13
THE SCIENCE SPACE OF ARTIFICIAL INTELLIGENCE KNOWLEDGE PRODUCTION

Global and regional patterns, 1990–2016

Dieter F. Kogler, Adam Whittle, and Bernardo Buarque

Introduction

How will automation affect human well-being? This question arguably dates back as far as the first industrial revolution with the invention of the steam engine, and beyond. It is also a question that has more recently recaptured the attention of social scientists following the advent of Industry 4.0, and in particular the emergence of Artificial Intelligence (AI). Most noticeably from an economics perspective, several recent publications have sought to evaluate how AI will transform labour productivity (Brynjolfsson et al. 2018), economic growth (Aghion et al. 2017), international trade (Goldfarb and Trefler 2018), and employment (Arntz et al. 2016; Agrawal et al. 2019).

To put it succinctly, while AI has the potential to result in economic growth, prosperity, and positive change, it could just as easily produce job displacement and income inequality (Korinek and Stiglitz 2017). That is, despite the excitement over its many promises, the recent debate regarding AI seems to focus on the future of work “in a world in which computer algorithms can perform many of the functions that a human can” (Furman and Seamans 2019: 161). Indeed, comparable to the emergence of any other disruptive general-purpose technology (GPT), the rise of AI will have a profound impact on our daily lives and well-being. As Buarque et al. (2020: 175) state, it will eventually “lead to significant shifts in employment and income distributions across and within society, particularly when the gains are concentrated in AI-producing regions or sectors”.

Irrespective of whether your point of view of AI is optimistic or pessimistic, it is unquestionable that policies enacted today will shape how AI impacts society tomorrow (Agrawal et al. 2019). As interests around AI have begun to increase, countries around the world have started developing their own AI programmes with an eye on becoming a market leader in this disruptive technology (Dutton 2018). Equally, scholars in world-class institutions are also actively working to examine, propose, and implement policies that can enhance the benefits offered by AI, while mitigating against any of its negative consequences. Instrumental in this arena was a conference organised by the OECD in 2017, “AI: Intelligence Machines, Smart Policies”, with the sole aim of mobilising social, economic, and political responses to the transformation of society brought on by the advent of AI technologies. Therefore, we are not only observing a
mere surge in AI advancement, but also rising political concern about how to respond to it, and moreover how it should be managed. However, before it is possible to design fit-for-purpose policies, we must first understand, in an in-depth manner, the evolution and diffusion of AI systems as well as their many socio-economic consequences that are just now unfolding.

It is strikingly obvious that AI has become a hot topic and a frequently used buzzword. Nevertheless, despite its growing popularity, one dimension which remains unclear is how we can accurately measure the creation and diffusion of AI. In fact, whilst there seems to be a consensus that AI will transform our daily lives, it remains to be seen how these transformations will manifest in space, that is, through economic growth or productivity. As a consequence, AI risks becoming a policy ahead of the theory initiative, based primarily on speculative analysis and anecdotal evidence. Moreover, to the best of our knowledge, the relevant literature continues to even lack a precise definition of AI. Therefore, it is critical for AI’s successful implementation that there is an accurate depiction of its creation and development. Further still, to produce reliable inferences about how AI impacts our economy and society, we also need to develop robust data on its spatial and temporal diffusion; otherwise, our understanding of AI will remain speculative at best.

Against this backdrop, the present handbook chapter is a structured attempt to inform AI discourse and provide both a review of the relevant literature as well as a novel methodological axiom to analyse the creation and diffusion of AI technology. Essentially, our objective is to construct a relational database composed of academia publications on AI derived from Web of Science (WoS) data. Thereafter, this information is used to graph the distribution of AI knowledge in both the global and EU scientific communities. In doing so, it is possible to address the issue of when and where AI is created, as well as to identify potential trends in the evolution of this new disruptive technological domain.

As previously mentioned, very few empirical studies have managed to accurately disentangle the relationship between AI and socio-economic outcomes, primarily due to the lack of necessary data required. Furthermore, those notable exceptions that have addressed this issue have primarily looked at the impact of automation on labour outputs by using proxy data for the local exposure of AI methods. For instance, Acemoglu and Restrepo (2017) and Graetz and Michaels (2018) use data from the International Federation of Robotics (IFR) to estimate the regional and industrial exposure to robots and thus determine their impact on the local economy. Yet, their dataset by no means captures all the dimensions of AI; it merely proxies a fast-growing technology for the presence of robots in the industry. Hence, it does not actually allow the authors to correctly infer the multiple aspects or consequences of Artificial Intelligence in an economy.

Another common approach is to measure the likelihood that certain occupations will become automated by advances in the field of AI. Frey and Osborne (2017) famously pioneered this method when they gathered data on the probability that different cognitive tasks would become “computerised” in the future. Thereafter, they combined this information with the O*NET dataset, which describes the dependency of 702 distinct occupations on each of these tasks. Using both sets of data, they were able to estimate the prospect of automation for all 702 occupations. The task-based approach to measuring the risk of automation has since become a popular strategy for scholars looking to evaluate the relationship between AI and the labour market (OECD 2017; Acemoglu and Restrepo 2019). For example, both Arntz et al. (2016) and Nedelkoska and Quintini (2018) applied this approach to study the likelihood of automation for the OECD member nations.

However, despite its merits, these task-based approaches all suffer from the same empirical shortcoming in that they do not enable a thorough and detailed investigation into the presence
and diffusion of AI across local economies. As a consequence, they are not suitable methods for evaluating the determinants of AI knowledge production and its many implications on society. In fact, these task-based approaches only permit a high-level view of how AI-driven automation “might” affect one specific aspect of the economy: the local labour market.

Given the previously stated, it therefore appears that existing data are not sufficient to fully appraise the spread of AI, and as a consequence we are severely limited in our capacity to understand the drivers of this change and its impact on our society. Notwithstanding these numerous shortcomings, authors have begun exploiting recent advances in the field of text-analytics to circumvent these issues. Namely, previous research used text-analysis to classify patents and other documents into unique “technological” groups and used this information to infer the extent of innovation in a given domain, such as measuring environment-related technologies (Haščič and Migotto 2015). Following this logic, Mann and Puttmann (2018) applied a machine learning algorithm to a dataset consisting of texts from American patents to sort them into automation and non-automation innovations. Once they identified the automation patents, the authors could (geo)locate them in time and space. Further still, they could begin to expose the relationship between the volume of automation patents and local employment outputs.

Similarly, Cockburn et al. (2018) conducted a keyword search on a corpus of publications and American patents to distinguish which “inventions” should be classified as symbolic systems, robotics, and deep learning. Thereafter, the authors compared the diffusion of these methodologies across fields, regions, and time. Moreover, at least for the scope of this analysis, Buarque et al. (2020: 176) employed a list of technological classes and keywords to identify European patents that are associated with AI methods. In doing so, their goal was to “build a comprehensive data set of AI patents, which will enable us to study AI knowledge production and how it is distributed across the different regions and technological sectors of the European economy”.

Following Buarque et al. (2020), the present investigation employs a list of identifying keywords from WIPO (2019) to map the creation of scientific knowledge. Therefore, unlike most prior studies the focus here is not on patents, but rather on the scientific literature that concerns Artificial Intelligence. As the initial step, we created a subsample of academic documents from the Web of Science (WoS), our primary source of bibliographic data. WoS indexes approximately 280,000 scientific journals, as well as several conference proceedings and books. As such, it provides valuable information on academic publications, authors, institutions, and citations. Most importantly, however, the WoS also collects data on the keywords for each document – as provided by the authors. Exploring the information available in these keywords, we performed a search algorithm to identify and classify all the AI-relevant documents. More precisely, we looked within the publications for keywords that describe an AI method, like “Neural Networks” or “Genetic Algorithms”. We then classified and sampled a document as AI whenever it includes at least one keyword associated with the technology. Next, we used the metadata of these AI publications to graph the development of Artificial Intelligence in space and time. Thus, providing a valuable map of the creation and diffusion of Artificial Intelligence among the global scientific community.

While the present investigation should prove very useful, it does not provide empirical evidence on the determinants of AI creation, nor does it provide estimates on its potential implication for local economies. Instead, the objective is to offer a first glance at the creation

1 The present analysis is based on data retrieved from the following Web of Science bibliographic databases: “1980–2017 – Annual Science Citation Index Expanded and Proceedings-Science Combined”. 243
and diffusion of AI methods in the scientific world and on a variety of spatial scales – global, national, and regional. In turn, we hope to inspire and support more detailed empirical investigations into this emergent and meaningful technology.

Taking advantage of the proposed data and methods, future investigations in the field might shine a light on how to foster the development of AI as well as produce essential estimates on the impact of AI on social inequality and human well-being. Following this approach, further analysis will potentially contribute not only to our understanding of this general-purpose technology, but also inform policymakers seeking to design “smart policies” in the age of Artificial Intelligence. Most policy briefings currently emphasise the economic opportunities brought about by AI systems, and the need to educate displaced workers for the jobs of the future (OECD 2017). The methodological framework developed by Buarque et al. (2020), which we also employ here, will allow one to touch – at least marginally – on both of those issues. First, mapping the evolution of AI would allow for the recognition of the sectors/regions with related core competencies most suited for building a development pathway into this rapidly growing technology field (Hidalgo et al. 2007; Kogler et al. 2017; Whittle 2020). That is, mapping the diffusion of AI across regions and sectors would enable one to recognise opportunities to “invest in and develop AI for its many benefits” (The White House, Executive Office of the President 2016). Second, the AI “knowledge-space” (Kogler et al. 2013) could identify which sectors/regions are more likely to be affected by this expanding technology. In other words, studying the diffusion of AI could help to identify those more vulnerable to job displacement and other negative consequences.

**Identification strategy**

The first phase of our analysis involves identifying AI documents. The database which supports our analysis is a raw Web of Science (WoS) corpus containing articles from over 46m journals, books, and conference proceedings. Following, we will use all documents in that database that have been published over the period 1990–2016, which is about 38m records. While WoS shares commonalities with other bibliometric databases, including Google Scholar, Scopus, or Microsoft Academic Knowledge Graph, two noticeable differences are pertinent for our investigation. Firstly, WoS has a proclivity to favour journal articles over other outlets. Further, WoS has a bias towards the “Hard Sciences” (i.e. the natural sciences, engineering, and biomedical research, at the expense of social science and arts and humanities) (Mongeon and Paul-Hus 2016). Nevertheless, for the purpose of the present investigation these issues are not that detrimental. In fact, given the substantive nature of AI and the fact that journal articles are the preferred outlet for dissemination across relevant disciplines, this bias may even serve to our advantage and increase our overall coverage.

Turning to the information contained within each document, WoS lists the titles, journal names, year of publication, authors, and their affiliations, among other data. Most importantly, WoS provides a list of keywords for each document, as determined by the authors. For these reasons, WoS can rightfully be considered a strong medium for analysing the creation, integration, and evolution of Artificial Intelligence throughout space and time.

Given its inherently fuzzy nature, there is no easy way to identify the AI documents. To tackle this problem, we adopt a commonly used technique in bibliographic studies and apply a keyword identifier to the WoS database. Namely, we search across the keywords section of each document for AI-specific terms, such as “Machine Learning” or “Supervised Learning”. In turn, this approach enables us to classify all documents in our database as either AI or non-AI.

Naturally, the choice of AI identifiers will heavily influence our results. For this analysis, we follow the example set by Buarque et al. (2020) and borrow a list of AI-related keywords
produced by the World Intellectual Property Organization (WIPO 2019). We believe WIPO’s identification gives a more recent and specific definition of AI than other alternatives. Seeking to identify AI-related patents in Europe, Buarque et al. (2020) employed a list of 43 n-grams that are indicative of modern AI technology. Thus, we adopt the same list of words to recognise AI-specific knowledge in the WoS database, which includes the terms: “Artificial Intelligence”, “Data Mining”, and “Learning Algorithm” (see Appendix A for a full list of terms).

Using this methodology, we identified 260,351 documents as AI – out of the 38 million possible documents in our WoS database. Thus, although everyone seems to be discussing AI, very few have been able to seriously engage with the technology so far. To explore this issue further, Figure 13.1 plots the 20 most frequent keywords associated with AI. The histogram is positively skewed with “Neural Networks” being the most frequently occurring keyword. This is immediately followed by “Genetic Algorithm”, “Data Mining”, and “Support Vector Machine”.

It should be noted that we have carried out a thorough and careful stemming and cleaning process on our entire list of AI documents. These processes are necessary given the fact that inconsistencies are a common feature with any large-scale dataset. Moreover, one of the most common difficulties is that different elements actually represent the same thing. For example, the same keyword may be reported in a variety of ways (e.g. Neural Network or Neural Networks or “NN”). For the analysis reported here, we have applied a stemming technique to all of the keywords with the aim to minimise duplications.

2 A n-gram consists of a list of “n” items from a sample text. For example, in this analysis, we say that “machine learning” is a 2-gram.
Additionally, we consider a document to be AI if at least one of its keywords matches any of our 43 unique n-gram identifiers. Therefore, documents which we classify as AI can also include non-AI keywords. As such, our final database, which contains only the “AI” documents also includes non-AI keywords. Indeed, we expect every document in our database to have a mixture of AI-specific keywords, as well as terms that are not related to AI. That being said, this is valuable information since it allows us to study the integration of AI knowledge across space, time, and subject matter.

Data and methods

Global focus

Once we identified the AI documents within our WoS sample, we plotted the total volume of AI publications in space and time to obtain a better picture of the evolution of Artificial Intelligence. Along these lines, Figure 13.2 provides a first glance at the growth in the production of AI documents by continent over the past three decades.

Initially, the number of AI publications grew very slowly, which is not surprising due to the novelty of that field paired with uncertainties of how such a radically new technology could be applied in the market place. This is evident once we considered that between 1990–1999 the global number of AI publications was equal to 21,531 – which represents only about 8% of all documents in our sample and 22% of documents produced in the following decade. Nevertheless, even at this primitive stage, it is still possible to identify the Western economies of Europe and the Americas as key regional players, whereby they account for 5,632 (26%) and 5,555 (25%) respectfully. Finally, although not as prominent in the early years of the first period, it is also possible to identify a nascent cluster in Asia (19%).

In the second period (2000–2009) the first “real” surge in AI publications is evident. During this time, a total of 95,813 papers were published, signifying an almost five-fold increase over the first period. From a path-dependence perspective, Europe and America continue to dominate the initial years of the decade, accounting for 28% and 23% respectfully. However, this position is transformed following the emergence, and thereafter the dominance, of Asia. Throughout the decade, the Asian economy accounts for almost half (50%) of the global output, and from 2002 onwards reports a doubling of its publication output relative to its Western counterparts.

Equally important, during this period, AI begins to formalise as a discipline. Frequency analysis based on the journals’ keywords (see the following subsection) indicates that topics such as “Neural Networks”, “Machine Learning”, “Genetic Algorithms”, “Pattern Recognition”, “Fuzzy Logic”, and “Data Mining” began to emerge in documents during this time. More formally, these advances served to establish the foundation on which modern AI is based and thus can be regarded as setting up the pre-conditions for the final stage.

The third and final stage refers to the years 2010–2016, and it is during this time period that the bulk of the data lies. During this stage, we observe a further increase in the divergence between Eastern and Western economies. Focusing on the global crisis of 2008, we see that Asia was initially impacted by this crisis, but quickly recovered. In contrast, whereas Europe and

3 Over the period of analysis, the mean number of keywords listed on journal articles remained constant at 4.8. Therefore, for an article to be identified as an AI document, one of its keywords would need to match our identifier, but the remaining three or four keywords do not. This approach enables us to create a robust picture of how AI is beginning to integrate with other research fields.
the Americas appeared to be less affected in the years preceding and during the crisis, they have struggled to increase their total number of publications since then.

**Country-level analysis**

Expanding the initial analysis beyond the continental level, it is possible to use the authors’ affiliations on each WoS document to map the spatial distribution of Artificial Intelligence across countries. Doing so enables us to view continents in terms of their countries rather than as a collective. To illustrate, Figure 13.3 (Panels a, b, and c) displays the spatial distribution of AI-specific documents across countries for the three time periods discussed previously. Using this approach, we observe that Europe’s dominance in AI is partially the result of five countries (United Kingdom, France, Spain, Germany, and Italy), which make up nearly 62% of the continent’s total output. A similar pattern emerges in Asia, where China (47%) clearly dominates while other countries like Taiwan (7%), India (12%), Iran (7%), and Japan (7%) also make a noticeable contribution. The same pattern does not emerge for the Americas where the United States is continually the primary producer of AI with 70% of the continent’s contribution.4

Comparing between periods, we can further observe the path dependency process mentioned in the previous section. Namely, those countries with a historical advantage in AI production, that is, those leading the development of AI in our first period, continue to dominate in terms of the overall share of documents. Our results, thus, seem to corroborate past work in

---

4 As with any country level analysis the problem with the United States is that its sheer size distorts the innovative potential of the individual US states. Given the space constraints here, we do not go into detail as to which states in the United States are responsible for its dominance in AI, but future work should engage with these questions.
Figure 13.3  Distribution of AI documents by country and time period
(a) 1990–1999
(b) 2000–2009
(c) 2010–2016

Source: The authors; based on Web of Science records derived from the “Science Citation Index Expanded” and the “Conference Proceedings Citation Index”. GIS shapefile source is EUROSTAT.
Evolutionary Economic Geography (EEG), which highlights that regional innovation is driven by a path-dependent process (Martin and Sunley 2006; Kogler 2016). Furthermore, it is in line with the findings of Buarque et al. (2020) who showed the same tendency amid the AI-specific patents of Europe; those regions that excel in computing technology in the early stages of the analysis are also those with the largest share of AI patents in the end period.

On the other hand, our maps show the emergence of Asian economies, China in particular, as key players in the scientific development of AI technologies. One might explain the rapid rise of China as a by-product of global geopolitics. After all, as China marches to become the largest economy in the world, it is only natural that it will control the technologies of the future, for example, AI. Nevertheless, it is more likely that the growth in China is a result of the country’s ambitious policies regarding the development of AI (Dutton 2018). Since 2017, when the “New Generation of Artificial Intelligence Development Plan” was implemented, China has even invested more in the production of AI for the future than it has in the past. Thus, China’s continuous dominance in this technology domain will most likely prevail for quite some time.

Country collaboration network

Measuring the distribution in time of AI documents across countries is surely informative. Nonetheless, to obtain a better picture of the creation and diffusion of AI, we also need to account for international collaborative efforts. The seminal contribution of Wuchty et al. (2007) highlights how journal articles are increasingly found to be the result of collaborative efforts involving teams of researchers. Taking an Evolutionary Economic Geography stand, one possible reason for their findings is that there is an upper echelon limiting the extent an individual scientist, firm, region, or country can create all the knowledge they require internally; this seems to be particularly valid for the creation of general purpose or complex technologies whose production is the result of the recombination of multiple parts (Whittle 2019). Besides, Buarque et al. (2020: 177) have further commented that “AI is best developed when well connected to other research and development activities within the larger regional knowledge production ecosystem”. From this perspective, one might expect that collaboration between institutions, regions, and countries is at the heart of developing Artificial Intelligence knowledge. Further, to understand the creation of AI, it is paramount that we study cross-country collaboration networks in our WoS subsample.

Along these lines, Figure 13.4 uses the information on the co-location of authors listed within the same WoS document to generate a global collaboration network for AI. Country nodes are coloured in shades of grey according to the continent they belong to, whereas their size indicates the number of AI publications in that country. To draw the networks, we used a force-directed algorithm to ensure the position of the nodes is proportional to their graph distances. Hence, from this network, we can deduce that countries that are closer together collaborate more frequently than those further apart. Moreover, as one might expect, the major AI producers we identified in the last sections (e.g., China and the United States), also occupy the centre of the collaborative network.5 That is, these big AI producers also are among the most influential nation-states in the international collaboration network.

Beyond this, we observe that a large proportion of international collaboration occurs between countries within the same continent. Cultural and social arguments from economic

5 One potential reason why Europe might appear as less innovative to America (which is in contrast to Figures 13.1 and 13.2) is because Europe is a collection of many individual countries whereas the Americas are largely dominated by the United States.
geography may further help to explain these trends. Particularly, we expect that collaborations and interactions are more likely to occur between agents (e.g., individuals, firms, countries) that share the same language, customs, and routines (Boschma 2005). Finally, we must acknowledge those countries scattered on the periphery of the collaboration network who have yet to establish a serious footing in AI and may have only published a few articles. Many of these are either African or those smaller Asian economies. From an evolutionary perspective, it will be interesting to see how these economies develop over the coming years and whether they will enter the global AI collaboration network.

**Evolution of keywords**

As a methodological axiom, co-occurrence analysis is a valuable research strategy. It has found residence in a wide variety of fields, including economic geography (Kogler et al. 2013), regional development (Hidalgo et al. 2007), and scientometrics (Leydesdorff 2007). Likewise, with the advent of big-data, co-occurrence analysis based on the frequency of (key)words that occur in the same publication has also been identified as a burgeoning research field. Indeed, while earlier research sought to measure similarity across authors and fields using the co-citations networks (McCain 1990), more recent analysis has drawn on advances in text-mining and text-analytics to map knowledge structures using the co-occurrence of words. These developments are particularly helpful when tracing the evolution and intellectual structures of emerging new fields, such as the Internet of Things (Yan et al. 2015) and Infometrics (Sedighi 2016). Further, the methodology can also be used to produce bibliometric data on particular journals (Ravikumar et al. 2015), and to produce systematic literature reviews (Zhu et al. 2019).
Following this line of inquiry, Figure 13.5 illustrates the keyword co-occurrence network for the AI documents in the sample. It splits the analysis into the previously discussed periods in order to examine the changing research frontier of Artificial Intelligence. Here, each node represents a keyword with its size being proportional to the frequency at which it occurs in journal articles. Like before, when drawing these networks, we ensured that keywords that frequently co-occur across our AI data sample are closer together than those that do not. Doing so reveals a core-periphery structure with the most focal concepts at the centre. At last, for a better visualisation, the ten most common nodes are highlighted in red and have been labelled.

Although Artificial Intelligence initially developed slowly (see previous subsection about “global focus”), it still produced a very dense network of approximately 1,600 nodes and 32,000 edges. A plausible explanation is that, due to its infancy, many researchers were actively experimenting and trying to find applications for such a radical technology. In Figure 13.5(a), the three largest nodes are “Neural Network”, “Genetic Algorithm”, and “Pattern Recognition”, which,
to this day, are considered frontier concepts in the field. Furthermore, other concepts such as “Optimisation”, “Fuzzy Logic”, and “Classification” have also begun to emerge – although in a far smaller capacity. Like before, these concepts are also inherent to AI and, in particular, its trial-and-error processes. Lastly, whilst it is often very informative to know which nodes are present, it is equally valuable to know which nodes have yet to appear. To these ends, concepts relating to “Decision Trees”, “Data Mining”, or “Unsupervised Learning” are still missing at this first stage.

Moving to Figure 13.5(b), what immediately becomes clear is that there are significantly more nodes. Indeed, even after filtering we observe 400 more nodes than previously, which indicates that the network has grown. Likewise, we observe a significant shift regarding how said networks are wired. That is, comparing the two periods, not only the keyword co-occurrence network has increased in size, but it also has become denser and more concentrated. As with the period before, “Neural Network” remains the largest node, which is unsurprising given its focal positioning in the study of Artificial Intelligence (see Figure 13.1). Other noticeable changes include the introduction of keywords like “Data Mining”, “Support Vector Machines”, and “Reinforcement Learning”. Colloquially, these tokens are commonly used to explain Artificial Intelligence, so it is not that surprising they appear here. Beyond this, there was also an obvious concentration around the concepts of “Fuzzy Logic”, “Classification”, and “Optimisation”, which again have a strong resonance with AI.

Figure 13.5b  The changing research frontier of AI and the evolution of AI keywords: 2000–2009
Source: Authors’ calculation/illustration
However, perhaps the most important shift between the two time periods was the concentration around the keyword “Genetic Algorithm”. Although inherently different from Neural Network, the two terms co-occur rather frequently on the same documents. Indeed, throughout the period in consideration, a little over 2,000 documents listed both terms as keywords – which represents nearly 10% of all “Genetic Algorithm” articles in the period. In turn, this event highlights the potential for combining different AI methods to achieve faster and better algorithms, particularly for these two widespread methods.

Moving to the last stage, one can see that the network has become increasingly dense. This is largely because while the shift from Figure 13.5(b) to 13.5(c) reports only a small increase in the number of nodes, it sees a massive increase in the overall number of edges, or connections between nodes, which now stands at over 157,000. The surge in edges implies that AI has become increasingly intertwined with other disciplines. Otherwise stated, the

---

6 Neural networks are a sub-form of deep-learning where the algorithms are inspired by the structure of the human brain. In short, neural networks are trained to identify patterns in data (text, audio, visuals, etc.) and then predict outputs for a new set of similar data. Genetic Algorithms on the other hand reflect the processes of natural selection where the fittest mutations are selected for producing the next generation.
theory and methods which traditionally have underscored AI are now finding residence in new areas, including mechanical engineering, medicine, finance, and automation. These changes are the driving factors behind the self-driving car, smart home technologies, and mechanical medicine. Finally, keywords such as Neural Network (50,237), Genetic Algorithm (26,242), and Support Vector Machine (15,250) remain vitally important. We also see they are now produced closer together in the network; that is, they appear together on the same publications, which is very different than observed in the initial time period where they were distinct and further apart.

**Regional focus**

In this final section, we shift the focus of our analysis to the subnational level. That is, we will focus on the temporal evolution and spatial distribution of AI across 318 European NUTS2 economies.
regions. In this context, Figure 13.6 compares the total number of AI articles in our dataset over the NUTS2 areas in the initial period 1 (1990–1999) compared to the final period 3 (2010–2016). Supplementary information about the top AI producing regions of Europe is also provided in Table 13.1.

Surprisingly, Figure 13.6(a) shows that most regions appear to have had an early start publishing in AI. During the first period, the median number of AI documents was 68, which in part illustrates the technology’s novelty. At the same time, it is possible to identify some early AI “hotspots” – with the Île de France (FR10) appearing as a driving force. By the same token, there are clusters of activity in the South-East of England (UK13), Northern Italy (ITC4), and in Central Spain (ES30).

Moving between Figures 13.6(a) and 13.6(b), several differences are immediately apparent. Early hotspots like Île de France, Madrid, Lombardy, and London retained their status as leaders.

![Map of Europe showing AI article counts](image)

**Figure 13.6b** The spatial and temporal evolution of Artificial Intelligence science across European regional economies: 2010–2016

*Source: Authors’ calculation/illustration*
Table 13.1 Top AI producing regions

<table>
<thead>
<tr>
<th>NUTS2 region</th>
<th>Period 1</th>
<th>NUTS2 region</th>
<th>Period 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Île de France (FR10)</td>
<td>250</td>
<td>Île de France (FR10)</td>
<td>1,516</td>
</tr>
<tr>
<td>Lombardia (ITC4)</td>
<td>127</td>
<td>Madrid (ES30)</td>
<td>1,264</td>
</tr>
<tr>
<td>Inner London – West (UKI3)</td>
<td>118</td>
<td>Andalucía (ES61)</td>
<td>1,123</td>
</tr>
<tr>
<td>Madrid (ES30)</td>
<td>101</td>
<td>Inner London – West (UKI3)</td>
<td>1,021</td>
</tr>
<tr>
<td>South Yorkshire (UKE3)</td>
<td>96</td>
<td>Cataluña (ES51)</td>
<td>863</td>
</tr>
<tr>
<td>South Holland (NL33)</td>
<td>92</td>
<td>Norte (PT11)</td>
<td>765</td>
</tr>
<tr>
<td>West Midlands (UKG3)</td>
<td>91</td>
<td>Lombardia (ITC4)</td>
<td>714</td>
</tr>
<tr>
<td>Rhône-Alpes (FR71)</td>
<td>87</td>
<td>Fermo (IT14)</td>
<td>684</td>
</tr>
<tr>
<td>Southern West Scotland (UKM3)</td>
<td>84</td>
<td>Rhône-Alpes (FR71)</td>
<td>662</td>
</tr>
<tr>
<td>Attica (EL30)</td>
<td>83</td>
<td>Mazowieckie (PL12)</td>
<td>645</td>
</tr>
</tbody>
</table>

Source: Authors’ calculation

However, new regions have emerged indicating a restructuring among the most influential AI producing regions.7

More generally, the median number of AI documents per region rose from 68 to 344, which represents nearly a five-fold growth. Further still, this rise is potentially indicative of achieving a critical mass across both time and regions. Indeed, beyond the hotspots listed in Table 17.1, many regions are actively attempting to establish themselves as centres of excellence and as key players in the production of AI in Europe. Visually, Dublin (IE02), South Holland (NL33), Eastern Scotland (UKM2), and Oberbayern (DE21) have all moved into the foreground of scientific AI research, despite their weak starting point.

Nonetheless, one must be careful when interpreting the results. Foremost, because we are using the total number of AI documents, surely there are other decisive factors behind the patterns observed. Namely, Île de France – our chief AI producer – is also the largest metropolitan area in the continent. It has the largest population density, the most researchers and academic publications, as well as more patents and firms. Thus, declaring the region as the AI hotspot in Europe seems premature. Indeed, if we consider how innovation scales with the urban population, it might just be that Île de France is an average AI producer (Betten-court et al. 2010).

Still, we believe the table and graphs presented earlier do reveal worthy patterns. Continuing from the previous paragraph, the largest NUTS2 economies in Europe seem to concentrate most of the AI production on the continent. These results are perhaps unsurprising given those are the places with the most resources to invest in the development of this nascent technology; but, it also mirrors recent evidence that academic research “concentrates disproportionately in large cities” (Balland et al. 2020: 248).

However, in contrast with the data provided by Buarque et al. (2020) which uses patents, we find that scientific publications in Artificial Intelligence are far more diffused throughout

---

7 The results concerning the spatial distribution of AI scientific knowledge production across European regions, in particular for period 3 (2010–2016), are not necessarily what we would have expected at the onset based on prior research on the distribution of AI technical knowledge production (Buarque et al. 2020). We’ve conducted some further investigation into the validity of these results; all records for region PT11 (Portugal, North) were reviewed manually to ensure consistency in our geocoding approach, which confirmed the findings.
the continent. Smaller regions, and those often considered marginal in the European market, are producing a lot more scientific knowledge on AI than one might expect from the patent data. In turn, this could reveal inherent differences on how far and fast scientific or practical knowledge travels.

While the previous paragraphs describe the distribution of AI documents in Europe, the present analysis seeks to go beyond this and examine how AI-specific knowledge connects to other sectors of the regional economy. That is, we wish to estimate how AI is embedded in the local knowledge-producing and innovative environment. Since many AI-promoting policies seek to specifically develop with an eye towards multiple commercial products and processes (Dutton 2018), we need to understand how AI interacts with other domains of knowledge.

To construct a measure of AI embeddedness, we follow the approach introduced by Kogler et al. (2013, 2017) and produce a regional knowledge space for each European region. Using data contained in patent documents, these authors discern a measure of technological relatedness based on the co-occurrence of “Cooperative Patent Classification” (CPC) classes. Focusing on the US metropolitan areas, they discovered a link between higher levels of technological relatedness and faster rates of patenting per worker. In a subsequent analysis, Rigby (2015) found that technologies related to a region’s pre-existing knowledge were more likely to enter the said region than those that were further apart from the region’s expertise. Since then, geographers and regional scientists have applied these general principles to examine the innovative ability of cities, regions, and countries using a variety of indicators (see Whittle and Kogler 2020 for an overview of these methods).

Nonetheless, for the present examination, we are particularly interested in the recent works by Feldman et al. (2015) and Buarque et al. (2020). The first examined the diffusion of rDNA technology and illustrated how cognitive and geographical proximity affects the spread of this revolutionary method. Whereas Buarque et al. (2020) illustrated how bibliographical analysis alongside embeddedness studies can be used to measure the creation and integration of Artificial Intelligence in Europe.

However, despite their value, these contributions have exclusively used patent data to compute the innovative performance of regions. Whilst earlier research has recognised patents as an excellent proxy for innovation, especially on a regional scale (Acs et al. 2002), they also have significant limitations. Particularly, it has been argued that patents are the result of R&D and therefore reflect the innovative output potential of a region. In turn, we wish to study more succinctly the inputs of knowledge creation. To this end, by looking at academic publications, which form the bedrock on which many patents are created, we have a more accurate picture of the creation and diffusion of AI in the regional economies.

We proceed by employing the methods mentioned in the above paragraphs, only this time, we do so to our AI sample of the Web of Science. In other words, using detailed information within each article, we can produce graphs that map the co-occurrence of keywords across all the documents in a NUTS2 region – a local science space. Every WoS entry has at least one keyword, but most have between four and five. These keywords are signifiers and provide a snapshot of the document’s underlying knowledge. Thus, by examining the frequency at which individual keywords occur together in our data sample, we can generate matrices of how related these AI keywords are to one another. Namely, as we did for all the AI documents, we can assume that keywords that often co-occur together are more related than those that do not.
Hence, we graphed the AI scientific space for all the 318 European NUTS2 in our sample. To be specific, to make these graphs, we used all documents flagged as AI that contained at least one author residing in the region at the time of publication. Next, we plotted the data from these documents in a network, where each keyword is a node and, each time two keywords show on the same publication, we create an edge between the two. Then, to understand how the production and integration of AI knowledge vary across Europe, we collect several network characteristics data for all the regions’ scientific space. In other words, we can measure how dense, centralized, clustered, and how long it takes to traverse each AI regional network. In turn, we expect these statistics will allow us to infer how embedded Artificial Intelligence knowledge is across the European regions. For example, we imagine that where the AI scientific space is denser and longer, the technology is more rooted in the local innovation environment – as it seems to be more connected to other sectors.

Along these lines, Figure 13.7 shows the distribution of two network statistics from the AI scientific spaces for all the regions with at least 100 documents published between 1990–2016. On the top graph, we show the distribution of network density – a measure of the proportion of existing edges out of all possible links in the network. And on the bottom, we display the degree centralization of these scientific spaces, which captures how central the most central node is, compared to all other nodes in the network. To measure centralization, we first calculate the sum in differences in centrality between the most central node and all other nodes, then divide this value by the theoretically largest possible difference in any network of the same size. As one may conclude, the European regions are quite different regarding their “AI embeddedness.” Indeed, it seems like AI-specific knowledge is far more centralized in some regions than others. The region’s AI scientific space is more concentrated in a few keywords, and it lacks some potential applications and alternative methods. On the other hand, some regions have a denser network – thus demonstrating a more connected AI knowledge space. The result is not surprising in light of research by Buarque et al. (2020) who measured a very skewed distribution for the AI centrality index in patents.

**Figure 13.7  AI science space network statistics**
The graphs display the density distribution for two network statistics collected from the science space of 213 European NUTS2 regions. We include in these figures all the places with at least 100 documents published during the 1990-2016 period. The x-axis shows the values for the descriptive statistics, while on the y-axis, we display the number of places presenting that range of values. We scale the y-axis between one and zero, where the statistic with the highest number of regions is equal to one. Network density refers to the proportion of existing edges out of all possible links in the network. We measure degree centralization according to the formula

$$\frac{\sum_{i=1}^{N} (c_i(p_i) - c_{\text{max}})}{\max \sum_{i=1}^{N} (c_i(p_i) - c_{\text{min}})}$$

The nominator represents the sum in differences in centrality between the most central node and all other nodes, and the denominator the theoretical maximum value of such differences.

We must remark, however, that it can be challenging to make meaning from these network statistics. First, it can be complex to interpret the results. For example, what conclusions can we make from finding that a region has a degree centralization of 0.5? Without a proper benchmark, these network variables do not tell us much. More importantly, there is ample evidence that network statistics vary significantly even when the networks come from the same model. And that these metrics are highly dependent on the networks’ sizes (Anderson et al., 1999; Faust and Skvoretz, 2002; Van Wijk et al., 2010). Along these lines, we understand that “to make comparisons, the measures must be on the same scale and have the same meaning across the various networks; however, when the networks being compared vary in size, often the values of these network statistics can be dominated by size effects” (Smith et al., 2016). Therefore, for better metrics, we follow the seminar work by Anderson et al. (1999) and first compare the regional networks to random graphs (Erdős-Rényi) of the appropriate size. That is to say, for every NUTS2, we generate 1,000 random networks containing the same number of nodes and density. We then collect the same statistics from these random graphs and normalize the metrics from our regional science space against their equivalent random distribution. We thus obtain scale metrics between one and zero, and Figure 13.8 displays the density distribution for two such values – network diameter and average path length. Both these statistics offer us information about how
Figure 13.8  AI science space network statistics scaled

Source: Authors’ calculation/illustration

Note: The plots display the density distribution for two network statistics from the science space of 214 European NUTS2 regions – those with at least 100 documents during 1990–2016. The x-axis shows the values for the descriptive statistics, where we scale these values against the distribution of 1,000 equivalent statistics from random graphs of similar sizes (Anderson et al., 1999). The y-axis represents the number of places presenting that range of values. We scale the y-axis between one and zero, where the statistic with the highest number of regions is equal to one. Network diameter refers to the shortest distance between the two most distant nodes of the network. And average path length is the average number of steps along the shortest paths of all possible pairs of nodes.
difficult it is to transverse the regional science spaces. And once again, we find the European regions have very different AI networks regarding their connectivity and span.

Given the significant divergence across the European regional networks, one might propose to examine how the level of AI embeddedness relates to the amount of AI knowledge produced by the NUTS2. To these ends, Buarque et al. (2020:186) sought to expose the correlation between the number of AI patents and their relative importance for the region’s knowledge space. Accordingly, they show “there is a positive correlation between those regions where AI patents are most prevalent and those for which AI is most embedded.” We must remark, nonetheless, that our methodology diverges from Buarque et al. (2020) in one vital detail. They estimated the value of AI patents in the regional knowledge space by artificially removing said patents when building the graphs and thus observing the impact of this exclusion on the networks’ structures. In contrast, the present analysis exclusively represents the scientific space for the AI documents. We consider only the co-occurrence of keywords within articles that we identified as AI and do not estimate the region’s overall scientific space - i.e., our focus concerns the AI-specific knowledge.

We must equally recognize that we are focusing on differences regarding the networks’ global statistics. These provide valuable insights into the overall structure of the regional science spaces and allow us to reflect on how inclusive or rooted is AI knowledge. But there are limitations to using these methods. For example, as mentioned before, the global statistics do not necessarily “yield robust results,” and “often fails in catching important local features” like communities and others (Tantardini et al., 2019:3). Of course, the literature offers several other techniques one could use to measure diversity across the AI science spaces, and Tantardini et al. (2019) provides a valuable review on these many alternatives. Although beyond the scope of this initial examination, one could readily employ any preferred “network comparing method” to strengthen our comprehension of how distinct are the European NUTS regarding their AI production. Likewise, they could measure how similar the regional networks are to one another. Hence, future research can profit from our approach coupled with robust statistics to examine what distinguishes the AI key players. What are the driving forces behind the different AI embeddedness? Is there notable structural variance across the regions – e.g., some are more focused on applied systems than others? And even study the role of social and geographical proximity in shaping these regional networks – i.e., do regions nearby exhibit more similar AI science spaces?

Along these lines, to further substantiate the value of our science space statistics, Figure 13.9 illustrates two regional AI networks we built using data from 2010-2016. On the top, we have Dublin (IE02) and on the bottom Vienna (AT13). For comparative purposes both regions belong to high-income countries, have a similar number of universities, and enjoy a very high standard of living. As depicted earlier, both regions are significant producers of Artificial Intelligence and have roughly the same amount of journal articles, IE02 (422) and AT13 (381). However, despite their commonalities, these regions produce very different network structures with AI occupying a more central position in Dublin.

In terms of the sheer number of AI-specific keywords, Dublin and Vienna are once again very similar. Of the 36 keywords8 listed by the WIPO (2019), Vienna has published with 23 of them, whereas Dublin published with 26. Although the volume of documents and keywords in a region is indicative of its capacity to produce AI knowledge, in order to understand the region’s full potential we must also account for the links between keywords and the interconnections among the different kinds of knowledge produced. Along these lines, comparing the

---

8 The World Intellectual Property Organization (WIPO 2019) provides 43 n-grams, which we used to identify the AI documents. We grouped the different n-grams into 36 keywords. Namely, we grouped together terms that refer to the same or very similar method – such as, “Supervised Training” and “Supervised Learning”.

261
two networks shown in Figure 13.9a, we first observe that Dublin’s scientific AI space has more nodes overall and is denser than Vienna’s. That is, visually, we can conclude that Dublin combines more industries and sectors into its AI network; thus leading to a more diverse and applied technology when compared to Vienna.

Furthermore, in both networks under consideration, we highlight “Neural Networks” and “Machine Learning” as the most traversed nodes, that is, the most relevant keywords. For Dublin, you can see that these keywords are closely connected, indicating they frequently occur in the same publications. Moreover, these keywords are also tightly surrounded by other nodes (both AI and non-AI), further demonstrating their recombinatorial potential. Recall from the previous sections how AI has become increasingly intertwining with other sectors of the economy (Frey and Osborne 2017) and technological frontiers (Buarque et al. 2020); it is precisely this recombination that is driving AI policy and regional development (Clifton et al. 2020; Acemoglu and Restrepo 2020).

Conversely, whilst “Neural Network” and “Machine Learning” are also the most connected nodes in Vienna’s network, they are not as embedded in the region’s scientific space. Thus, it seems that the region has been unable to connect distinct research frontiers in AI, which significantly hampers its ability to harness the capabilities of Artificial Intelligence. Insights from
Evolutionary Economic Geography (Kogler 2016) further substantiate this point, illustrating that though Vienna might have the necessary building blocks, it fails to connect them in a meaningful way and as a result their network remains sparsely connected.

**Conclusion**

Artificial Intelligence is currently one of, if not the most, widely debated science-technology fields in business and policy circles, and the rush to develop and market AI-related technology is palpable. Since its emergence in the early 1990s, governments around the world have been keen to develop strategies in order to harness and capitalise on its obvious societal and economic potential. In this context, the purpose of the present contribution is to provide insights into the spatial and temporal evolution of AI scientific knowledge production over the past three decades. Following this vision, the objective we set out with was to make a series of connected contributions to the relevant literature, both theoretical and empirical, all of which should inspire and support further work on this relevant topic. In particular, the study provides insights into three aspects that should advance this line of inquiry.
Firstly, to overcome a lack of precise definition, we augmented the methodology developed by Buarque et al. (2020) to identify AI-based journal articles and indexed approximately 260,000 such documents. To establish a solid foundation, we began by conducting an exploratory analysis in order to examine the path-dependence of countries (continents), as well as to identify the jurisdictions that were the leading AI producers throughout each period in time.

Secondly, utilising the keywords listed on journal articles we examined the changing research frontier of AI. Through a detailed analysis of their co-occurrence, it was possible to explore how AI is simultaneously becoming more concentrated and diverse. Concentrated by virtue of the fact that the core concepts – neural networks, genetic algorithm, and machine learning – that define AI are appearing more frequently on journal articles over time. Diverse in terms of the number of non-AI keywords that are also appearing alongside them. This indicates AI’s recombinatory potential whereby theories and methods that traditionally have underscored and defined core AI research are now also finding residence in new areas, including mechanical engineering, medicine, finance, and automation. As mentioned previously, it is precisely these principles that have led to the creation of the self-driving car, smart home technologies, and mechanical medicine.

Finally, we positioned these AI documents into the scientific knowledge space (Kogler et al. 2013, 2017) of two capital EU regions and developed a methodology for describing how embedded AI is in these regions. The results reveal that although Dublin (IE02) and Vienna (AT13) are very similar in terms of their overall number of publications and AI keywords, by the end these two places produce very different scientific knowledge production networks. A preliminary finding here is that AI knowledge production is more central in Dublin’s network, and as such, that Dublin might be better equipped to further harness its capabilities. On the other hand, whilst Vienna has the necessary building blocks to potentially exploit AI scientific knowledge, it has yet to connect these in a meaningful way to its broader network structure (i.e. other non-AI subjects).

In terms of next steps, an obvious direction would be to extend the methodology that was utilised in this study and to include information embedded in the relevant publications regarding authors and their institutions. The addition of this micro-dimension would permit a more thorough and detailed analysis of both the creation and diffusion of AI focusing specifically on those actors involved. In doing so, it would be possible to analyse not just which countries are collaborating, but also the institutions and individuals embedded in these countries. Such an analysis would be of critical importance in identifying those institutions that are at the forefront of AI scientific knowledge production and could be used by policymakers and funding agencies when targeting specific investment opportunities. Similarly, by focusing on authors, it would possible to discern (at the institutional/departmental level) who is collaborating with whom. For example, if a researcher in computer science (i.e. AI research), is engaged in a collaborative process with a colleague in medicine, the expectation then would be that scientific AI knowledge gets applied to a specific problem, and in turn provides inputs to generate a solution. It is these inter-disciplinary collaborations that provide the opportunity to produce recombinant knowledge with the potential to push forward technological change and the research frontier in the science/technology knowledge space (Kogler et al. 2013, 2017; Kedron et al. 2020).

This point also speaks more broadly to a crucial methodological contribution in the present investigation. In particular, we further substantiate the viability of text-matching and text-analysis methodologies for identifying and analysing the creation and diffusion of science and technologies when they are not easily identifiable by traditional means. Haščič and Migotto (2015) provide an additional example that follows this approach where a text-matching algorithm is employed in order to identify “green” technologies. We hope that the present study
will inspire other scholars to further explore AI scientific knowledge production as well as other non-standardised technology fields at the intersection of traditional domains.

References


## Appendix A

<table>
<thead>
<tr>
<th>Table 13.2 List of AI keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>artificial intelligence</td>
</tr>
<tr>
<td>bayesian network</td>
</tr>
<tr>
<td>deep learning genetic</td>
</tr>
<tr>
<td>machine learning</td>
</tr>
<tr>
<td>unsupervised learning</td>
</tr>
<tr>
<td>inductive logic</td>
</tr>
<tr>
<td>learning algorithm</td>
</tr>
<tr>
<td>gradient tree boosting</td>
</tr>
<tr>
<td>logistic regression</td>
</tr>
<tr>
<td>latent dirichlet allocation</td>
</tr>
<tr>
<td>neural network bayes</td>
</tr>
<tr>
<td>connections</td>
</tr>
<tr>
<td>natural language generation</td>
</tr>
<tr>
<td>generation</td>
</tr>
<tr>
<td>semi-supervised learning</td>
</tr>
<tr>
<td>tree</td>
</tr>
<tr>
<td>inductive program learning</td>
</tr>
<tr>
<td>algorithm</td>
</tr>
<tr>
<td>transfer learning</td>
</tr>
<tr>
<td>adaboost</td>
</tr>
<tr>
<td>xgboost</td>
</tr>
<tr>
<td>logistic regression</td>
</tr>
<tr>
<td>latent dirichlet allocation</td>
</tr>
<tr>
<td>source: WIPO (2019)</td>
</tr>
</tbody>
</table>

268