types of artifacts are being developed in this field? (iii) How do agents interact and behave in the techno-economic space that is considered?*

Moving from a micro-based perspective that allows us to capture both pertinent activities in the AI technological domain and the involved agents, the present work aims to provide statistical and quantitative answers so as to measure the AI TES as a complex techno-economic system from the perspective of an *agent-artifact space*. The three aforementioned research questions lead to the analysis of the AI TES in three dimensions. The first dimension is related to the mapping of the involved agents, in terms of location (at the country level) and organisation types. This initial part is fundamental for the comprehension of the considered complex system's structure. The second dimension is related to the investigation of the technological domain. The technological subdomains are identified within the AI domain using natural language processing methods. Hence, each subdomain represents one category of *artifacts* that the *agents* have developed. The third dimension is related to the technological collaborations in which the agents are involved. Hence, the core elements of the agent-artifact space model are addressed, namely agents' interactions.

3. Methodology

The methodological approach that we propose consists of three parts and can be implemented in any technological domain.[3] In this work, the proposed approach is developed for AI. The first part includes data collection, text pre-processing, and the identification of agents. The main elements of this methodological part are outlined in Subsection 3.1, and a detailed description is presented in Appendix A. In the second methodological part, we detect thematic knowledge subdomains through topic modelling. This is described in Subsection 3.2 and represented in Fig. 2. The corresponding results are discussed in Subsection 5.2. Finally, the third and last part focuses on the investigation of R&D collaborations and network centrality statistics. This is implemented through a geo-based network, as described in Subsection 3.3 and represented in Fig. 3. The corresponding results are discussed in Subsection 5.3.

[3] The present study is based on previous work in the domain of photonics (Samoili, Righi, Lopez-Cobo, Cardona, & De Prato, 2018).

3.1. Construction of the graph database

The first methodological part in this work consists of the collection and pre-processing of the information that allows us to answer the research questions. Some elements are highlighted in this paragraph and represented schematically in Fig. 1. This method is fully described in Appendix A, as it is out of the scope of this study to focus on these methodological details. The numeration of the steps that are described below correspond to the numeration of Appendix A and Fig. 1. Initially, data regarding relevant activities are collected and the corresponding textual information is pre-processed with text-mining approaches (step a). Second, the textual information is normalised (step b.1) and a keyword extraction algorithm is implemented to select the most relevant technological terms (step b.2). Third, agents' locations (at the city level) are corrected or completed where applicable (step c.1). Agents' names are then disambiguated, as they can be simultaneously detected by multiple data sources (step c.2). In addition, information about the organisational type (firm, governmental institution, or research institute) of each agent is determined (step c.3). Finally, the graph database is generated (step d).

We consider agents to be economic entities located in a specific city, which implies that if the same economic entity has multiple locations, a distinct agent is identified in each location (see Appendix A.c.1). The retrieval of agents' locations at high granularity (city level) allows the investigation of the system at different geographic levels (e.g., countries, regions). Second, the agents' organisation types are classified according to the literature on Triple Helix (Etzkowitz & Leydesdorff, 2000), a theoretical framework for the study of institutional interactions that are able to foster economic and social development. Agents are distinguished in three types: firms, governmental institutions, and research institutes (the initial capital letters F, G, and R are used to represent agents' organisation types in Fig. 1). More details are provided in Appendix A.c.3. Finally, distinctions are made between documents and activities in this study. Namely, a document carries only textual information about the corresponding economic activity. For instance, while a patent indicates an R&D activity, its textual component (the document) refers to the abstract, title, and keywords. Similarly, the supply of goods/services indicates an industrial activity, and the firm's business description constitutes the corresponding document. It is necessary to clarify the differences between these components for the semantic analysis that is presented in Subsection 3.2, in which technological knowledge subdomains are identified (research question (ii), methodology in Subsection 3.2 and results in Subsection 5.2) and the technological subdomains per geographic location and organisation type are further assessed.

Based on the collected data, the results of the agents' location mapping and the organisation types' attribution are discussed in Subsection 5.1. It should be noted that the data sources are distinguished in two cases, and a different data collection strategy is implemented for each case. These two cases are:

case 1. - vertical sources: import technologically specific data sources according to experts' input, i.e., databases exclusively including information about agents and activities related to the considered technology,

case 2. - horizontal sources: query non-technologically specific data sources with terms relevant to the technology, i.e., databases including information about agents and activities not necessarily related to the considered technology.

Based on this data collection strategy, we capture agents' involvement in (i) industrial activities (i.e., production, services, and trade by firms), and R&D activities in the form of (ii) patent applications and (iii) publications.

3.2. Detection of thematic knowledge subdomains

In order to detect agents' involvement in the knowledge subdomains of an emerging technology in a non-heuristic and objective method, a semantic analysis is performed on the documents of the agents' activities.[4] With this analysis and through the agents, we can quantify the involvement and thematic nature of countries' activities regarding the detected subdomains. In particular, to identify a technology's subdomains in the absence of a standard classification, such as the International Classification for Standards (ICS), for emerging technologies, such as AI (See Footnote 1), we use a natural language processing method called topic modelling, more specifically the Latent Dirichlet Allocation (LDA) model. LDA enables the clustering of all documents associated with industrial and R&D activities (i.e., firms' descriptions, patent applications, and publications). These activities are collected through a procedure that ensures that they sufficiently represent the technological domain (description in Appendix A). In this manner, these clusters can be associated, in a following step, with technological subdomains.

In literature, topic modelling is found to be used for the technological mapping of knowledge subdomains (Börner, Chen, & Boyack, 2003; Boyack, Klavans, & Börner, 2005; Leopold, May, & Paaβ, 2004; Spitters, Verbruggen,& van Staalduinen, 2014; Suominen & Toivanen, 2016), and in recent technological forecasting studies, as an objective method to analyse and classify the technological areas and sub-areas of R&D activities (Hu, Fang, & Liang, 2014; Lee et al., 2015; Suominen & Toivanen, 2016; Suominen, Toivanen, & Seppänen, 2017; Venugopalan & Rai, 2015). It is also employed to extract technology intelligence through a patent analysis at the industry level (Suominen & Toivanen, 2016), and with additional related statistical methods (Latent Semantic Indexing and Principal Components Analysis (PCA)) for publications and patent abstracts (Zhang, Porter, Hu, Guo, & Newman, 2014). In another study, LDA is applied to investigate technology convergence with patent data as an essential step in the pursuit of innovation and economic growth (Lee et al., 2015). Topic modelling is again invoked to address the subjectivity of patent classification and the non-unique mapping of industries to classes and vice-versa (Venugopalan & Rai, 2015). A combination of LDA, integrated k-means, and PCA is employed for

[4] It should be noted that the document is the textual information part of the activity (see Appendix A.a).

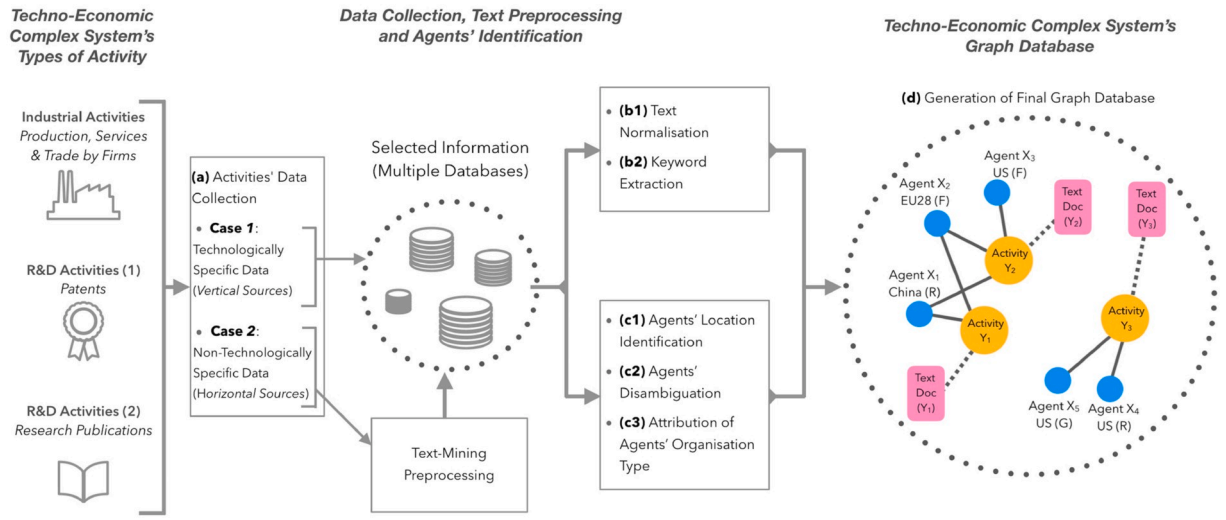


Fig. 1. Schematic representation of the first part of the developed methodology. Starting from the conceptualisation of the activities characterising the techno-economic complex system (left part), a series of procedures are implemented as described and enumerated in Appendix A (middle part), producing as a final output a graph database (right part). Notes: (i) in the right part, in parenthesis the agents' organisation types are reported ("F" for firms, "R" for research institutes and "G" for governmental institutions); (ii) the example of agents' location is here at the level of geographic areas, but the agents' identification is at city level.

the automatic classification of patents and the development of a knowledge organisation system (Hu et al., 2014).

Topic models are preferred to traditional clustering algorithms in information retrieval and have been found to outperform cluster models (Wei & Croft, 2006; Zhao, Zou, & Chen, 2014). More specifically, the advantage of topic modelling is that a training process is not required for this unsupervised learning algorithm. Hence, LDA can automatically discover latent semantic patterns, namely topics (Blei, Ng, & Jordan, 2003; Griffiths & Steyvers, 2004), in any given set of documents (Blei et al., 2003; Griffiths, Steyvers, & Tenenbaum, 2007) without predetermined classes through a generative process.

In particular, LDA, a commonly used probabilistic generative topic model, semantically clusters the content of a set of documents, called a corpus, into topics that can be identified only through LDA - explaining the use of "latent" in the naming of the model. The generative process includes the assignment of each document's terms to random variables.[5] These terms are semantically clustered into topics using an iterative probabilistic process and Dirichlet distribution (multivariate generalisation of the Beta distribution) (Blei, 2012). Through LDA, the probability distributions of topics in documents (topics' probabilities) and of words in topics (terms' probabilities) are estimated. This means that a probability distribution is created for each document, indicating the extent to which it belongs to each of the topics, and a probability distribution is created for each word of the corpus, indicating the extent to which it belongs to a topic. The model returns the topics that are more likely to describe each document adequately, and as each document is part of the entire collection of documents (corpus), the most probable topics that better represent the corpus.

As LDA does not require the implication of experts in the process of identifying topics, the risk of unintentional bias is minimised (Blei et al., 2003; Jurafsky & Martin, 2014; Steyvers & Griffiths, 2007). However, it is noteworthy that in order to classify the documents into topics, the LDA algorithm requires the user to specify a fixed number of topics. Although the authors who proposed LDA in the context of machine learning attempted to provide a statistical indicator to compute the optimal number of topics that fits a given corpus, called the perplexity value (Blei et al., 2003), the evaluation of the fit is of limited value as the topics' interpretability by humans is difficult, even with standard quantitative measures of fit (Chang, Gerrish, Wang, Boyd-Graber, & Blei, 2009). In our approach, we first investigate the approximate number of proposed subdomains of the technology to be investigated with a literature review of research, policy, and market publications. We then follow a trial-and-error approach to determine the number of topics that better describe the corpus of the technology in question, testing with numbers of topics around the boundaries for the number of subdomains found, as in Suominen and Toivanen (2016). To evaluate the output, a list of the most probable terms per topic are examined for coherence through the term probability distributions per topic and are then compared to real-world subdomains of the technology reported in the literature.

The advancement of our implementation of topic modelling in comparison to relevant literature consists of identifying emerging technological subdomains with the titles, abstracts, and keywords (when available) of both R&D and industrial economic activities (firms' descriptions, patent applications, and publications). This part of the analysis aims to (i) quantify countries' involvement in each subdomain in order to assess their technological competences in the future and (ii) inquire into the composition of each subdomain per

[5] Random variable α: controls the topic distribution per document (the probability that a document contains a topic). Random variable β: controls the word distribution per topic (the probability that a topic contains a word).

organisation type of agents. In Fig. 2, a schematic explanation is presented regarding the ways in which topic modelling enables the two aforementioned points of this analysis by uncovering latent technological patterns.

The initial information considered concerns agents' involvement in activities. In Fig. 2.a agents are represented as X_1, X_2, X_3, X_4 and X_5 (in blue), and activities as Y_1, Y_2 and Y_3 (in yellow). The knowledge space that the agents' activities create is here accounted as the textual information contained in each activity, which corresponds to a different document represented by $Doc(Y_1)$, $Doc(Y_2)$, $Doc(Y_3)$ (in pink). The terms contained in the documents describing the content of agents' activities are then clustered with LDA into thematic groups, also referred to as topics (Cronen & Pearce, 1980; Ding, 2011; Hacklin, 2007), which are represented in Fig. 2b-d as *Topic 1*, *Topic 2* and *Topic 3* (in multiple colours). The documents and their terms are assigned to each topic with a probability, so each document is described by a topic with a higher or lower probability.[6] Fig. 2.c illustrates the step during which the topics are connected to the agents through the activities. Consequently, in Fig. 2.d, the thematic profile of each country is established by taking into consideration information regarding the agents' locations.

Finally, the following procedure is followed regarding the correspondence of topics to thematic subdomains of the studied technology. First, the top terms in the probability distributions of topics are automatically summarised. These summaries constitute the initial titles of the topics. To connect them to a technological subdomain, their correspondence to real-world topics is evaluated by comparing them to thematic subdomains found in the literature relevant to the technology. This literature is also used to identify the most representative number of topics. The titles of topics are then revisited and described in a more natural language. In the final step, they are evaluated by an expert who is provided with a list of the most frequently used terms per topic to assess each topic's correspondence to a technological subdomain.

3.3. A geo-based network of R&D collaborations

Network approaches have long been used in the field of economics to investigate innovation dynamics. The main examples in the literature concern the study of different roles in networks of innovation systems (Ibarra, 1993; Lundvall et al., 1988, pp. 349–369; Powell et al., 2005), regional knowledge clusters (Saxenian, 1994), the role of strategic alliances in firms' capabilities to innovate (C. Freeman, 1991; Mowery & Teece, 1996, pp. 111–129; Owen-Smith & Powell, 2004; Powell et al., 2005; Gilsing, Nooteboom, Vanhaverbeke, Duysters, & van den Oord, 2008), attractiveness in research collaborations (Abbasi, Hossain, & Leydesdorff, 2012; Balland, 2012), and the emergence of technologies (Breitzman & Thomas, 2015; Cho & Shih, 2011; Verhoeven et al., 2016). In addition, a large amount literature investigating trade flows with networks is present (Serrano & Boguñá, 2003; Garlaschelli & Loffredo, 2005; Fagiolo, Reyes, & Schiavo, 2008, 2009, 2010).By focusing on the connective and interactive nature of the phenomena observed, these studies have been able to reveal insights regarding the topological properties of the agents involved and of the global properties of the system considered.

Given the point that we aim to address in this final methodological part, agents' interactions, we implement network analysis. The goal is to study the network not at an individual level, but to explore differences at a specific geographic level. The objectives are multiple. The first is to analyse the level of internal and external collaborations observed by area. The second is to collect insights on different patterns in the composition of these collaborations. The third is to infer the relevance of each area in the worldwide landscape. To answer these questions, we construct a geo-based network starting from agents' locations and their participation in activities, hence at the micro level of analysis. It is also important to note that the proposed geo-based network, in addition to enabling an investigation at the geographic area level, allows us to overcome the lack of connections between components in the initial network, as represented in Fig. 3. As agents are grouped based on their locations, it is very likely that the computed geo-based network will present a single component.[7] This solves potential problems of analysis and interpretability when analysing the topology of the nodes.[8]

In order to investigate the structure of technological collaborations, the following steps are implemented. First, in the graph database (see Fig. 3a), only agents and activities are selected (i.e., blue and yellow nodes, respectively) in order to obtain a bipartite network structure (see Fig. 3b). Industrial activities (supply of technological products and services) are not considered in this part of the analysis, because this step intends to capture the structure of the most technologically advanced collaborations, i.e., R&D activities,

[6] If, for example, document $Doc(Y_1)$ has a 70% probability of belonging to *Topic 1*, 1% to *Topic 2*, and 29% to *Topic 3*, then the activity Y_1 is mostly speaking about *Topics 1* and *3*.

[7] A network consists of a single component when at least one path of connection exists between any couple of nodes.

[8] If the network is disconnected, i.e., presents more components, some centrality measures (like closeness) cannot be computed (Wasserman, Faust, et al., 1994; Opsahl, Agneessens, & Skvoretz, 2010), and others (like betweenness) assess the nodes' position with respect only to the nodes located in the same component (instead of assessing the position of the node with respect to all of the nodes belonging to the system).

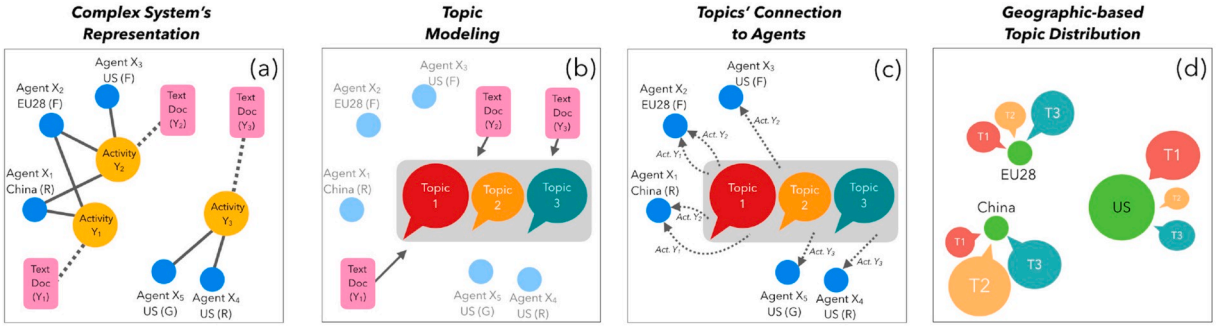


Fig. 2. Schematic representation of network of agents and their activities with mutual technological language. Technology-related collected activities are notated as Y_1, Y_2, Y_3, their textual information as documents $Doc(Y_1)$, $Doc(Y_2)$, $Doc(Y_3)$, and agents as X_1, X_2, X_3, X_4 and X_5. In **(a)** the technology's complex system is illustrated with the connections of the activities to the agents (characterised by geographical location in terms of area). In **(b)** the thematic topics are extracted from the textual information of the activities through topic modelling. In **(c)** the activities are connected again to the agents carrying the thematic information. In **(d)** the thematic profile of each country is presented.

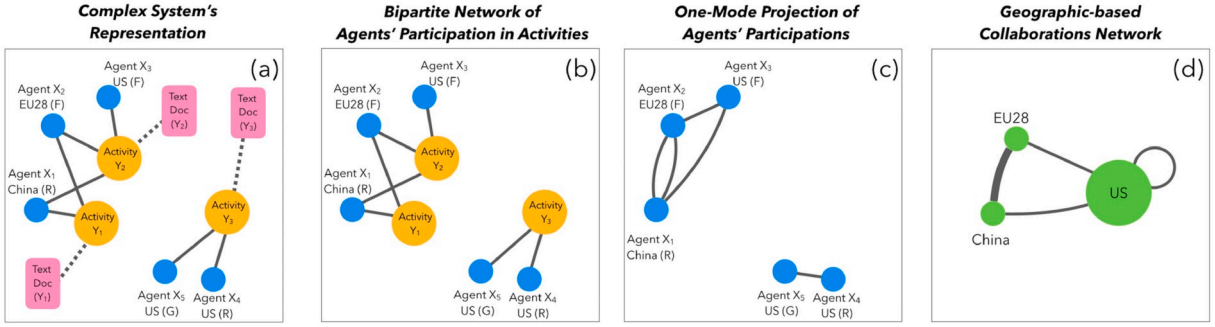


Fig. 3. Schematic representation of the generation of the geo-based network. The complex system, made of agents (blue nodes), activities (yellow nodes) and textual information (pink nodes) is represented in **(a)**. In **(b)**, the bipartite network consists of two types of nodes, representing agents and activities. Agents are connected to the activities in which they participate. In **(c)**, the one-mode projection based on agents is generated by connecting agents who participate in the same activity. In case of multiple common activities, multiple edges are generated. Finally, in **(d)** the geographic-based network is built by merging together the nodes belonging to the same geographic area. Edges are weighted based on the number of collaborations between agents of the corresponding areas and, for collaborations involving agents of a same area, self-loops are generated. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

which are oriented to generate new developments in the agent-artifact space.[9] Given the bipartite structure of the considered network,[10] agents are not directly connected to each other. Instead, they are indirectly connected to the common activities in which they participate. It is only possible to analyse their relationships after the network is projected in its corresponding one-mode network.[11] The resulting weighted adjacency matrix[12] describes the number of common activities in which each couple of agents participate (i.e., the one-mode projection based on agents). The initial bipartite network is thus transformed into a network in which agents are directly connected by means of their activities (Fig. 3c). In the case of an activity developed by a single agent, no connection is generated in the one-mode projection of the network, as no collaboration is established. In the case of n common activities, $n \in \mathbb{N}^+$, between the same couple of agents, n edges are generated between the couple. The only assumption that is made in the transformation from a bipartite network to its one-mode projection is that each agent involved in a common activity is directly collaborating with all the other agents involved. In order to investigate R&D collaborations from the perspective of geographic areas, a second

[9] Other network analyses are possible through exploiting the data resulting, among others, from ownership structure, venture capital investments, etc. These extensions are under development.

[10] A bipartite graph is a network in which two types of nodes exist - in this case agents and activities - and connections are present only between one node of the first type and one node of the second.

[11] This is possible by multiplying the binary incidence matrix describing the bipartite network with its transpose matrix.

[12] The adjacency matrix, i.e., a squared matrix describing connections among the nodes of a network, in this case is set with null diagonal. Let A be the initial incidence matrix describing the bipartite network with the number of rows equal to the number of agents and the number of columns equal to the number of activities, then the $diag(A \cdot A')$ describes the number of activities in which each agent is involved. However, in graph theory, this implies the presence of self-loops, i.e., connections connecting one node with itself. Therefore, after the described transformation, all the elements of the diagonal are set equal to 0.

transformation is implemented, leading to the geo-based network as represented in Fig. 3.d. Agents are grouped by their locations, resulting in a network in which (i) the nodes represent different geographic areas, (ii) the edges represent collaborations between agents of different areas or between agents of the same area (the latter are represented as self-loops[13]), and (iii) the edges are weighted based on the number of collaborations between agents of the corresponding areas.

This final network is analysed in three ways. The first is through the study of the volume of collaborations, namely the amount of collaborations within and between geographic areas, as represented by Fig. 7 for the case of AI. The second is the study of the composition of these collaborations, distinctions between the collaborations in terms of the location of the peers (within or between areas) and their organisation types, as presented in Fig. 8. This enables an investigation of the extent to which the different combinations of types of agent are involved in the observed collaborations. The third is the study of the level of control that is exerted by each geographic area on the rest of the network by means of the computation of a network statistical indicator, namely the weighted betweenness centrality (WBC) (L. Freeman, 1977, 1978; Brandes, 2001). The betweenness centrality (BC) of a node $v \in V$, where V is the set of nodes of the considered system, is defined as $BC(v) = \sum_{s \neq t \neq v} \frac{\sigma_{s,t}(v)}{\sigma s,t}$, where $\sigma_{s,t}$ indicates the number of shortest paths between s and t, where $\sigma_{s,t}(v)$ indicates the number of shortest paths between s and t in which v lies, and with $s \in V$ and $t \in V$. In order to account for weighted connections, the computation of the shortest paths is set to select the path with the largest[14] sum of weights. As information about time is present for R&D activities (patents and publications) (see Appendix A.a); nodes' WBC is computed on a yearly basis, considering only the set of collaborations occurring in each year at a time. The statistical indicator is calculated for the entire network of R&D activities, and for its two subnetworks separately (i.e., the subnetwork of patents, and the subnetwork of publications)). For each of these three cases, the year-by-year WBC is normalised in the interval $[0, 1]$.

4. Data collection for the AI TES

The process of data collection for the study of the AI TES is presented in this section, following the methodology that is outlined in Subsection 3.1 and Appendix A. In order to target the AI TES and collect data about AI-related industrial and R&D economic processes, we selected (i) a list of AI-specific sources (case 1 - vertical sources), and (ii) a list of horizontal databases containing information regarding any technology (case 2 - horizontal sources). The sources that are queried and collected are detailed in Appendix B. For publications in the AI domain in particular, only the top 10 international AI conferences are considered, following the findings of a recent study (Frank, Wang, Cebrian, & Rahwan, 2019) that AI researchers present increasing interest in publishing their work in topic-specific conferences instead of academic journals. The final list of AI conferences that are considered can be found in Appendix C.

In order to identify the activities that are relevant to the considered technological domain, horizontal sources (case 2) are queried with the following list of terms that resulted from the process described in Appendix A.a: *artificial intelligence, automated vehicle, chatbot, computer vision, deep learning, face recognition, knowledge representation and reasoning, machine learning, natural language processing, neural network, reinforcement learning, robotics, speech recognition, swarm intelligence,* and *virtual assistant*. The text-mining procedure is based on Elasticsearch engine version 2.4.6 (Lucene library). Both for vertical sources (case 1) and horizontal sources (case 2), only activities that occurred in the period 2009–2018 are considered. The last decade was selected for two reasons. First, this scope allows us to focus on the recent AI spring. Second, the data suffers from a limitation; the vertical data sources describe the core business of firms currently involved in AI without providing information about the year in which they first became involved in AI-related processes.[15] This prevents us from extending the period of time for the analysis. Therefore, we set the considered time period to 2009–2018. It is important to note that in the present work, we consider the time dimension only when analysing R&D activities, as time-related information is present for such activities (year of registration for patent applications and year of publication for research articles).

There are 56,318 activities collected by considering vertical sources and queried horizontal sources referring to 2009–2018. However, in 3706 cases, activities belonging to the vertical sources present no textual information regarding the description of the core business, only listing the name of a candidate agent. As one of the scopes of this work is to perform a textual analysis, we remove these observations. Therefore, the number of activities considered is 52,612. In these activities, 35,276 agents are initially detected. As discussed in Appendix A.c.1, we identify agents based on their geographical location at the city level. Complete geographical information is only available for 60% of agents (in the form of structured variables or in the form of addresses that are processed to extract information about country, region, subregion, and city). For 15% of agents, geographical information is retrieved based on their name. For 11% of agents, geographical information is obtained online. For 3% of agents, geographical information is imputed based on a probabilistic distribution of cities, considering that partial information is available (i.e., the country, region, and subregion of the agent). The remaining 11% of agents that do not have any kind of geographical information and for which it is not possible to retrieve such information online, are discarded. Once the geographical location is completed for all agents, the process of disambiguation is implemented (see Appendix A.c.2). Out of a set of 31,396 agents resulting from the preceding steps, 2146 groups of potential duplicated agents, overall involving 9438 agents, are detected. After a manual check of the detected cases, 1809 groups (out of the

[13] While the diagonal of the adjacency matrix obtained in the previous transformation was set equal to zero, the diagonal of the adjacency matrix of the graph describing connections between geographic areas is computed based on the number of collaborations among agents of the same area.

[14] Edges' weight in social networks usually measures a distance between two connected nodes. In this work, it is the opposite. As the weight of the edges indicates the number of collaborations, the more two areas collaborate, the closer we consider them.

[15] These sources were collected at the end of 2018.

2146 detected) are identified to contain duplicated agents, and this leads to the deletion of 2115 agents.[16] Additionally, 60 cases are refined to control for major agents (e.g., Google, IBM, Microsoft, Amazon, Toyota, Mitsubishi, Sony, EPFL, Carnegie Mellon University, Max Planck Institute, MIT). Thus, another 187 agents are eliminated because they are considered to be duplicates. In the end, the number of agents is reduced to 29,094. It is important to note that the agents considered as duplicates are those that: (i) are detected to represent the same economic entity and (ii) are located in the same city.

During the disambiguation process, the activities in which duplicated agents are involved are associated with the "original" non-deleted agent. If the same agent is associated with multiple industrial activities, these activities are merged. This led to the removal of 1915 industrial activities, resulting in a total of 50,697 industrial and R&D activities. On the other hand, if the disambiguation process leads to the association of multiple R&D activities with the same agent, the activities are kept distinct.

Regarding the attribution of the 29,094 agents' organisation type (see Appendix A.c.3), in 73% of the cases, the organisation type is univocally detected from the data sources. In order to homogenise the different categorisations of organisation types detected in the data sources, all the different types are converted to the considered types.[17] Three percent of agents present more than one organisation type and are manually disambiguated. For the 7262 agents without organisation type, we infer the missing data based on the presence of meaningful part-of-words in their standardised name.[18] As 3312 agents are still not allocated to any organisation type after these steps and specific patterns are not detected in their names, we consider them to be firms. With this last step, we finally define the 2009–2018 AI techno-economic complex system as a system made of 29,094 agents involved in 50,697 AI-related activities containing textual information.

Table 1 presents, per type of activity, the number of activities and agents collected and used for this study. The percentage of agents detected in more than one type of activity is low because of the following two cases. First, the adopted definition of agents, as described in Appendix A.c.1, considers the combination of an economic entity and its location at the city level. Therefore, if an economic entity is located in several sites, then each site is usually associated with different types of economic processes (e.g., legal offices, research laboratories, production plants). This generates distinct agents and no overlap is detected. Second, infrequent overlap between agents detected through patenting activities and agents detected through industrial activities is expected. Only technologically specialised agents are detected in the latter case, i.e., big high-tech companies and small-medium companies mainly providing AI solutions and services, because agents involved in industrial activities are detected based on the core description of their business. Agents making only partial use of AI in the products they supply, e.g., companies belonging to the automotive industry, are not detected among AI industrial activities, as AI is not explicitly included in the core description of their business. Nevertheless, these agents are detected when they filed AI-related patents or authoring AI-related articles.

5. Results and discussion

The section is divided into three subsections in accordance with the three dimensions that are investigated. First, in Subsection 5.1, an overview of the AI techno-economic complex system is presented. In particular, we focus on the distribution of agents in terms of their location and organisation type. Second, in Subsection 5.2, the technological subdomains emerging in the AI domain are presented. In addition, we assess the agents' involvement in each AI subdomain by geographic area and organisation type. Third, the technological collaborations in R&D are investigated in Subsection 5.3. In particular, we analyse the volume of internal and external collaborations, their composition in terms of the organisation type of the agents involved, and the control level exerted by each country in the considered network and its sub-networks.

For the implementation of the analysis, we group countries into geographic areas, which better represent geo-political areas. These are: Africa, Canada, China, the EU28, India, Japan, Middle East, Oceania, Other American countries, Other Asian countries, Other European countries, South Korea, and the US. Hence, the EU28 is presented as an individual area throughout the study. In order to ensure that we examine the most recent AI advancements, the time span considered in the analysis is the period from 2009 to 2018. This means that only the activities that occurred during that period, and the involved agents, constitute the system under analysis.

5.1. AI landscape: AI agents' worldwide distribution

The detection and localisation of AI TES agents is the first output of the present study. Table 2 presents the percentages of AI agents per geographic area and organisation type. The US is the leading country in AI agents (30.0%), followed by China (25.5%), then the EU28 (18.7%), and two other Asian countries, namely India (which has a quarter of the agents of the EU28) and South Korea (one-sixth

[16] If a group is considered to contain duplicated agents, all agents but the one considered to be the "original" are deleted. Therefore, as each group can contain more than two agents, the final number of agents eliminated exceeds 1,809.

[17] The type "Firms" includes companies, firms, and start-ups. The type "Governmental Institutions" includes all public and local economic entities with administrative roles, public institutes not associated with research activities, and state-owned utility monopolies. The type "Research Institutes" includes universities, public and private research centres, and higher or secondary education organisations. Although the name of the categories "company gov non-profit" and "gov non-profit" would suggest that they are governmental institutions, they are in fact public research institutes. These two categories are both found in the patent data source.

[18] For instance, agents are considered as governmental institutions if their names contain: *agencija, agenzia, ayuntamiento, bundes, comune, feder, governmen, minist, ministry, municip, national, navy, provincia, region, regional, state, etc. Agents are considered as research institutes if their name contains: academy, center scientifique, central science, depart, ecole, fachhochschule, national lab, recherche, resear, school, univ, vyskumny, wissenschaften*, etc.

Table 1
AI 2009–2018 techno-economic complex System's descriptive statistics.

Types of Activities	Number of detected Activities[b]	Number of detected Agents
Industrial activities	16,731	16,731
Production, Services and Trade related to AI goods & services[a]		
R&D Activities (1)	29,247	10,873
AI-related patents		
R&D Activities (2)	4,719	1,856
AI-related research publications		
	Final Number of Activities[c]	Final Number of Agents[d]
	50,697	29,094

[a] The number of activities initially detected in the category Industrial Activities is 18,646. Due to the disambiguation of agents, some agents are associated to multiple industrial activities. However, as each agent can be associated to only one industrial activity, the multiple industrial activities associated to the same agent are merged.
[b] Each activity corresponds to one document, which contains the textual information of the associated activity.
[c] Sum of number of detected documents.
[d] The final number of agents is not equal to the sum of number of agents detected by type of activities, as the same agent may be involved in different types of activities.

Table 2
AI Agents by geographic area and organisation type.

	N. of Agents [a]	AI Presence Intensity[b]	N. of Firms [c]	N. of Governmental Institutions [c]	N. of Research Institutes [c]
US	30.0%	0.646	32.1%	1.9%	12.2%
China	25.5%	0.512	25.5%	87.0%	39.9%
EU28	18.7%	0.367	18.6%	1.4%	20.5%
India	4.3%	0.207	4.6%	0.9%	1.3%
Other Asian countries	3.5%	0.221	3.4%	1.4%	4.2%
South Korea	3.2%	0.705	3.0%	0.9%	5.1%
Other European countries	3.1%	0.281	2.8%	5.6%	5.9%
Canada	3.0%	0.730	3.1%	0.5%	2.3%
Japan	2.2%	0.168	2.2%	0.0%	2.5%
Middle East	2.1%	0.125	2.1%	0.0%	1.9%
Other American countries	2.0%	0.061	2.0%	0.0%	1.9%
Oceania	1.6%	0.439	1.6%	0.0%	1.5%
Africa	1.0%	0.289	1.0%	0.5%	0.9%
Total N. of Agents & Total N. of Agents by Type	29,094		26,158	215	2,721

[a] % over Total.
[b] Number of Agents over GDP PPS 2015.
[c] % over Total by Type.

of EU28 agents). Modestly following these five are players located in Canada and Japan. The worldwide AI landscape is therefore mainly dominated by the US, China, and the EU28, and there is a significant gap between the leading group and the followers.

The composition of the three major areas differs by organisation type. The US leads in firms, with one third of all firms detected worldwide (32.1%). It is remarkable that, although the overall presence of governmental institutions is relatively low (215 agents out of 29,094), China leads in this specific organisation type (87.0%). Although the EU28 ranks third in the number of firms, it has the second greatest presence in AI research institutes (20.5%) after China. These observations show that AI activities in the US are strongly oriented to business activities, an area in which the EU28 is less present. China's large number of agents in each of the three types suggests that its involvement in the AI domain is a structured and complex phenomenon including a variety of agents. In addition, the strong presence of governmental institutions seems to confirm the Chinese Government's strong commitment to sustaining the development of AI (China Institute for Science and Technology Policy at Tsinghua University, 2019). Compared to the US and China, the EU28 has a much greater presence of research institutes in the field, showing a very active research environment.

In addition, in Table 2, under the assumption that the number of agents is correlated with the size of a geographic area's economy, we define the indicator *AI presence intensity* as the number of agents over Gross Domestic Product (GDP). GDP is measured in EUR purchasing power standards (PPS) to account for differences in prices, allowing cross-country comparisons. This indicator allows the number of AI agents in a country in comparison to its GDP (base year 2015) to be evaluated. In other words, it tentatively measures countries' institutional and industrial efforts to become important agents in the AI landscape. In particular, it is notable that South Korea and Canada arise as the countries with the highest AI presence intensity.

This subsection has described the AI landscape with reference to the number of agents detected. Although this part of the work does not weight agents according to their size, large businesses with locations in different cities have a larger presence in the data than small businesses with a single location because we define a single economic entity over multiple locations as multiple agents. We are developing a method to account for more specific weights in further analyses.

5.2. Thematic subdomains in AI

In view of the absence of a standard classification of AI (as previously explained, see Footnote 1), the probabilistic topic model applied on the large collection of documents and described in Subsection 3.2 identified six non-exclusive thematic topics in the AI technological landscape for the period 2009–2018. These topics are *non-exclusive*, as they consist of technological terms that may be found in multiple topics. This is because a natural language term may belong to multiple topics. For example, the term "neural network" appears in multiple AI thematic topics, as neural network algorithms are used by multiple AI subdomains.

5.2.1. Identified AI thematic topics of industrial and R&D activities

Six topics are selected to represent thematic subdomains in the AI domain after implementing LDA and testing several numbers of topics to determine the number that better describes the technological domain. To further elaborate, the number of topics is determined by (i) inquiring the potential bounds of real-world subdomains in the corpus with a literature review of 55 publications about AI from research, policy, and market perspectives (exhaustive list in Appendix E), which results in the identification of a minimum of two and a maximum of seven separate AI subdomains, (ii) applying an LDA trial-and-error approach (Suominen & Toivanen, 2016; Zhao et al., 2015), with an input number of topics around the boundaries of the number of AI subdomains found in the literature,[19] (iii) evaluating each topic of each run for coherence with the terms' probability distribution of each topic,[20] and correspondence to real-world subdomains reported in the literature.[21]

After the identification of the number of topics, an automatic summary of the top 10 terms in the probability distribution of each topic provides the initial titles of the topics. These are revisited and shortened to allow the description of each topic in more natural language. The titles are finally evaluated by an in-house AI expert to form the final titles of the thematic subdomains of the inter-topic distance map presented in Fig. 4. The top-10 list of terms per topic is provided in Appendix D, for the assessment of the final titles' correspondence to the AI thematic subdomains. The AI thematic subdomains detected using our method are: (i) "speech recognition, natural language processing (NLP) & synthesis", (ii) "face & image recognition", (iii) "Platform & Software Services (PaaS & SaaS)", (iv) "theoretical methods", (v) "automation processes and robotics", and (vi) "connected & automated vehicles" (CAVs). More details about the description of these subdomains can be found in Appendix D.

5.2.2. Closeness of AI thematic subdomains

To explore the relevance of the AI thematic subdomains and their relationships, we use LDAvis, a system for visualising and interpreting topics estimated by LDA (Sievert & Shirley, 2014).[22] This visualisation system allows each topic to be explored separately with a comparison of the terms' frequency in a topic and in the corpus, rather than only providing insights regarding the corpus in the form of word clouds per topic or bar plots per document (Chaney & Blei, 2012; Gardner et al., 2010; Snyder, Knowles, Dredze, Gormley, & Wolfe, 2013).In Fig. 4, the result of this visualisation is adapted to illustrate both the topics resulting from the topic model and the titles of the corresponding AI subdomains over the entire study period in the AI industrial and R&D activities. The topics' areas are analogous to the topics' prevalence in the corpus. For example, the subdomain of AI "theoretical methods" is the largest; hence, there are more terms associated with this subdomain than with the others. This implies that there are more activities mentioning AI theoretical methods than activities mentioning other subdomains, which is to be expected, as all the other subdomains adopt AI methodological terms to describe the development of their corresponding activities.

All the illustrated AI thematic knowledge subdomains contain applications and methods from interdisciplinary domains, which explains the closeness in distance in certain cases. For topics that are depicted as close in distance, more operations and terms are used interchangeably than for topics depicted at greater distances. We interpret that the thematic subdomains "speech recognition, NLP & synthesis", "face & image recognition", and "connected & automated vehicles" are found close because speech and face recognition applications are used in CAVs from the automobile industry. On the contrary, the greater distance between the operations ensuing from the thematic subdomains "Platform & Software Services (PaaS & SaaS)" and the CAVs subdomain can be understood by examining the applications and theoretical advancements of the two. For example, decision management applications for profit increase are not related to CAVs decision management. AI "theoretical methods" is distant from the other subdomains on the PC2 axis, which can be

[19] The number of topics given as input to the LDA is a discrete set of values with the lower bound set to three (as the distinction of only two AI subdomains is deemed as too limited to describe the variance of subdomains in the technology), and the upper bound set to ten (so as to explore the knowledge space in the collected documents with a larger value than the number of AI subdomains found in the literature, i.e., seven, according to the literature in Appendix E. Hence, LDA is run eight times with the input of each of the values of this range.

[20] For each run, the 50 most frequent terms in the term probability distribution of each topic are examined for coherence. The run for three to six numbers of topics are found to be thematically coherent.

[21] The topics' content of each run for three to six topics are assessed for correspondence to the thematic subdomains of AI in the studied literature. We find that 24 of the 55 analysed documents corroborate the existence of six topics, as the six identified topics adequately represented the AI subdomains reported in these documents. Therefore, the topic modelling of six topics is selected to proceed for further analysis.

[22] The topics' centres are plotted in a two-dimensional plane through the Jensen-Shannon divergence and PCA. Jensen-Shannon divergence is used in probability theory to measure the similarity between probability distributions. As topics are probability distributions, the visualisation system that we applied (LDAvis system (Sievert & Shirley, 2014)) computes inter-topic differences to provide inter-topic distances. The topics' area is proportional to the frequency at which their contained terms are found in the set of documents. The inter-topics' distance is proportional to the relations between topics. Therefore, a visualisation of the topics would require a multi-dimensional space. Principal Components Analysis is used to project inter-topic distances, as well as the frequency of a topic illustrated proportional to its area.

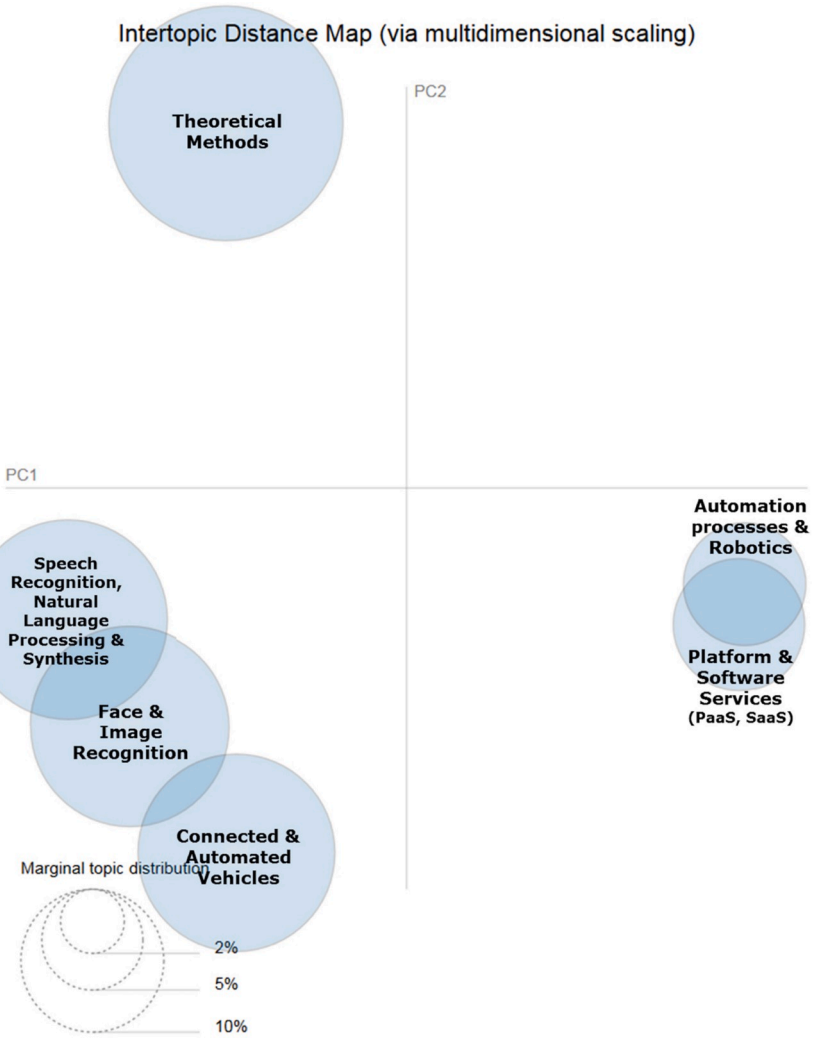


Fig. 4. Inter-topic distance map of AI subdomains for collected industrial and R&D AI activities (2009–2018). The nodes represent the detected topics, and their area is proportional to the frequency at which their contained terms are found in the set of documents. The inter-topics' distance is proportional to the thematic relations between topics. The two axes represent the two dimensions that the PCA identified as capturing the largest amount of variance in the data. Note: For the visualisation the *LDAvis R package* is used. The map is interactive, a static representation is provided due to this medium.

interpreted as a distinction between theoretical developments and applications.

5.2.3. Geographic distribution of AI thematic subdomains

The most active geographic areas in each of the six identified AI thematic subdomains are analysed in this subsection. In this scope, Fig. 5 illustrates the activities in geographic areas per topic, including industrial and R&D activities. All six thematic subdomains are led by China or the US. The EU28 holds a strong position in two subdomains: "automation processes and robotics" and "Platform & Software Services (PaaS & SaaS)". South Korea ranks third in "speech recognition, NLP & synthesis" and "connected and automated vehicles". The main leaders for each thematic subdomain are discussed below.

In "speech recognition, NLP & synthesis" (Fig. 5a), the US and China lead with a share of almost 30% of worldwide activities each. South Korea follows with one-third fewer activities, ranking third. Japan and the EU28 are moderately active in this subdomain. In "face & image recognition" (Fig. 5b), China has a strong presence with two-thirds of activities worldwide in the thematic subdomain. The US's involvement in this subdomain is relatively moderate. South Korea again ranks third in this thematic subdomain, with Japan and the EU28 following closely. In "connected and automated vehicles" (Fig. 5c), China is strikingly active, with a concentration of more than half of worldwide activities in the thematic subdomain. The US follows with considerably lower activities. The EU28 and South Korea are also moderately present, with half of the US's number of activities. Japan and other Asian, European, and American countries are present due to their firms' activities in the thematic subdomain (further explained in Fig. 6c). In the subdomain of "automation processes and robotics" (Fig. 5d), the majority of activities are located in the US. The EU28 is in a competitive position,

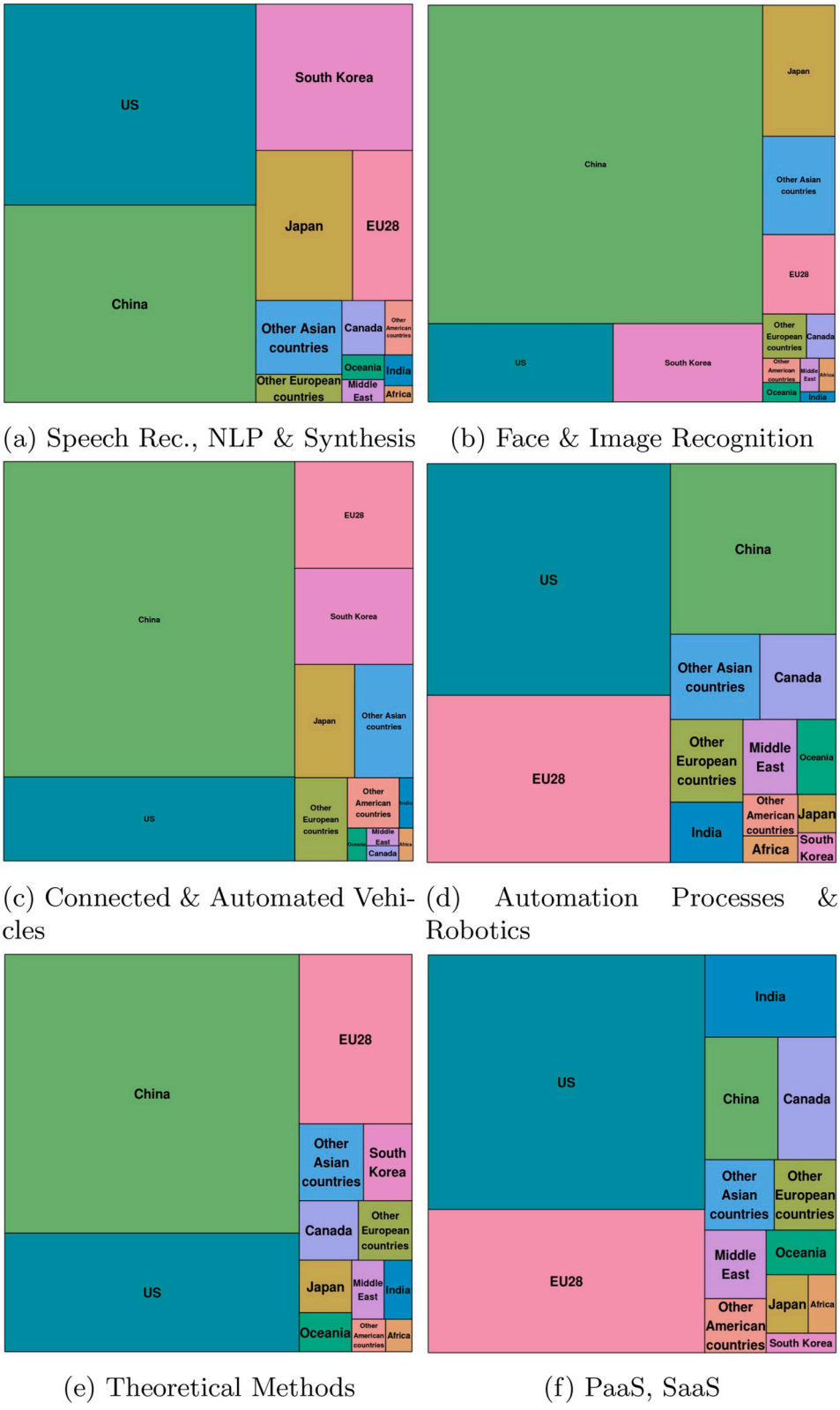


Fig. 5. Geographic distribution of AI activities in each thematic subdomain, 2009–2018.

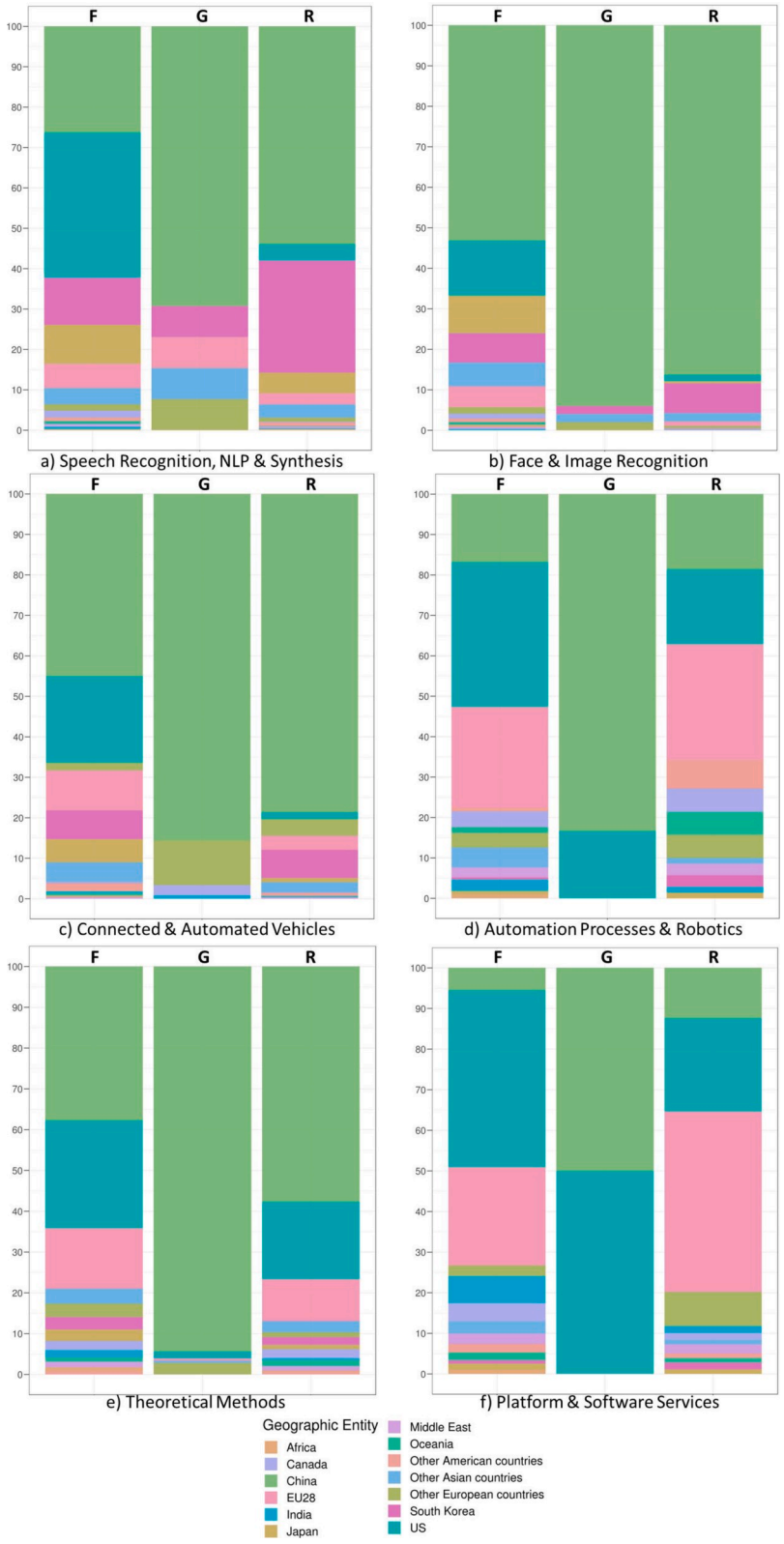
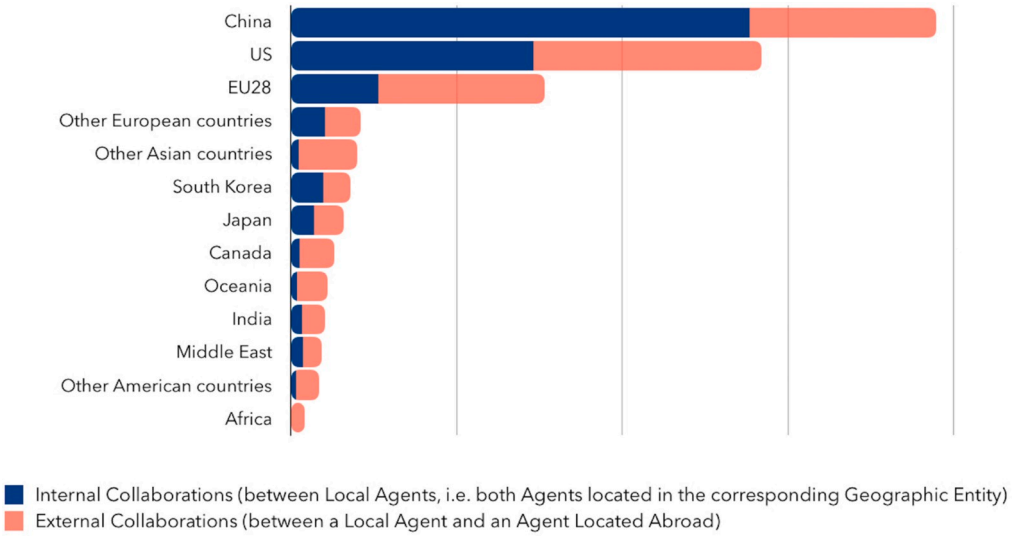
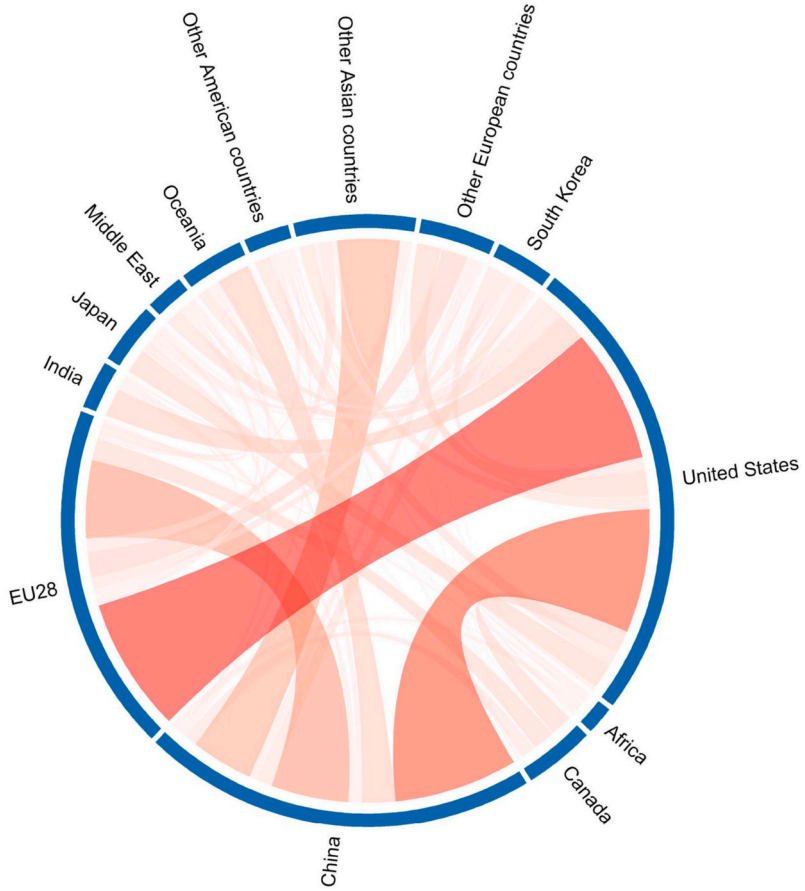


Fig. 6. Geographic distribution (in %) of AI activities per thematic subdomain and organisation type of the agents involved, 2009–2018. ”F” is the abbreviation for ”Firms”, ”G” for ”Governmental Institutions” and ”R” for ”Research Institutes”.



(a)



(b)

Fig. 7. **(a)** Internal vs. external collaborations of agents. **(b)** Collaborations of agents between geographic areas.

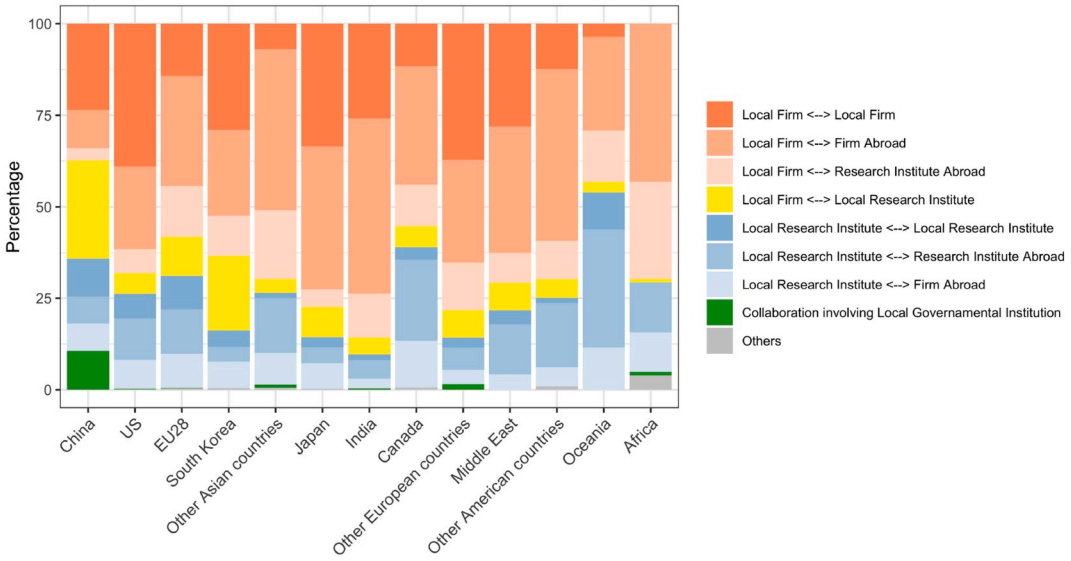


Fig. 8. Collaborations by geographic area and type of the involved peers. With "Local" we refer to agents that belong to the corresponding geographic area, while with "Abroad" we refer to agents that do not belong to the corresponding geographic area. Geographic areas are in decreasing order (from left to right) of R&D activities in which their agents are involved.

ranking second with 25% of the worldwide activities in the topic. China is also present, though moderately active in the subdomain with one-fifth of worldwide activities. Other countries and the most active geographic areas in this thematic subdomain are India and other Asian countries, as well as Canada and other European countries, which follow with a more modest number of activities. In the subdomain "theoretical methods" (Fig. 5e), which represents the "know-why" of AI technology as it concerns methodological knowledge, half of the agents developing related activities are located in China, with fewer activities in the US and the EU28. Activities located in South Korea or other Asian countries have a moderate presence. The thematic subdomain of "Platform & Software Services (PaaS & SaaS)" (Fig. 5f) captures the US's interest with 43% of the worldwide activities in this subdomain. The EU28's focus is also very close, with one-fourth of worldwide activities in the subdomain. India ranks third, with less than one-tenth of the worldwide activities in the thematic subdomain, closely followed by Chinese and Canadian activities.

5.2.4. AI thematic profile of organisation types worldwide

Fig. 6 illustrates, the geographical distribution of the AI activities developed by agents according to their organisation type for each of the identified thematic subdomains. The most compelling results are the following.

The majority of activities performed by governmental institutions correspond to agents located in China, a pattern that holds for every identified subdomain. Only in the thematic subdomain "Platform & Software Services (PaaS & SaaS)" (Fig. 6f) is there a fair representation of US governmental activities, and a more modest one in the subdomain "automation processes and robotics" (Fig. 6d). In all thematic subdomains, the share of the US is higher for firms' activities than for research institutions' activities (Fig. 6).

In the subdomains of "automation processes and robotics" (Fig. 6d) and "Platform & Software Services (PaaS & SaaS)" (Fig. 6f), a significant percentage of the total worldwide research-led activities is located in the EU28. In these subdomains, the EU28 also has a strong presence from firms, with approximately 25% of the worldwide firm activities in the corresponding subdomains. A moderate part of firms' activities in the EU28 are "know-why"-oriented, as they are identified in the "theoretical methods" subdomain (Fig. 6e). Firms' activities on "connected & automated vehicles" cover a modest part of the activities worldwide (Fig. 6c).

In the subdomain "speech recognition, NLP & synthesis" (Fig. 6a), South Korea holds approximately 30% of research-led activities worldwide, with a moderate presence in firms' activities and governmental activities. In the subdomains of "face & image recognition" (Fig. 6b) and "connected & automated vehicles" (Fig. 6c), South Korea is represented almost equally in research and firms' activities, with a share of nearly 8%. Japanese firms are modestly present among firms' activities worldwide in the thematic subdomains of "speech recognition, NLP & synthesis" (Fig. 6a), "face & image recognition" (Fig. 6b), and modestly in the "connected & automated vehicles" (Fig. 6c) subdomain. Among all thematic subdomains, Indian firms appear to have a modest participation in worldwide firm activities in "Platform & Software Services (PaaS & SaaS)" (Fig. 6f).

5.3. Network of AI technological collaborations

The objective of this section is to investigate agents' interactions oriented to the development of AI-related technologies. Therefore, two types of R&D activities are considered: (i) patent applications and (ii) publications in top AI conferences. As described in the methodological Subsection 3.3, a worldwide network is generated from the collected data in which connections represent technological collaborations between agents of different geographic areas, and in which self-loops represent collaborations between agents of

the same area (Fig. 3d).

5.3.1. Collaborations within and between geographic areas

The distinction between internal and external collaborations is the first important contribution of this work with regard to the investigation of technological collaborations of AI agents worldwide. Fig. 7.a presents the amount of internal and external collaborations by area. The three leading areas, China, the US, and the EU28, show a composition that is in line with that discussed in Subsection 5.1. Though China does not have the largest number of agents, it is the most active area, presenting significantly more internal than external collaborations..[23] Then, while the activities of the US are balanced between internal and external collaborations, those of the EU28 show a large propensity for external collaborations. Therefore, the three most important areas –the US, China, and the EU28 –present different balances between internal and external collaborations. Though we do not address the implications related to this finding in this work, it will be the object of future investigations.

In addition to this distinction, external collaborations between geographic areas concerning AI activities are presented in the chord diagram (Fig. 7b), where the width and colour intensity of each chord is proportional to the number of collaborations between the two geographic areas involved. The most intense technological connections, in decreasing order, are observed between: the EU28 and US, which represent almost half of the external collaborations of the EU28, and almost one-third of the external collaborations of the US; China and the US, which correspond to almost one-third of all external collaborations for both areas; the EU28 and China, representing almost one-fourth of all external collaborations for both areas; other Asian Countries and China, corresponding to almost more than one-half of all the external collaborations of the first area and more than one-eighth of the external collaborations of China.

Therefore, the strongest connection observed in the AI techno-economic system is that between the EU28 and the US. Nevertheless, the role of China should not be considered as secondary. China is always present in all the top connections apart from the first one. Considering the number of internal collaborations, as presented in Fig. 7.a, China appears as a leading force that is very well-connected with the rest of the world and has intense internal activities.

5.3.2. Collaborations by type and location of the peers

When further the organisation types of the agents involved in the connections are considered, different structures emerge in the considered areas. Fig. 8 shows the percentage distribution of collaborations by location and type of agents involved. One of the most relevant elements is the outstanding percentage of Chinese collaborations involving local firms and local research institutes (in yellow). Again, China is the only area in which there is a significant presence of collaborations involving governmental institutions (in green). The US's structured and established AI industry is shown through the large percentage of collaborations between local firms (dark orange). The EU28 is rather weak in this collaboration type. On the other hand, the collaborative activities of the EU28's research institutions are good but not outstanding. The sum of the blue stacks corresponding to the EU28 results in a good amount of research collaborations, but the difference from the other leading areas is not remarkable. What is less expected is the large number of collaborations between EU28 local firms and firms abroad (second shade of orange). Regarding the other areas, South Korea presents a large percentage of internal collaborations between firms and research institutes, and Japan and India present similar profiles, showing large relative amounts of collaborations between local firms, and between local firms and firms abroad (dark orange and second shade of orange, respectively).

5.3.3. R&D network dynamics

In Fig. 9, from left to right, the following statistics are plotted for three of the most relevant areas. In the first column, the number of agents (by organisation type, in different black lines) involved in R&D activities and the number of R&D activities per type (patent applications in red and research publications in green) are reported. In the second column, the number of internal (within the considered geographic area) and external collaborations (with different geographic areas) (blue and red line, respectively) are shown. In the third column, the weighted betweenness centrality (WBC), standardised in range $[0, 1]$, for the whole network of technological collaborations and for its two subnetworks (in red for patent applications, and in green for research collaborations) is represented. This representation does not include data for 2018 due to a delay in the registration of patent filings. The consideration of this year would have led to erroneous conclusions, as a significant decrease would have been observed for patenting activity in 2018. Though the data of 2017 are also affected by this delay, the trend is not significantly affected. Finally, the values of the y-axis in columns 1 and 2 of Fig. 9 are separately scaled (for each country), in order to better evaluate changes in the country trends.

This final computation shows that, in general, the upward trend in the number of agents involved in AI R&D activities and in the number of AI R&D activities (column 1 in Fig. 9) started around 2012. In this year, the following elements can be detected: (i) the beginning of a rapid increase in patenting activities in China (red line); (ii) an increase in the number of firms involved in AI R&D collaborations and an increase in the number of publications in the EU28 (black dotted line with filled circles and green line, respectively); and (iii) a sharp increase in the number of firms and in patenting activities in the US (black dotted line with filled circles and red line, respectively).

Regarding the trends of internal and external collaborations (Fig. 9 column 2), China started to increase its internal collaborations in 2012 (blue line), and this dynamic was followed by greater external collaborations (red line) with a delay of three years. While in Fig. 7.a it is possible to observe that over the whole considered period, Chinese agents collaborate more internally than externally,

[23] This also reflects the different patenting structure put in practice by Asian countries in comparison to Western countries (WIPO, 2019).

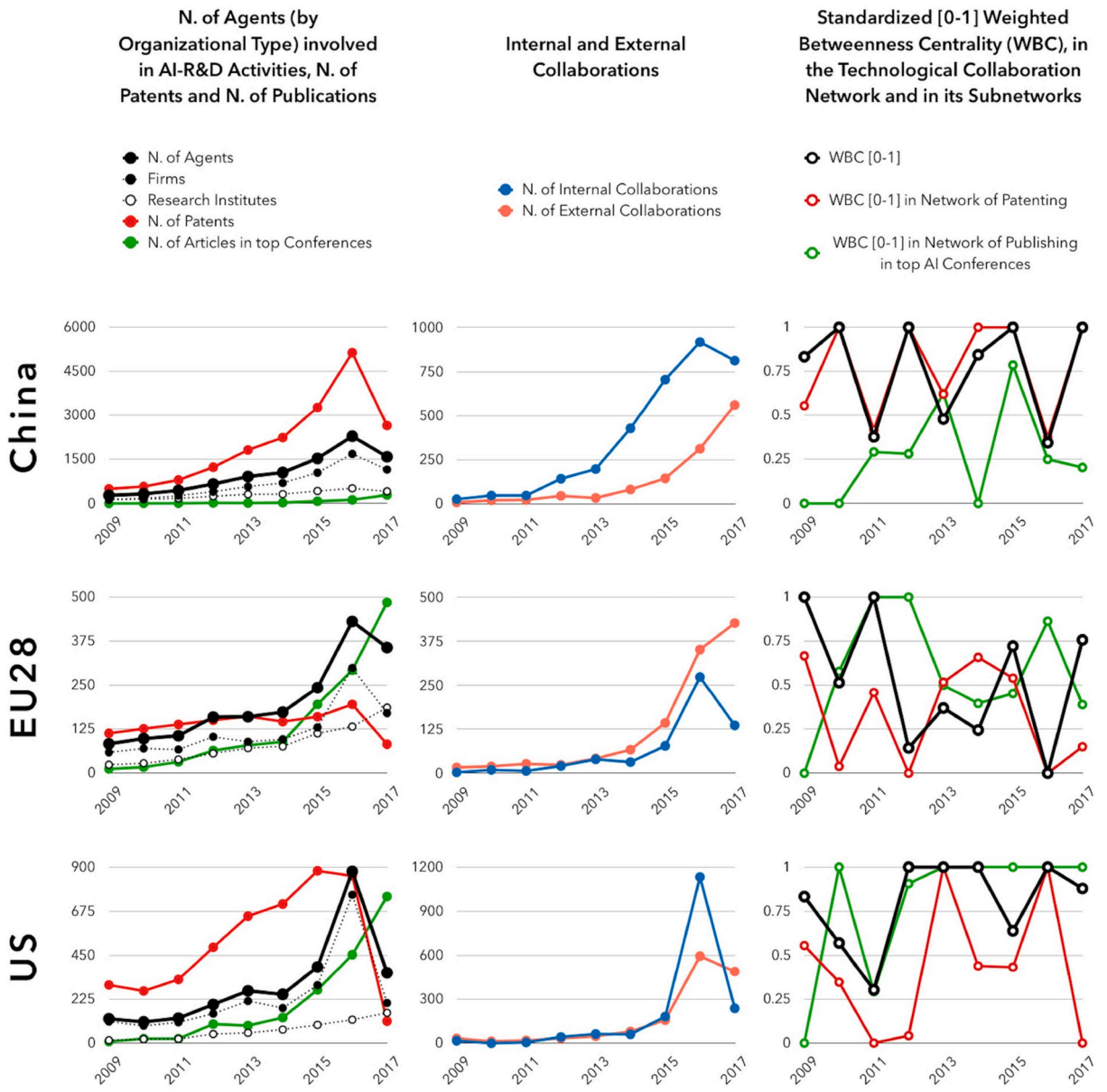


Fig. 9. Statistics regarding AI R&D network dynamics for China, the EU28 and the US, from 2009 to 2017. Note: 2018 is excluded because of the delay in patents' filing registration.

Fig. 9 (column 2) shows that this has been the case in each of the considered years. In the US, the number of external collaborations follows a similar trend to the internal ones, even if in 2016, the number of internal collaborations was double that of external collaborations. Though the EU28 has taken a similar path, it is noteworthy that the region's number of external collaborations (i.e., between an EU28 agent and an agent located out of the EU28) almost always exceeds the number of internal collaborations (i.e., between EU28 agents).

Information about the relevance of the considered areas in the global AI network (and subnetworks) of technological collaborations is provided (Fig. 9 column 3). China's overall centrality (black line), and hence its capacity to influence the rest of the network, fluctuates over time with downward trends in 2011, 2013, and 2016. This trend seems to be mainly determined by China's centrality in the patent applications' subnetwork (red line), which presents the same dynamic. The centrality in the subnetwork of conference publications (green line) increases over time, apart from in 2014, but remains lower than that of the patenting applications subnetwork (red line). The EU28 shows fluctuating trends in the patent applications and conference publications subnetworks. The latter presents larger values throughout time. This suggests that conference publications contribute more significantly than patenting applications to the maintenance of the EU28's significant position in the network of R&D collaborations (black line). In the subnetwork of patent applications, the US's centrality (red line) is higher than the corresponding centrality of the EU28, but lower than that of China.

Nevertheless, the US's centrality throughout the period (black line) is very high. The major contribution to the US's overall centrality is the country's centrality in the subnetwork of conference publications (green line), which is the largest observed for almost all the studied periods.

In summary, the number of Chinese patents (red line in the first column of graphs) significantly increased over time, reaching more than 4500 applications in 2016. The centrality of China (red line, column 3) in this subnetwork is considerable but inconsistent. The US and the EU28 do not have the same trends and scale for patent applications, as the number of patents they develop is much more limited than that of China (red line, column 1). The performance of the US and EU28 in terms of strategic position in this subnetwork is secondary, but remarkable (red lines, column 3). This is probably a consequence of different partnership strategies, namely more openness from the US and the EU28 to collaborate with agents from external areas, and for China, more collaborations within national borders. Both the EU28 and the US perform better than China concerning conference publications. The large number of publications developed by EU28 and US agents (green line, column 1) and their more open collaborative structure (orange line, column 2) allow the two geographic areas to lead in terms of centrality in conference publications (green line, column 3). Finally, the combined analysis of the two R&D subnetworks (black line, column 3) reveals alternate leaderships over time, with only one year presenting two leading geographic areas (in 2012, China and the US) in terms of overall centrality. EU28 closely follows the US and China, and outranks them in 2011. In addition, of the three geographic areas, it is the only one with a higher number of external than internal collaborations (column 2). However, leadership is contended between the US and China. The former fluctuates less and is more steadily located in the most central position (especially from 2012 to 2017). The latter has a less constant trend, but it is able to re-affirm its primary role after a partial decrease, as demonstrated by the peaks in 2012, 2015, and 2017, which were all preceded by at least one year of weighted betweenness centrality (WBC) lower than 0.5.

6. Conclusions and perspectives

Artificial intelligence is undergoing a period of intense progress, and its increasing pervasiveness is expected to have disruptive impacts on economies and societies. Thus, AI is attracting the attention of a growing number of governments. However, like other transversal and dynamically evolving technologies, there are still no mutually agreed-upon systems for classifying or measuring AI. In this work, we investigate and measure the AI techno-economic segment from the perspective of an *agent-artifact space* by collecting and analysing multiple micro-data sources. We use a structured and non-heuristic methodology that can be also implemented for the study of technologies other than AI. This methodology is used in the present work to uncover elements regarding (i) the agents participating in this evolving complex system, (ii) the technological subdomains of AI, and (iii) the structure and the dynamic of collaborations in patents and publications. All of these points have been analysed from the perspective of the location of the agents involved and their organisation type. Although this work does not make specific policy recommendations, it provides essential information for the discussion and design of future policy interventions. Appropriate knowledge of the shape of the system under observation can enable more effective and efficient measures to be developed.

Future work that builds on this research should first introduce new elements for consideration, such as business structures, the connection of the different branches of multinational companies, and information regarding revenues and capital stocks. These elements are expected to contribute to the enrichment of the analysis. Nonetheless, the described methodological framework allows the fundamental aspects of the considered complex system to be mapped. Second, the use of additional statistical analyses, e.g., predictive analyses, would help uncover the emergence of relevant aspects in the techno-economic landscape in future research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.telpol.2020.101943.

References

Abbasi, A., Hossain, L., & Leydesdorff, L. (2012). Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. *Journal of Informetrics, 6*(3), 403–412.
Amendola, M., & Gaffard, J.-L. (1988). *The innovative choice. An economic analysis of the dynamics of technology*. Tech. Rep.
Amendola, M., & Gaffard, J.-L. (1998). *Out of equilibrium*. Tech. Rep.
Arthur, B. (1999). Complexity and the economy. *Science, 284*(5411), 107–109.
Arthur, B. (2007). The structure of invention. *Research Policy, 36*(2), 274–287.
Arthur, B. (2009). *The nature of technology: What it is and how it evolves*. Simon and Schuster.
Arthur, B., Durlauf, S., & Lane, D. (1997). *The economy as an evolving, complex system II* (Vol. 24, pp. 59–61). Reading, MA: Addison-Wesley.
Balland, P.-A. (2012). Proximity and the evolution of collaboration networks: Evidence from research and development projects within the global navigation satellite system (GNSS) industry. *Regional Studies, 46*(6), 741–756.
Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84.
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.
Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology, 37*(1), 179–255.
Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics, 64*(3), 351–374.
Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology, 25*(2), 163–177.
Breitzman, A., & Thomas, P. (2015). The emerging clusters model: A tool for identifying emerging technologies across multiple patent systems. *Research Policy, 44*(1), 195–205.

Breschi, S., & Malerba, F. (1997). Sectoral innovation systems: Technological regimes, schumpeterian dynamics, and spatial boundaries. *Systems of innovation: Technologies, institutions and organizations*, 130–156.

Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlström, P., … Trench, M. (2017). *Artificial intelligence: The next digital frontier*. McK- insey Global Institute.

Chaney, A. J.-B., & Blei, D. M. (2012). Visualizing topic models. In *Sixth international AAAI conference on weblogs and social media*.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (pp. 288–296).

China Institute for Science and Technology Policy at Tsinghua University. (2019). *Development report* (Vol. 2, p. 15), 2018.

Cho, T.-S., & Shih, H.-Y. (2011). Patent citation network analysis of core and emerging technologies in Taiwan: 1997–2008. *Scientometrics, 89*(3), 795–811.

Craglia, M., Annoni, A., Benczur, P., Bertoldi, P., Delipetrev, P., De Prato, G., … Vesnic Alujevic, L. (2018). *Artificial intelligence: A European perspective*. European Commission: Tech. Rep. https://doi.org/10.2760/11251.

Cronen, V., & Pearce, B. (1980). *Communication, action and meaning: The creation of social realities*. New York: Praeger Publishers.

De Prato, G., Lopez-Cobo, M., Samoili, S., Righi, R., Miguel, V.-P. B., & Cardona, M. (2019). *The AI techno-economic segment analysis. Selected indicators (tech. Rep.)*. European Commission: Joint Research Centre.

Ding, Y. (2011). Community detection: Topological vs. topical. *Journal of Informetrics, 5*(4), 498–514.

Dosi, G. (1982). Technological paradigms and technological trajectories: A suggested interpretation of the determinants and directions of technical change. *Research Policy, 11*(3), 147–162.

European Commission. (2018). *Artificial intelligence for Europe - COM* (p. 237), 2018 ec.europa.eu/transparency/regdoc/rep/1/2018/en/com-2018-237-f1- en-main-part-1.pdf. (Accessed 25 April 2018). Online.

European Commission. (2018). *Coordinated plan on artificial intelligence - COM* (p. 795), 2018 ec.europa.eu/transparency/regdoc/rep/1/2018/en/com-2018-795-f1-en-main-part-1.pdf. Online; 7 December 2018.

Etzkowitz, H., & Leydesdorff, L. (2000). The dynamics of innovation: From national systems and "mode 2" to a triple helix of university–industry– government relations. *Research Policy, 29*(2), 109–123.

Fagiolo, G., Reyes, J., & Schiavo, S. (2008). On the topological properties of the world trade web: A weighted network analysis. *Physica A: Statistical Mechanics and Its Applications, 387*(15), 3868–3873.

Fagiolo, G., Reyes, J., & Schiavo, S. (2009). World-trade web: Topological properties, dynamics, and evolution. *Physical Review E, 79*(3), 036115.

Fagiolo, G., Reyes, J., & Schiavo, S. (2010). The evolution of the world trade web: A weighted-network analysis. *Journal of Evolutionary Economics, 20*(4), 479–514.

Frank, M. R., Wang, D., Cebrian, M., & Rahwan, I. (2019). The evolution of citation graphs in artificial intelligence research. *Nature Machine Intelligence, 1*(2), 79.

Freeman, L. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35–41.

Freeman, L. (1978). Centrality in social networks conceptual clarification. *Social Networks, 1*(3), 215–239.

Freeman, C. (1991). Networks of innovators: A synthesis of research issues. *Research Policy, 20*(5), 499–514.

Gardner, M. J., Lutes, J., Lund, J., Hansen, J., Walker, D., Ringger, E., et al. (2010). The topic browser: An interactive tool for browsing topic models. In *Nips workshop on challenges of data visualization* (Vol. 2).

Garlaschelli, D., & Loffredo, M. I. (2005). Structure and evolution of the world trade network. *Physica A: Statistical Mechanics and Its Applications, 355*(1), 138–144.

Georgescu-Roegen, N. (1971). *The entropy law and the economic problem*. Harvard: Harvard University.

Gerken, J. M., & Moehrle, M. G. (2012). A new instrument for technology monitoring: Novelty in patents measured by semantic patent analysis. *Scientometrics, 91*(3), 645–670.

Gilsing, V., Nooteboom, B., Vanhaverbeke, W., Duysters, G., & van den Oord, A. (2008). Network embeddedness and the exploration of novel technologies: Technological distance, betweenness centrality and density. *Research Policy, 37*(10), 1717–1731.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(suppl 1), 5228–5235.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211.

Hacklin, F. (2007). *Management of convergence in innovation: Strategies and capabilities for value creation beyond blurring industry boundaries*. Springer Science & Business Media.

Hicks, J. (1973). *Capital and time: A neo-Austrian theory*. Oxford University Press.

Hu, Z., Fang, S., & Liang, T. (2014). Empirical study of constructing a knowledge organization system of patent documents using topic modeling. *Scientometrics, 100*(3), 787–799.

Ibarra, H. (1993). Network centrality, power, and innovation involvement: Determinants of technical and administrative roles. *Academy of Management Journal, 36*(3), 471–501.

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing*.

Lane, D. (2011). Complexity and innovation dynamics. In *Handbook on the economic complexity of technological change* (Vol. 63).

Lane, D. (2016). Innovation cascades: Artefacts, organization and attributions. *Philosophical Transactions of the Royal Society B, 371*(1690), 20150194.

Lane, D., & Maxfield, R. (1997). Foresight, complexity, and strategy. In *The economy as an evolving complex system II* (pp. 169–198).

Lane, D., & Maxfield, R. (2005). Ontological uncertainty and innovation. *Journal of Evolutionary Economics, 15*(1), 3–50.

Lane, D., Maxfield, R., Read, D., & van der Leeuw, S. (2009a). From population to organization thinking. In *Complexity perspectives in innovation and social change* (pp. 11–42). Springer.

Lane, D., Pumain, D., van der Leeuw, S., & West, G. (2009b). *Complexity perspectives in innovation and social change* (Vol. 7). Springer Science & Business Media.

Lane, D., van der Leeuw, S., Sigaloff, C., & Addarii, F. (2011). Innovation, sustainability and ICT. *Procedia Computer Science, 7*, 83–87.

Lee, W. S., Han, E. J., & Sohn, S. Y. (2015b). Predicting the pattern of technology convergence using big-data technology on large-scale triadic patents. *Technological Forecasting and Social Change, 100*, 317–329.

Lee, C., Kang, B., & Shin, J. (2015a). Novelty-focused patent mapping for technology opportunity analysis. *Technological Forecasting and Social Change, 90*, 355–365.

Lee, C., Kwon, O., Kim, M., & Kwon, D. (2018). Early identification of emerging technologies: A machine learning approach using multiple patent indicators. *Technological Forecasting and Social Change, 127*, 291–303.

van der Leeuw, S. E. (2008). Agency, networks, past and future. In *Material agency* (pp. 217–247). Springer.

van der Leeuw, S. E., & McGlade, J. (1997). *Time, process, and structured transformation in archaeology* (Vol. 26). Psychology Press.

Leopold, E., May, M., & Paaß, G. (2004). Data mining and text mining for science & technology research. In *Handbook of quantitative science and technology research* (pp. 187–213). Springer.

Lundvall, B.-A., Dosi, G., & Freeman, C. (1988). *Innovation as an interactive process: From user-producer interaction to the national system of innovation*, 1988.

Mowery, D. C., & Teece, D. J. (1996). *Strategic alliances and industrial research*. Engines of innovation: US industrial research at the end of an era.

National Academies of Sciences Engineering and Medicine and others. (2017). The fourth industrial revolution. In *Proceedings of a workshop—in brief*.

Nelson, R. R. (1993). *National innovation systems: A comparative analysis*. Oxford university press.

Newman, M. E. (2011). *Complex systems: A survey*. arXiv preprint arXiv:1112.1440.

Nilsson, N. J. (2014). *Principles of artificial intelligence*. Morgan Kaufmann.

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks, 32*(3), 245–251.

Owen-Smith, J., & Powell, W. W. (2004). Knowledge networks as channels and conduits: The effects of spillovers in the Boston biotechnology community. *Organization Science, 15*(1), 5–21.

Powell, W. W., Grodal, S., et al. (2005). *Networks of innovators* (Vol. 78). The Oxford handbook of innovation.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach*. Malaysia: Pearson Education Limited.

Samoili, S., Righi, R., Lopez-Cobo, M., Cardona, M., & De Prato, G. (2018). Unveiling latent relations in the photonics techno-economic complex system. *Communications in Computer and Information Science*, 900.

Saxenian, A. (1994). *Regional networks: Industrial adaptation in silicon valley and route 128*.
Schwab, K. (2017). *The fourth industrial revolution* (UK).
Serrano, M. A., & Boguñá, M. (2003). Topology of the world trade web. *Physical Review E, 68*(1), 015101.
Sievert, C., & Shirley, K. (2014). Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63–70).
Snyder, J., Knowles, R., Dredze, M., Gormley, M., & Wolfe, T. (2013). Topic models and metadata for visualizing text corpora. In *Proceedings of the 2013 NAACL HLT demonstration session* (pp. 5–9).
Solow, R. M. (1957). Technical change and the aggregate production function. *The Review of Economics and Statistics*, 312–320.
Spitters, M., Verbruggen, S., & van Staalduinen, M. (2014). Towards a comprehensive insight into the thematic organization of the tor hidden services. In *2014 ieee joint intelligence and security informatics conference* (pp. 220–223).
Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis, 427*(7), 424–440.
Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology, 67*(10), 2464–2476.
Suominen, A., Toivanen, H., & Seppänen, M. (2017). Firms' knowledge profiles: Mapping patent data with unsupervised learning. *Technological Forecasting and Social Change, 115*, 131–142.
Venugopalan, S., & Rai, V. (2015). Topic based classification and pattern identification in patents. *Technological Forecasting and Social Change, 94*, 236–250.
Verhoeven, D., Bakker, J., & Veugelers, R. (2016). Measuring technological novelty with patent-based indicators. *Research Policy, 45*(3), 707–723.
Wasserman, S., Faust, K., et al. (1994). *Social network analysis: Methods and applications* (Vol. 8). Cambridge university press.
Wei, X., & Croft, W. B. (2006). LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international acm sigir conference on research and development in information retrieval* (pp. 178–185).
WIPO. (2019). *WIPO technology trends 2019: Artificial intelligence*.
Young, O., Berkhout, F., Gallopin, G., Janssen, M., Ostrom, E., & van der Leeuw, S. (2006). The globalization of socio-ecological systems: An agenda for scientific research. *Global Environmental Change, 16*(3), 304–316.
Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., & Newman, N. C. (2014). "Term clumping" for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change, 85*, 26–39.
Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., et al. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics, 16*, S8.
Zhao, W., Zou, W., & Chen, J. J. (2014). Topic modeling for cluster analysis of large biological and medical datasets. *BMC Bioinformatics, 15*, S11.