A Semantic Analysis of the Recent Evolution of AI Research

Juan Mateos-Garcia, Joel Klinger, Konstantinos Stathoulopoulos and Russell Winch

November 2019

Executive summary

Fast-improving Artificial Intelligence (AI) systems are being applied in a growing number of areas, from internet search and social media to the analysis of health scans and management of power grids. Economists are hailing AI as a general purpose technology that will revolutionise our economy and policymakers are putting in place national strategies to spur its development and diffusion.

Powerful deep learning networks that identify patterns in vast datasets and reinforcement learning algorithms that learn through trial and error in synthetic environments, have overtaken previous AI approaches that programmed logic into computers and taught them from experts. Although these new methods have achieved sensational breakthroughs, they also have important limitations that could restrict their applicability and benefits, and/ or create risks when they are deployed, for example in terms of discrimination against vulnerable groups, manipulation by malicious actors and unexpected outcomes when they

interact with each other in the wild. An increasing awareness of these issues means that policies to encourage more AI activity and its diffusion – in particular, to tackle big societal challenges through innovation missions – are going hand in hand with research on fairness, accountability, transparency and safety, regulatory changes and a proliferation of ethical charters to encourage responsible innovation. Taken together, these efforts amount to what we call a third-wave Research, Development and Innovation (R&D&I) policy, concerned not just with the levels of AI activity but also its direction.

This policy programme needs to be informed by relevant, inclusive, trusted and open data and indicators, which go beyond aggregate measures of the volume of AI research and how it is evolving, to consider its composition, inclusion, diffusion, geography and purposes: we need smarter data about smarter machines.

We have collected and enriched data from arXiv, an open repository of research widely used by the AI community. We combined it with several other sources and analysed it using data science methods to map trajectories in AI research and explore its drivers, illustrating how novel data sources and methods can inform novel policy frameworks to steer AI in societally-beneficial directions.

Our analysis shows that AI research has grown rapidly in recent years: 77 per cent of AI papers in arXiv were published in the last five years. This is not just about computer science: other fields have also experienced fast increases in the number of papers that use AI methods to tackle important scientific challenges. Growth in activity has been accompanied by shifts in its composition, with deep learning algorithms and applications such as computer vision overtaking symbolic and statistical methods: the share of papers about deep learning has multiplied four-fold since 2012, while the share of papers using statistical methods has halved. These thematic changes have been accompanied by shifts in the geography of the field, with China trebling its share of global AI research since 2012 and some European countries falling behind, especially in cutting edge methods.

In forthcoming work, we will explore how various factors such as gender diversity, corporate participation, regional clustering and the involvement of countries with different political values in AI research are shaping its trajectories. This analysis will illustrate how smarter data about smarter machines can inform activist AI R&D&I policies to steer AI in a direction where its benefits are more widely shared and its risks more wisely managed.

1. Introduction

An AI revolution (again)

Building intelligent machines has been one of the driving ambitions of computer science since the days of Alan Turing and significant efforts have been devoted to this purpose in the decades since (Dyson 2012). Previous strategies to achieve Artificial Intelligence (AI) had limited success, however. Symbolic approaches to program logic into computers in the 1950s, and expert systems that learn rules of behaviour from human experts in the 1980s were too difficult to scale to the variety of situations where an AI system might be expected to operate (Markoff 2016). Important aspects of our intelligence and how we perceive and behave in the world were found to be too hard to codify and therefore implement in AI systems (Russell 2019).

Machine learning approaches that bypass the challenge of programming intelligence in machines by, instead, training them from examples or letting them learn through trial and error in synthetic environments, have overcome some of these challenges and delivered AI systems that are able to function effectively in a variety of situations; in some cases even outperforming humans (Russell 2019; Goodfellow, Bengio, and Courville 2016; LeCun, Bengio, and Hinton 2015; AI Index 2017). Some domains where AI systems have experienced rapid improvements include game playing, machine translation, information retrieval, image classification and speech recognition. Differently from previous AI booms, today's AI systems are already powering mainstream applications in search engines, social media networks, translation systems, voice assistants and (partially) self-driving cars (Brynjolfsson and McAfee 2014; McAfee and Brynjolfsson 2017).

AI impacts and risks

It is widely believed that AI systems could transform many domains beyond the technology sector including health, manufacturing, transport or scientific research (McAfee and Brynjolfsson 2017). They could also help tackle some of society's biggest challenges, such as the prevention and treatment of chronic diseases, environmental sustainability and the decline in productivity in scientific research, to name a few (Topol 2019; Rolnick et al. 2019; Agrawal, McHale, and Oettl 2018). The broad relevance of the capability that AI systems promise to deliver ('to behave appropriately in many different situations') has led economists to herald AI as the latest example of a general purpose technology that will define a new economic era in the same way that steam, electricity or the transistor did in past times (Cockburn, Henderson, and Stern 2018; Klinger, Mateos-Garcia, and Stathoulopoulos 2018; Trajtenberg 2018). AI's potentially pervasive impact has raised concerns about disruption in labour markets as smarter machines encroach on an expanding range of tasks and occupations (Acemoglu and Restrepo 2018; Restrepo and Acemoglu 2018; Ford 2015), perhaps even to the point where humanity itself is rendered obsolete by exponentially improving machine 'super-intelligences' (Bostrom 2017).

In the shorter term, there are increasing concerns about the risks of AI systems that could entrench inequality if they learn the biases in historical data or make mistakes that disproportionately impact minorities and vulnerable groups (Bostrom 2017; Noble 2018; Eubanks 2018; Buolamwini and Gebru 2018), be gamed by malicious actors (Brundage et al. 2018), turned into weapons or tools of surveillance that abuse personal data to monitor and exploit users and citizens (Zuboff 2019), or create barriers for entry in increasingly concentrated markets (Furman and Seamans 2018). AI systems could also behave in unsafe ways if the metrics they seek to optimise are not well aligned with human goals (Amodei et al. 2016), or if they create dangerous emergent phenomena when they interact with each other in complex environments, like high frequency trading algorithms did during the 'flash crash' of the New York Stock Exchange in 2010.

Some of these risks stem from modern AI systems' reliance on deep learning algorithms that learn increasingly abstract patterns from large amounts of unstructured data, such as video or text (we will sometimes use the term 'connectionism' to refer to this programme of research). Although deep learning systems have strong predictive power inside the domains where they are trained, they lack robustness, interpretability and common sense. This tends to break down when exposed to new situations, including strategic behaviours by users seeking to manipulate them. It can also be difficult to understand their internal operation and outputs. Further, they can create safety issues when they greedily optimise performance metrics independently of the actual goals of their programmers, users and wider society (Mateos-Garcia 2018; Marcus and Davis 2019). Their reliance on large datasets and computational power could make them anti-competitive, privacy-infringing and environmentally unsustainable (Amodei and Hernandez 2018). Some have argued that these limitations require new approaches to AI that combine modern connectionism with ideas from previous (symbolic and rule-based) AI eras (Marcus and Davis 2019; Marcus 2018).

Third-wave research, development and innovation policies for AI

There is increasing recognition of the need for policy action to manage the processes through which AI systems are developed and deployed, leading to a proliferation of `AI strategies' around the world (Stilgoe et al. 2013; Jobin et al. 2019).[1] These Strategies generally seek to nurture national high-growth AI industries, and encourage the diffusion of AI systems into other sectors in ways that are safe and consistent with ethical and political values. Meanwhile, private sector organisations, non-governmental organisations and the research community are creating ethical charters and guidelines to encourage responsible innovation (Jobin, Ienca, and Vayena 2019), and fairness, accountability and transparency, and safety research groups, are exploring technical and institutional solutions to various AI risks.

Together, these activities represent an example of what we have described in previous work as a 'third wave' research, development and innovation (R&D&I) policy framework (Nesta 2019). Differently from older (first wave) approaches that focused on increasing research and development investment levels without paying attention to its purpose (first wave), and (second wave) models aimed at increasing the transfer of knowledge from university to industry, the third wave of R&D&I policy is directional: it seeks to steer the development and diffusion of new technologies mindful of their purposes and impacts (Stirling 2009, 2014). This approach is informed by a growing body of research in evolutionary economics, complexity economics, and science and technology studies suggesting that increasing returns (growing momentum) in the deployment of new technologies can create sub-optimal scenarios where second best (or worse) technologies end up being adopted (Aghion, David, and Foray 2009; W. Brian Arthur 1994; David 1985). There are many potential reasons for this including random events, strategic behaviours (e.g. investments on marketing or lobbying) and the preferences of lead developers and adopters early in the lifecycle of a technology or industry (Brian Arthur 2014; Garud and KarnÃže 2001). Once a technology gains an early advantage against its competitors, network effects and sunk investments in complementary assets such as infrastructure and skills could make its success irreversible even if it is inferior in the longer run.

An implication of this is that the emergence and deployment of new technologies does not have a single equilibrium. Instead, we could imagine a collection of parallel universes, each of which is dominated by a qualitatively different technology. Activist R&D&I policymakers try to identify, among all these technological universes, which is more societally desirable and put in place interventions to bring it about (Mazzucato 2015, 2018; Kattel and Mazzucato 2018; Cantner and Vannuccini 2018). Some examples include:

- Stronger levels of public engagement during the development of R&D&I policies.
- Policies to increase inclusion in the R&D&I workforce.
- Mission-oriented innovation policies to support R&D&I activities to tackle specific social challenges.
- Developing norms and practices to incorporate ethical considerations into technology development.

Smarter data about smarter machines

We argue that in order to be effective, third-wave R&D&I policies to steer AI in societally-beneficial trajectories need to be informed by new, ‘smarter’ data (Bakhshi and Mateos-Garcia 2016; Nesta 2019). By this, we mean data that is:

- **Relevant**: Smarter data for AI policies should capture AI R&D&I activity with high timeliness and resolution, helping to measure not only aggregate levels of AI activity but also their composition in terms of the technological trajectories that are being pursued and deployed (Teece 2008; Dosi 1982). It should also capture geographical, institutional and relational dimensions of R&D&I: where the activity happening, and what organisations and networks are involved in it. Ultimately, this data should enable analyses that not only describe, but also explain the dynamics of AI R&D&I, thus helping formulate policies to shape those dynamics.
- **Inclusive**: We need to understand what social groups and types of organisations participate in AI R&D&I, and what communities and interests are absent and therefore potentially neglected when AI technologies are selected, opportunities pursued and risks highlighted or downplayed. This requires timely indicators of socio-demographic, sectoral and spatial inclusion in AI R&D&I, and their links with the nature of technologies that are developed and the goals they seek to achieve.

Traditional data sources and indicators based on publication and patent counts, number of AI businesses or university graduates with AI skills are insufficient to deliver the relevant and inclusive evidence that AI policymakers need. New data sources that capture the creative and collaborative mechanisms used in AI R&D&I – crucially including open source, data and dissemination channels as well as conventional Intellectual Property Rights – and its diffusion have much to contribute. Data science methods that extract quantitative patterns from text descriptions of AI R&D&I activities and enrich them with additional information about the identities of participants, their locations, affiliations, goals and networks can help capture the micro-dynamics of AI R&D&I and its drivers (Bakhshi and Mateos-Garcia 2016). This illustrates how AI and allied techniques such as machine learning and natural language processing can recursively transform its own analysis.

However, like in other domains where AI is being applied, these methods also come with challenges. In particular, there is the risk that complex analytical methods to map AI may yield results that are difficult to explain, interpret or use to make policy decisions, or that the size and proprietary nature of the data sources they rely on and their technical sophistication create barriers for their replication and expansion. We believe that in order to address these significant risks, smarter data for AI policy also needs to be:

- **Trusted**: In order to be used, new data sources and methods to measure and map AI need to be trusted first. The best way to create this trust is through rigorous validation with better known and understood data sources and domain experts, and the use of qualitative methods to make new results interpretable, meaningful and actionable. It is also vital that the results of AI mapping efforts based on new methods are reproducible. The results of machine learning, natural language processing and clustering analyses can be sensitive to the datasets used for training, and the selection of parameters by the analyst (or the algorithm). Understanding the robustness of experimental results to different contexts, assumptions and model specifications is critical for determining where and how they can be used to make decisions.
- **Open**: The best way to build trust in new data and methods is by making them openly available (while subject to constraints around the release of sensitive information such as personal data) so that other researchers can review, validate and build on them (Peng 2011; Burgess et al. 2016). This strategy makes it easier to improve methods, detect errors and combine sources to triangulate findings and explore new questions. It also reduces inefficiency in research and lowers barriers to the adoption of new techniques, making the field of AI mapping more inclusive too.

In this report, we present a pipeline for data collection, processing and analysis of data about AI research that fulfils these features of relevance, inclusiveness, trust and openness with the goal informing directional AI policies. Before describing its pipeline, we summarise relevant work.

Relevant work and our contribution

The notion of technological trajectory was put forward by evolutionary economists in the 1990s as a challenge to mainstream economics' aggregate, undifferentiated 'black box' view of technological change, where innovation is conceptualised as a productivity-enhancing investment in knowledge, disregarding the fact that this could take many different forms and bring out wildly-varying outcomes in terms of economic structures, distributions of benefits and costs, technological risks etc (Dosi 1982). Influenced by economic history, sociology of science and complexity science, the analysis of technological trajectories pays strong attention to the historical process through which new technologies emerge and evolve, and the preferences, worldviews and goals of those involved, as well as their economic incentives (Rosenberg and Nathan 1994, 1982; Kuhn 2012; W. Brian Arthur 1994; W. B. Arthur 1999; Garud and KarnÃže 2001). It acknowledges that historical phenomena are path-dependent and sometimes irreversible, potentially leading to sub-optimal outcomes (David 1985).

Some of these ideas are echoed in the work of AI researchers and practitioners who have described the evolution of AI as a collection of parallel trajectories involving various technologies and markets that are integrated and interact as 'Comprehensive AI Systems' (David 1985; Drexler 2019). Others have called for better models of AI progress that measure the links between inputs (computation, data and skilled workers) and outputs (advances in AI capabilities) in order to inform technology foresight (Brundage 2016; Prediger 2017), and expressed concerns about how discrimination in the AI workforce may embed discrimination in the (path-dependent) AI systems that are deployed (Stathoulopoulos and Mateos-Garcia 2019; Myers West, Whittaker, and Crawford 2019). The majority of these analyses have until now remained conceptual and qualitative.

Parallel to them, we have started to see a growing number of studies that use novel methodologies to measure AI R&D&I activity. Some examples include maps of the AI research and development landscape using publications and patents (Elsevier 2018; Intellectual Property Office, n.d.; Cockburn, Henderson, and Stern 2018; Mann and Püttmann 2017), in some cases linked to business activity (Centre 2018). Other initiatives such as the AI index have adopted a broad-based approach to the measurement of AI R&D&I using a variety of indicators that also capture skills supply and improvements in AI performance metrics among other factors (Index 2017, n.d.). Most of these analyses are descriptive, capturing the evolution of activity in AI R&D&I, its diffusion in different academic fields and its geography. Although several of them use natural language processing methods, such as topic modelling or keyword co-occurrence analysis to create topical maps of AI research and the trajectories that different countries specialise on, so far there has been limited effort to understand the reasons for the patterns that are identified and their theoretical or policy implications. There is a lack of standardisation in the strategies used to define and operationalise AI, creating the risk of contradictory findings, for example regarding the position of the European Union in 'global AI rankings'.

We have previously contributed to this literature through analyses of AI R&D in the arXiv database (about which we will have more to say later) with a specific focus on the geography of deep learning and its drivers (which we characterised using topic modelling) (Klinger, Mateos-Garcia, and Stathoulopoulos 2018), and a study focusing on gender diversity in AI research, helping build the evidence base about (lack of) inclusion in the field (Stathoulopoulos and Mateos-Garcia 2019). In addition to publishing our results, we have released the data and code for our analysis so that other researchers can reproduce and build on our efforts.[2]

Here, we present new results of our analysis of AI activity in arXiv data with a particular focus on the research trajectories followed in 'open' AI R&D and their drivers. Our goal is to provide a detailed account of the recent evolution of the field – in particular, how it has been transformed by the advent of deep learning – and to illustrate opportunities to generate policy-relevant, smarter data about smarter machines, using data science methods and new combinations of open data sources. In doing this, we seek to provide an empirical grounding for current perceptions and concerns about the evolution of the field, providing a rationale for directional AI R&D&I policies. This report will be followed by a collection of case studies focusing on:

- The connection between AI technological trajectories and gender diversity in research teams.
- Participation of the private sector in AI research and its link with the research trajectories that are pursued.
- Regional concentration of AI research and its links with the geography of automation (with a focus on England).
- Participation of illiberal countries in AI research with a particular focus on their involvement in the development of controversial visual surveillance AI technologies.

Together with the report, we have also published arXlive, an open-source, real-time tool to monitor AI research trends found in the arXiv repository, providing a source of data to update our analysis and undertake new ones.[3]

Section 2 outlines our data and methodology. In Section 3 we summarise recent trends in the evolution of the field, and Section 4 is our conclusions.

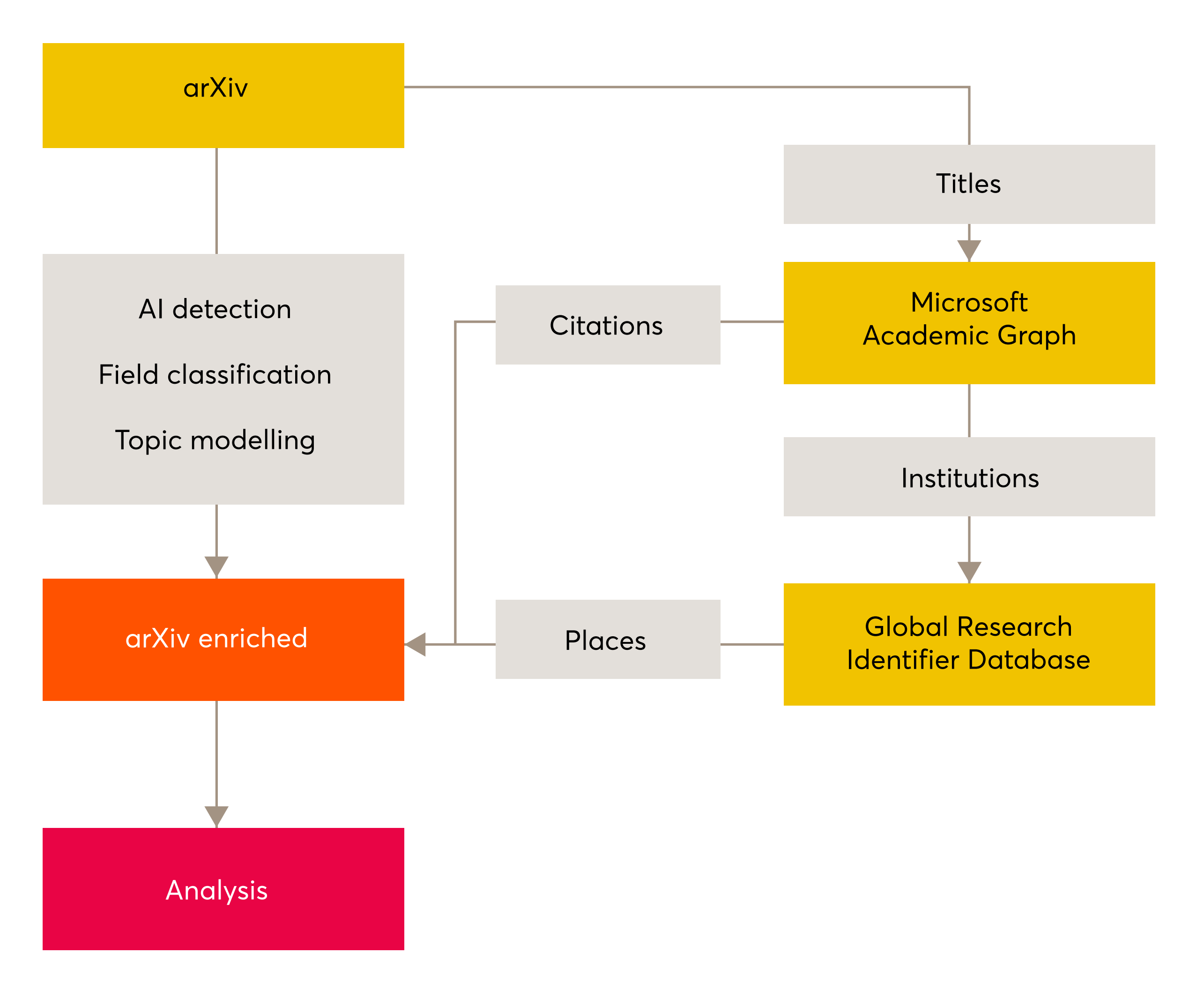
2. Data and methodology

This section introduces our data sources – how we collected them and enriched them – as well as key components of our analysis.

Data sources and processing

Our analysis involves a complex assemblage of data sources and methods. Figure 1 represents this pipeline.

Figure 1: Data sources and process



The core dataset we use in our analysis is arXiv, an open pre-prints website with 1.6 million papers that is widely used by various Science, Technology, Engineering and Mathematics research communities.[4] In recent years, arXiv has become an important channel for the dissemination of AI research in academia and the private sector. As an example, almost all research papers by DeepMind and OpenAI, two leading AI research labs, are available from arXiv.[5] Just under 60 per cent of the documents referenced in import ai, an influential newsletter monitoring AI research trends, are in arXiv.[6]

We collect data from arXiv and enrich it with information about the institutional affiliation of AI researchers and their location. Klinger et al (2018) and Stathoulopoulos et al (2019) provide a detailed account of the methodology used for this. Here we summarise.

The institutional and geographical analysis involves two fuzzy matching steps. We match arXiv papers with the Microsoft Academic Graph (MAG), a publications database, on titles. This gives us access to additional information about the papers in arXiv, such as the outlet where they were published, if they were published through an outlet (this includes conference proceedings, an important dissemination channel in computer science), their citation counts, authors, and in particular their institutional affiliation. We then match institutional affiliations with the Global Research Identifier Database (GRID), an open database of research institutions with information about their location and character (e.g. whether they are an educational, government or third-sector organisation, or a private sector company).[7] This matching process leaves us with 2.7 million unique paper-author pairs with detailed institutional and geographical information (including the geographical coordinates of each institution).

Semantic analysis

We undertake four streams of semantic analysis with the abstracts available in the arXiv data.

First, we use an expanded keyword search to identify AI and AI-related papers in the arXiv corpus. Full details of this analysis are available in Stathoulopoulos and Mateos-Garcia (2019). In summary, we expand an initial seed list of keywords related to AI with those that are semantically close (.ie. appear in similar contexts) in a vector space estimated with the word2vec algorithm (Mikolov, Yih, and Zweig 2013). We then tag as 'AI' those papers where those keywords appear after removing uninformative keywords (i.e. those that appear frequently in the whole corpus). This way, we identify just over 72,000 AI papers in the corpus. Manual validation of a random sample of observations suggests good classification performance, with 90 per cent precision and 90 per cent recall.

We are interested in reporting and comparing differences between research disciplines in our analysis but the taxonomy provided by arXiv has too many elements to do that easily, and papers are in any case labelled with multiple categories, hindering their classification. To address this, we cluster arXiv categories based on their co-occurrence in papers using the Louvain community detection algorithm, resulting in 25 research fields. We then create a labelled dataset of papers with categories in a single field and train a multi-label classification model to predict those categories. This gives us a vector of probabilities for each paper where every value indicates the probability that a paper belongs to a field. We classify each paper into its top field according to the probabilities predicted by the model, noting that our analysis could be expanded to consider explicitly the interdisciplinary nature of papers based on this classification exercise.

We use topic modeling to obtain a detailed understanding of the composition of AI research and loosely associate the topics extracted through this analysis with the notion of research trajectories – consistent collections of ideas potentially capturing methods, tools, analytical and technical strategies and application areas for AI research. Through our analysis of their evolution, linkages and drivers, we seek a policy-relevant sense of the direction of AI research and the interests shaping it. We estimate these topics with topSBM, a topic modelling algorithm that exploits the network structure of text corpora (the fact that it is possible to draw bipartite graphs of topics based on their co-occurrence in documents, and networks of documents based on the words that co-occur in them) in order to extract topics (communities of keywords in the aforementioned network) and estimates the weight of each topic in a document (Gerlach, Peixoto, and Altmann 2018).

TopSBM has some important advantages over other popular topic modelling such as Latent Dirichlet Allocation. It makes less stringent assumptions about the distribution that generates the data, automatically selects a suitable number of topics and generates a hierarchy of topics in the data at different levels of detail. We focus our analysis on the lowest level of resolution, with 290 topics. In order to facilitate reporting and interpretation later, we cluster these topics into higher-level aggregates using community detection on a binary topic co-occurrence matrix, which we label manually. One limitation of topSBM is that in its current implementation it is difficult to scale up to large corpora of text so we train it in a random sample of 25,000 AI papers (over a third of the population of AI papers that we have identified), and focus our analysis on those papers.

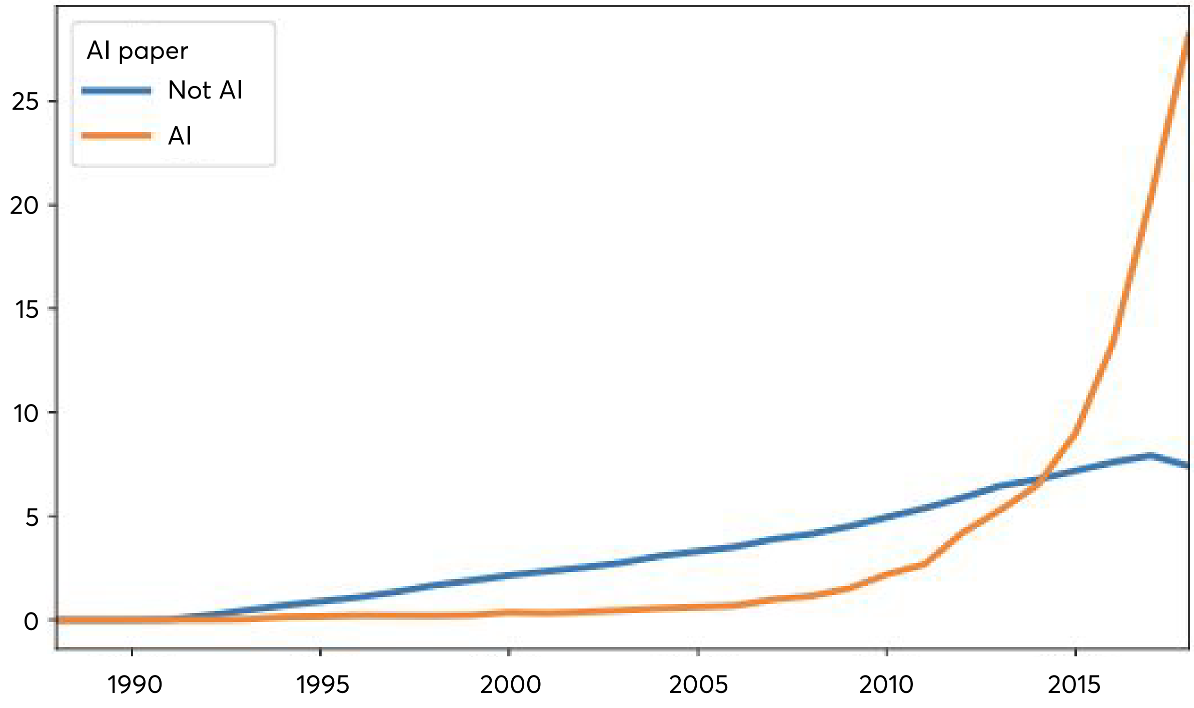
3. Results: the state of play in AI open research

We begin our analysis by studying the presence and evolution of AI activity in arXiv and its diffusion into other scientific fields beyond computer science and statistics (thus testing the idea that AI is an 'invention in the methods of invention' with broad applicability to scientific research and development (R&D) problems). We also want to measure qualitative changes: how has the composition of the AI field changed with the arrival of deep learning? Do we see evidence of a 'paradigm shift' as AI researchers and developers adopt new techniques able to solve problems that previous (symbolic and statistical) approaches were less suitable for? And how has disruption in the thematic content of AI research been associated with disruption in its geography?

Evolution of activity

Figure 2 confirms the idea of rapid expansion in the levels of AI activity in arXiv, particularly since the mid 2010s. Seventy-seven per cent of all the AI papers in our corpus have been published since 2014. The rate of activity in AI has grown much faster than the rest of the corpus, confirming that our findings are not simply driven by an increase in the popularity of arXiv as an outlet for research dissemination, or even of STEM disciplines as an area of research.

Figure 2: Evolution of AI activity in arXiv

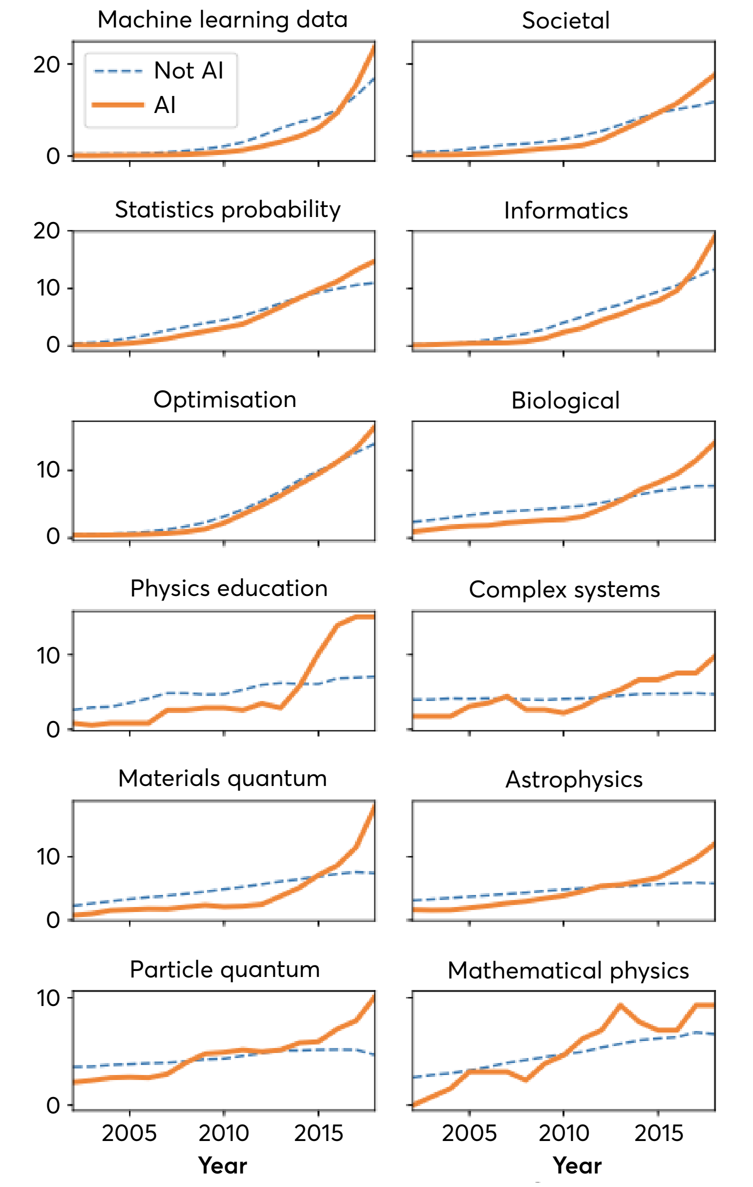


Diffusion of activity

The increase of AI activity is not confined to computing and data related disciplines - when we measure the share of AI papers in different fields, we find an increase not only in the 'machine learning and data' topic community but also in other fields, including physics, biology and materials science.

As Figure 3 shows, AI research activity has grown faster than the average in each field, consistent with the idea that AI is an 'invention in the methods of invention' that could revolutionise how R&D is conducted, for example by enabling the analysis of larger datasets and the automated exploration of more hypotheses. We find many interesting examples of AI applications outside of computing, from predictive models of solar radiation in astrophysics to optimisations of radiotherapy treatment in the medical sciences and the modelling of molecular dynamics in materials science.

Figure 3: AI activity by research field

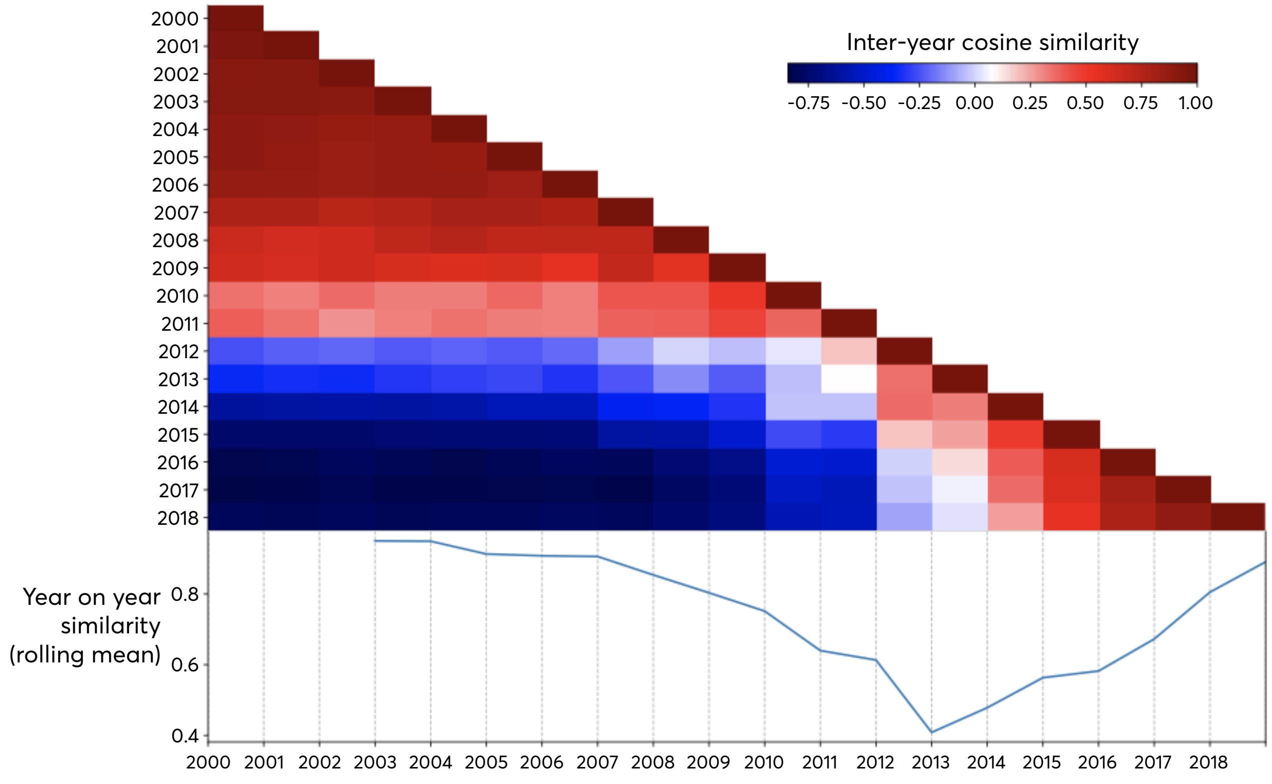


Structural change

Recent years have not just seen a quantitative expansion in the levels of AI activity, but also changes in its composition with a revival of interest in neural networks since the early 2010s (LeCun, Bengio, and Hinton 2015). 2012 in particular was a watershed moment for the field when a deep neural network significantly outperformed alternative methods in the ImageNet image classification competition (Krizhevsky, Sutskever, and Hinton 2012). Do we see this change in the data?

We begin to explore this question by calculating year-on-year semantic similarities in the composition of the field using the topic vectors we created through topic modelling.[8] In short, this involves aggregating, for each year, the topic vectors of all papers in the year and standardising them to control for secular increases in total levels of activity. We then calculate the pairwise cosine similarity between vectors in different years. Figure 4 presents results in the top panel.

Figure 4: Structural change in AI research



The colour of each cell represents the similarity between a year and those after it. Darker reds imply similar topic compositions and darker blues dissimilar compositions.

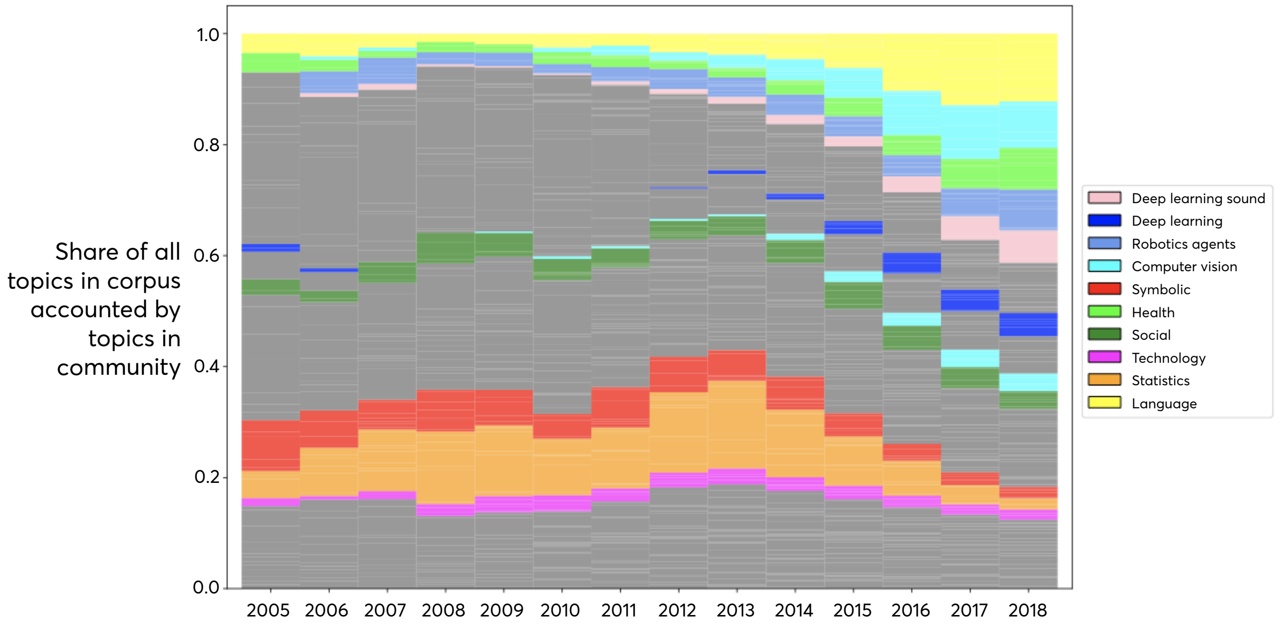
We see that each year is identical to itself (the diagonal). In the earlier period, each year tends to be quite similar to the years after, suggesting incremental, gradual changes in the evolution of the field.

Between 2010 and 2014 we detect a sudden discontinuity in the semantic composition of the field, in line with the idea of a 'revolutionary event': 2011 is not very similar to 2012 or 2013. This break in the evolution of the field is also visible in the line chart in the bottom panel, where we show the three-year rolling average of semantic similarities for every year (in other words, its average thematic similarity to the years around it). This series drops around 2012 and then starts to increase, suggesting an inflexion point in the field with the emergence of a new research trajectory that starts to stabilise and eventually to develop more incrementally.

What has driven these changes in the composition of AI research?

Figure 5 presents the relative importance of various topic communities amongst all topics in AI research (that is, how often a topic appears in AI appearances normalised by total topic presence in the corpus).

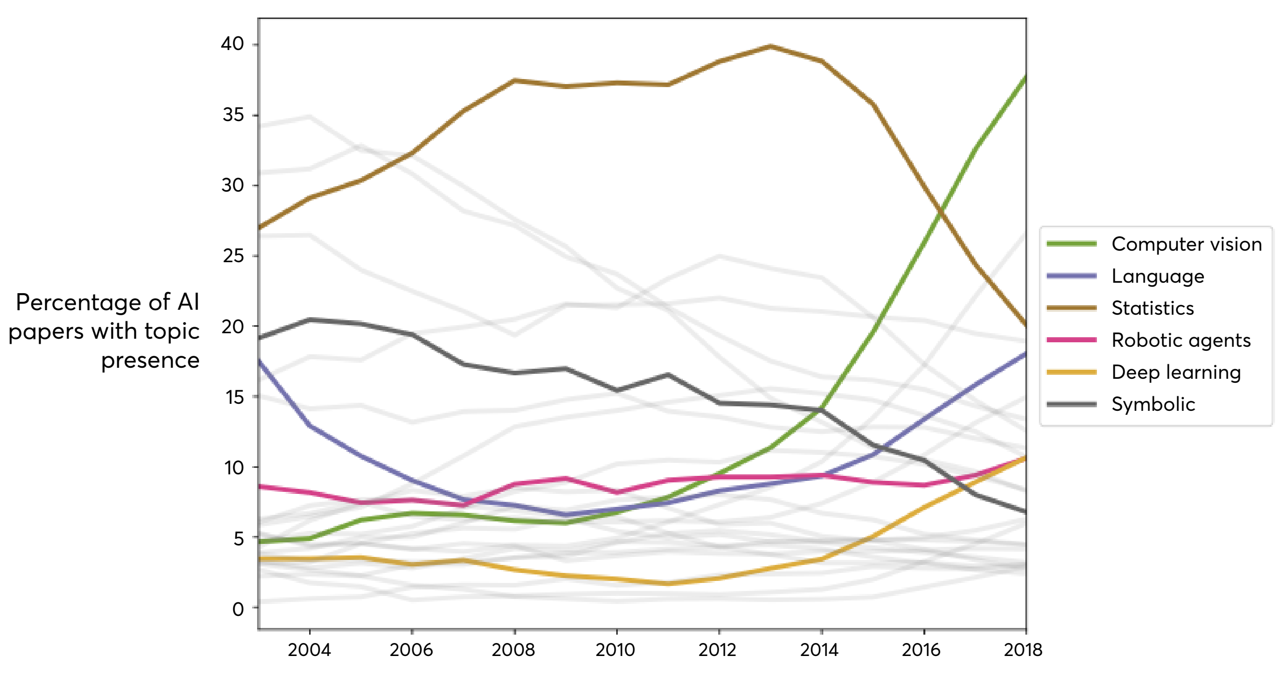
Figure 5: Evolution in the composition of AI research (topic distribution)



We see a clear decline in the relative importance of symbolic approaches since the beginning of the period and of statistical machine learning since the early 2010s. Meanwhile, computer vision, deep learning, and robotics and agents (which includes topics related to reinforcement learning) start growing since the early 2010s, consistent with the narrative of a new paradigm in AI research and the evidence of a discontinuity presented in Figure 3. We note with interest that the computer language and visions topics already had some presence the 2000s and only regained importance in the 2010s.

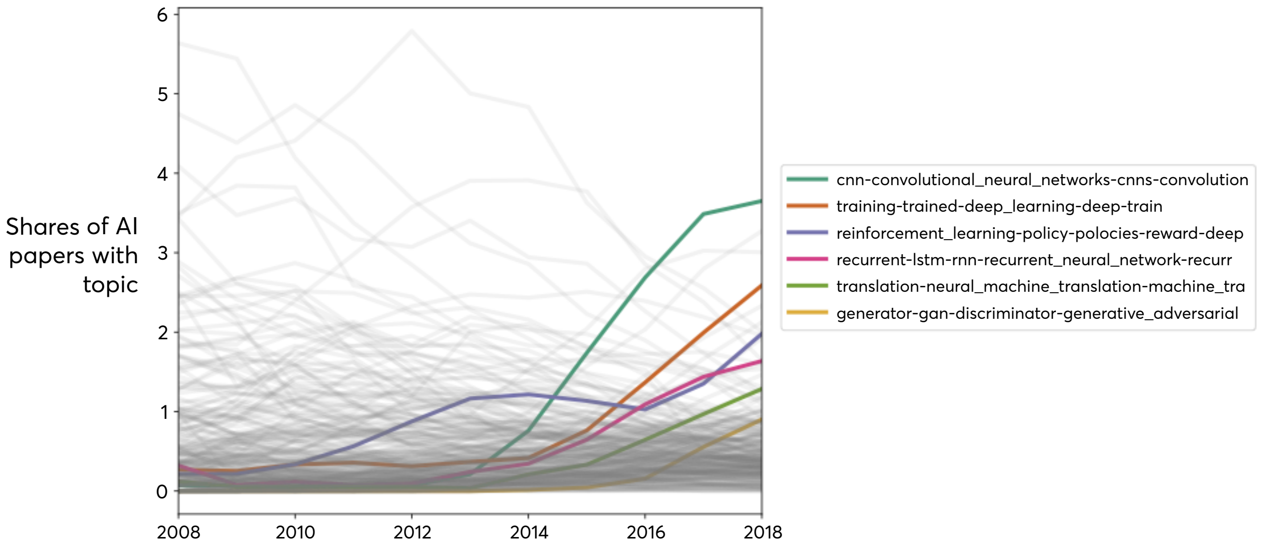
One potential issue with the figure above is that it does not take into account the secular increase in the number of topics in AI research or variations in the significance of different topics. To address that, Figure 6 shows the number of papers where topics in various topic communities have some significance. The figure confirms the results of figure 5, while highlighting additional patterns of interest: for example, we see that the computer language topic field languished during the 2000s and only regained importance in the 2010s. This coincides with the arrival of deep learning algorithms that have created significant breakthroughs in text classification, translation, captioning and generation (Young et al. 2018).

Figure 6: Evolution in the composition of AI research (topic presence in papers)



We can also study the evolution of more detailed topics. Figure 7 focuses on some key topics in what we will refer to as 'state of the art' topics in modern AI research. They include convolutional neural networks that have become the workhorse of modern computer vision research (Voulodimos et al. 2018), deep learning methods, reinforcement learning algorithms that have greatly contributed to important milestones in AI game-playing (Arulkumaran et al. 2017), recurrent networks used in language modelling, translation based on deep learning (Arulkumaran et al. 2017; Young et al. 2018), and generative adversarial networks that compete with each other to generate synthetic data with a high degree of verisimilitude (Goodfellow et al. 2014). We see rapid increases of activity in all these topics since the early 2010s. In the rest of the report, we combine these topics in some cases into a state of the art category that represents key techniques in modern AI research.

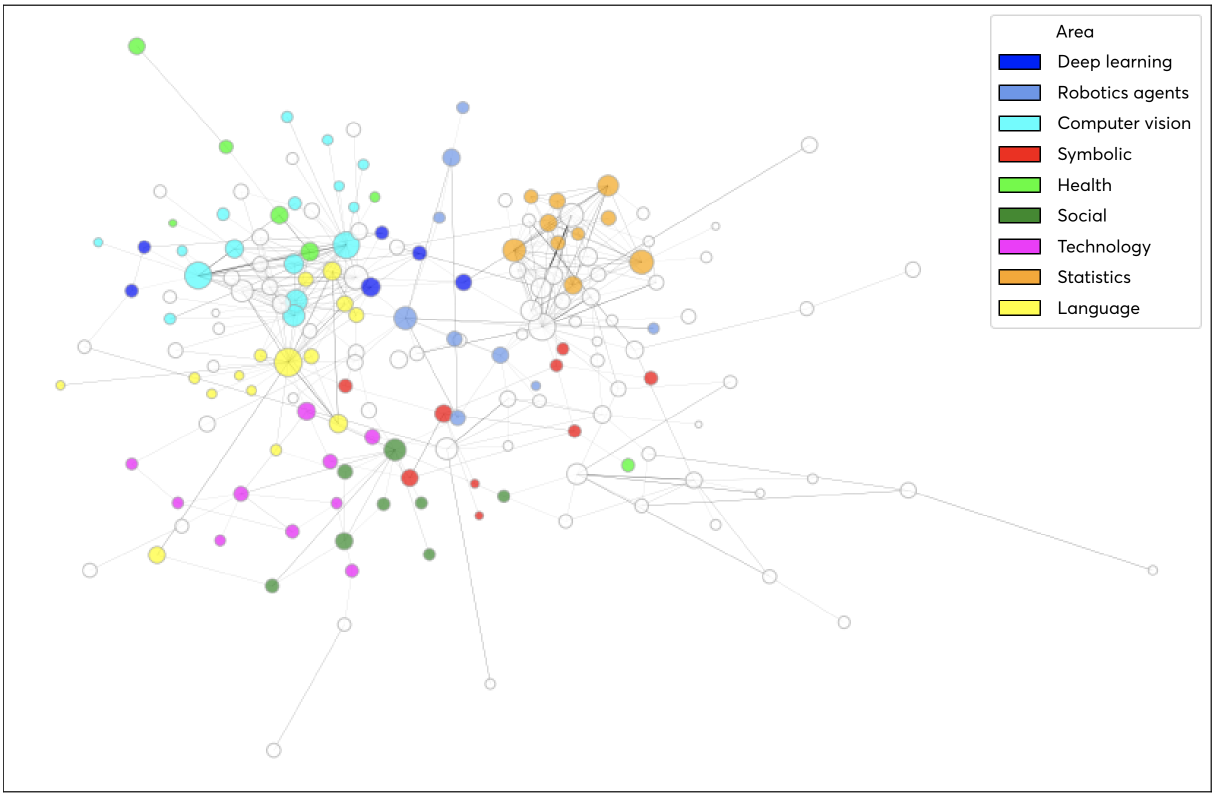
Figure 7: Evolution in the composition of AI research (detailed topics)



We can also explore the structure of AI research and its recent evolution by visualising it as a topic co-occurrence network displaying relations between topics and their centrality. Here, we are particularly interested in measuring the connectivity between novel and older topics in order to understand whether the new AI paradigm is building on previous approaches or disconnected from them.

In Figure 7, each node represents a topic and the edges between nodes are instances where topics co-occur in papers during the whole period. We have coloured some topics of interest based on the topic community they belong to. The size of the nodes is proportional to the number of papers where the topic appears.

Figure 8: Topic co-occurrence network (whole corpus)



The network seems to be split between a cluster of deep learning related AI research topics about computer vision, language and deep learning in the left, and topics related to statistical machine learning and symbolic methods to the right. One interpretation is that there is limited flow of knowledge and ideas between both communities, consistent with the idea of a 'break' in the evolution of AI research.

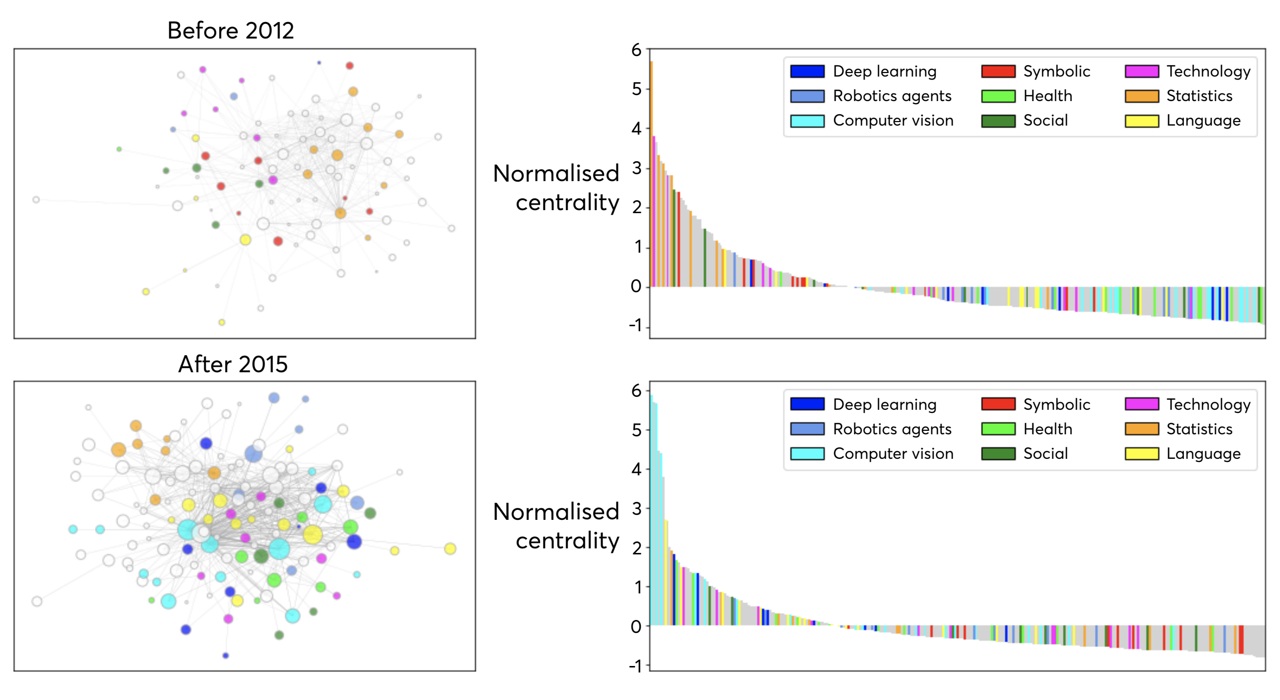
In the bottom of the graph we find a cluster of application domains including social media and technology. Interestingly, some of these are connected to symbolic methods, which could capture historical use of those methods in practical applications of AI, or perhaps the fact that some features of those methods – such as explainability – are valuable when developing real-world AI applications. Health applications are closer to the modern AI cluster because they often use computer vision algorithms to analyse medical scan data (Miotto et al. 2017).

Robotics and agents topics appear as bridges between various topic communities. We believe that this stems in part from this community's composition, including connectionism-related methods such as reinforcement learning and more broadly defined robotics topics that have also been pursued with symbolic and statistical approaches.

Figure 9 considers changes in the structure of the topic co-occurrence network between an initial period involving papers published before 2010 and a later period after 2015. The network graphs on the left column are interpreted in the same way as Figure 8.

The bar plots in the right column show the eigenvector centrality of the nodes (topics) in the network, coloured by the topic community they belong to with the same colour scheme as before. The eigenvector centrality of each node is based on the number of connections it has with other highly connected nodes and therefore captures its 'influence' (which here we interpret as its importance as a widely-adopted technique or widely-targeted application domain R&D&I during the period being considered).

Figure 9: Changes in the topic co-occurrence network

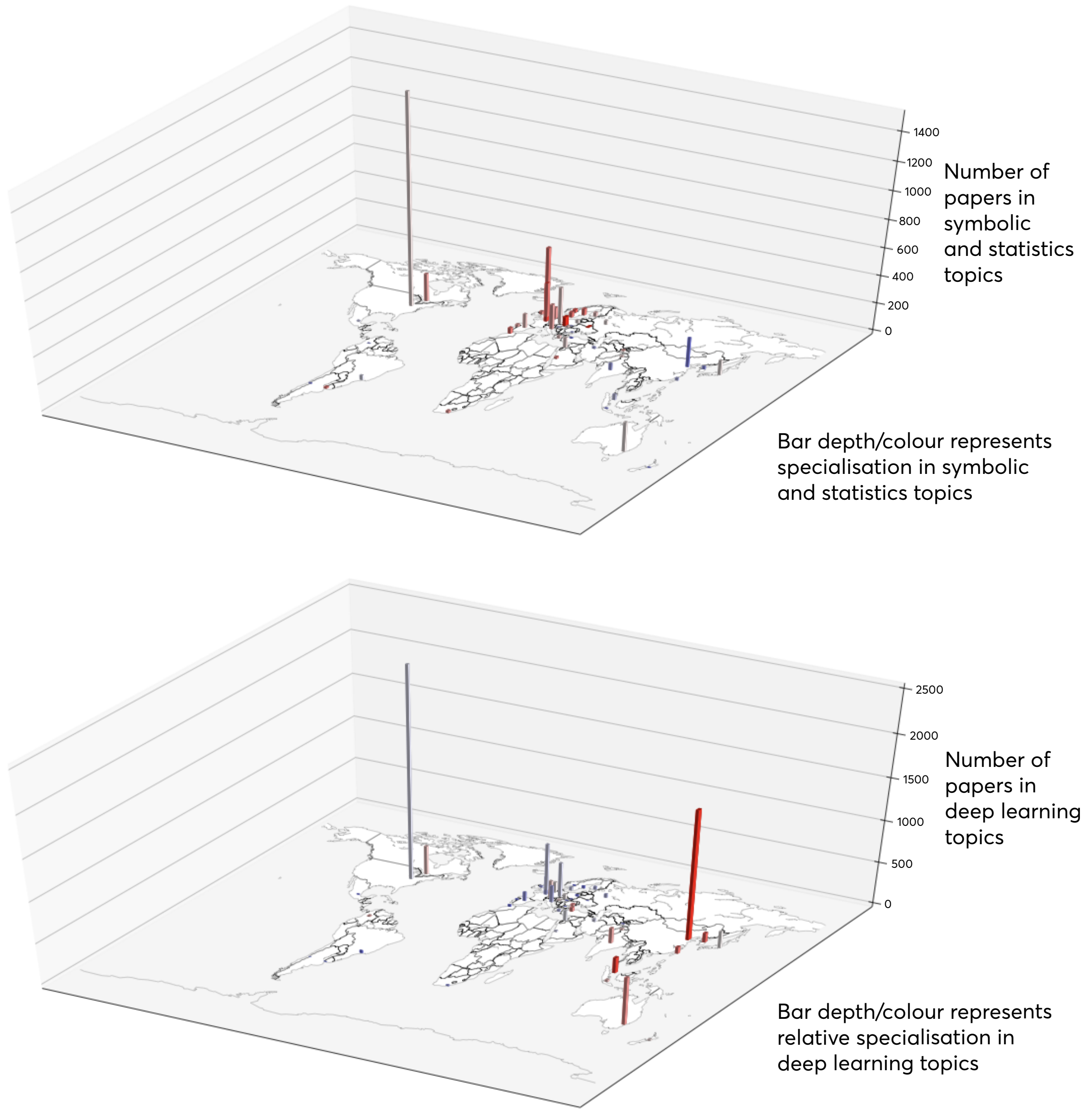


Geographical change

We conclude our analysis of the 'state of play' of AI research in arXiv by considering its geographical evolution: have increases of activity in AI and changes in its thematic composition been associated to shifts in its geography?

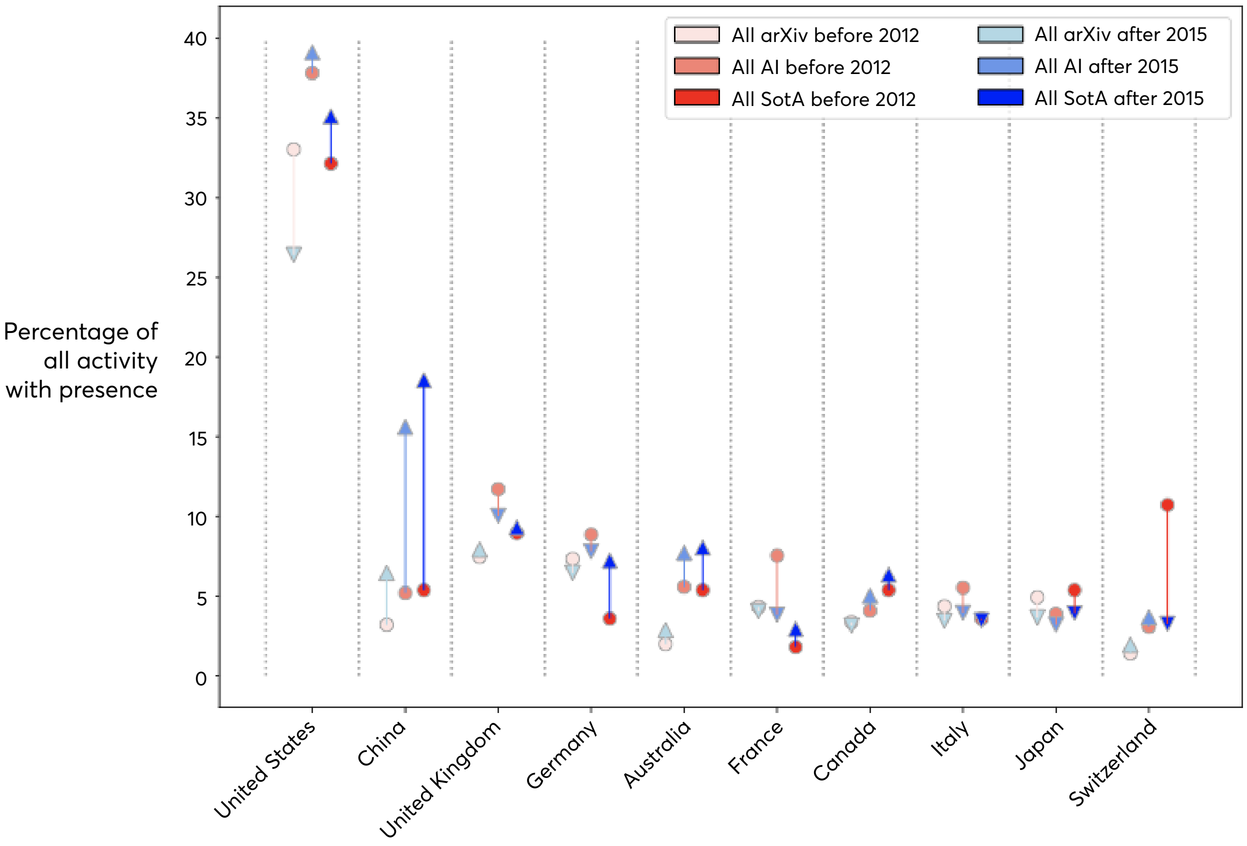
The two 3D maps in Figure 10 show the level of AI research activity in 'classical' topics related to symbolic and statistical methods (in the top) and topics related to deep learning (in the bottom). Its results are consistent with findings of previous research where we provided evidence for China's comparative advantage in AI research. By comparison, European Union (EU) countries appear relatively specialised in classical and symbolic methods.

Figure 10: National research activity in various AI topics



The results in figure 11 could be partly explained by compositional changes (i.e. the fact that China joined the AI research field more recently, when the focus of activity had shifted to deep learning related topics). In Figure 12 we try to account for this by comparing a country's share of activity in all of arXiv, all AI research and State of the Art (SotA) AI topics in the period before 2012 and the period after 2015, focusing on the top ten countries by total levels of AI activity. We want to measure changes in countries' importance in each of these fields and compare volatility across fields.

Figure 11: Changes in shares of activity in AI research for top ten countries



The figure shows that the US is dominant in the three fields. While its relative importance in the overall arXiv corpus has declined as other countries start publishing more research there, its importance in AI research and in SotA topics in this area has increased over time. China has experienced rapid growth in recent years, almost trebling its participation in AI research – especially in SotA (deep learning) topics. We do not observe a comparable increase in China's arXiv general activity, supporting the idea that it has a strategic focus (or revealed comparative advantage) in AI and especially SotA AI topics.

Changes in other countries have been less drastic. AI and SotA topics are slightly overrepresented in the United Kingdom, Australia and Canada while other EU countries such as France, Germany and Italy are underrepresented in these. Having said this, both Germany and France have increased their presence in the AI SoTA topics, suggesting a thrust to catch up in cutting edge areas of AI research.

We have also measured the geographical volatility of AI research by calculating the variance in national representation for the ten countries considered above, once again distinguishing between arXiv activity overall (the baseline), AI research and research in SotA topics. This analysis reveals higher variance of growth rates in AI and especially SotA topics compared to the arXiv benchmark (the respective variances and growth rates are 0.48, 0.75 and 0.16).[9] This suggests that some of the disruption AI topics that we have documented in this section are also manifested in changes in its geography. Determining the relationship between both vectors of change will be an important topic for future research.

4. Conclusions

Implications

Our analysis of AI research trends show a field that is being revolutionised by a revival of interest in neural network techniques and in particular, deep learning. It is hard to think of another area of science that has been so thoroughly overturned over such a short period of time, and where the translation of novel findings into practical application has been quicker. These shifts underscore the need for Artificial Intelligence (AI) maps that consider the composition of AI research activity: similar growth rates in levels of AI activity between two countries could mask significant differences if one of them concentrates on symbolic methods while another specialises in deep learning approaches.

The transformation that we have evidenced is a consequence of deep learning's success in a variety of domains, ranging from computer vision to language modelling and game playing. However, the dramatic rate of change we observe, the rapid stabilisation of the field on its new trajectory, and the low levels of overlap between current AI research and previous statistical and symbolic approaches, could raise concerns about premature lock-in and a loss of diversity in the field. As more researchers join the new paradigm, it can build a momentum of its own, driven by network effects as much as scientific and technological performance. There is still much uncertainty about the limitations and risks of new AI techniques, so it may be desirable to preserve a plurality of approaches – as Canada did when it continued supporting research in neural networks in the 1980s, when many other countries abandoned it, disappointed by its lack of progress. This provided the foundation for the deep learning revolution that we are witnessing today. Funding programmes to encourage collaboration between research communities working with connectionist methods and other techniques that are less data hungry, more explainable and more robust, could also help build AI systems bringing together the best of both worlds.

Next steps

The analysis that we have presented in this report is descriptive and focused on recent research trends in AI research. As a next step, we will publish discrete analyses of the drivers of these trends, including researcher diversity, participation of corporations in AI research, the regional distribution of AI research and its link with automation, and evolution of activity in controversial 'dual-use' surveillance AI technologies. We will also consider in further detail the policy implications of our analysis.

Going further, it will be important to consider other data sources in this work beyond arXiv, including research activity in traditional scientometric databases, as well as patenting, open source software development and business activity to name a few. Doing this will allow us to validate findings based on this experimental data source, and to understand how AI research is diffusing from laboratories into application and impact. We would also like to further develop our analysis of AI research trajectories, paying more attention to how over time multiple topics become part of, or splinter from, a trajectory. A longitudinal analysis that considers how topics co-evolve, merge and branch would provide us with a richer understanding of AI research trajectories. Here, it would be particularly interesting to distinguish more robustly between theoretical and applied contributions, perhaps using the full text of papers and other relevant information such as the data, diagrams and figures that they use.

Our analysis also fails to consider the actual goals or purposes of research, a critical component of the analysis of directionality. Although richer, full-text data might give us a better understanding of the intended goals of a research paper or AI system (e.g. predictive performance, robustness, explainability, safety, labour automation or labour augmentation etc), this will have to be complemented with qualitative assessments involving non-technical experts, who might have to face the systems in real-world situations. This is an example of the kind of mixed-methods research opportunities made possible by the granular data we are using here.

As this discussion shows, smarter data about smarter machines offer many opportunities to advance our understanding of the development and diffusion of novel technologies such as AI, and to inform policies seeking to ensure that the benefits of those transformations are widely shared.

References

Acemoglu, Daron, and Pascual Restrepo. 2018. 'Artificial Intelligence, Automation and Work.' National Bureau of Economic Research.

Aghion, Philippe, Paul A. David, and Dominique Foray. 2009. 'Science, Technology and Innovation for Economic Growth: Linking Policy Research and Practice in "STIG Systems."' Research Policy, Special Issue: Emerging Challenges for Science, Technology and Innovation Policy Research: A Reflexive Overview, 38 (4): 681–93.

Agrawal, Ajay, John McHale, and Alex Oettl. 2018. 'Finding Needles in Haystacks: Artificial Intelligence and Recombinant Growth.' 24541. National Bureau of Economic Research. http://www.nber.org/papers/w24541

Amodei, Dario, and Danny Hernandez. 2018. AI and Compute.

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan ManÃ©. 2016. 'Concrete Problems in AI Safety.' arXiv Preprint arXiv:1606.06565.

Arthur, W. B. 1999. 'Complexity and the Economy.' Science 284 (5411): 107–9.

Arthur, W. Brian. 1994. Increasing Returns and Path Dependence in the Economy. University of Michigan Press.

Arulkumaran, Kai, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. 2017. 'A Brief Survey of Deep Reinforcement Learning.' arXiv Preprint arXiv:1708.05866.

Bakhshi, Hasan, and Juan Mateos-Garcia. 2016. 'New Data for Innovation Policy.' Nesta Working Paper Series. London: Nesta.

Bostrom, Nick. 2017. Superintelligence. Dunod.

Brian Arthur, W. 2014. Complexity and the Economy. Oxford University Press, USA.

Brundage, Miles. 2016. 'Modeling Progress in AI.' In AAAI Workshop: AI, Ethics, and Society.

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, and Bobby Filar. 2018. 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.' arXiv Preprint arXiv:1802.07228.

Brynjolfsson, Erik, and Andrew McAfee. 2014. The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. W. W. Norton & Company.

Buolamwini, Joy, and Timnit Gebru. 2018. 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.' In Conference on Fairness, Accountability and Transparency, 77–91.

Burgess, Lucie C., David Crotty, David de Roure, Jeremy Gibbons, Carole Goble, Paolo Missier, Richard Mortier, Thomas E. Nichols, and Richard O'Beirne. 2016. 'Alan Turing Institute Symposium on Reproducibility for Data-Intensive Research–Final Report.' St. Hugh's College, Oxford.

Cantner, Uwe, and Simone Vannuccini. 2018. 'Elements of a Schumpeterian Catalytic Research and Innovation Policy.' Industrial and Corporate Change 27 (5): 833–50.

Centre, Joint Research. 2018. 'Artificial Intelligence: A European Perspective.' Seville: JRC.

Cockburn, Iain M., Rebecca Henderson, and Scott Stern. 2018. 'The Impact of Artificial Intelligence on Innovation.' National Bureau of Economic Research.

David, Paul A. 1985. 'Clio and the Economics of QWERTY.' The American Economic Review 75 (2): 332–37.

Dosi, Giovanni. 1982. 'Technological Paradigms and Technological Trajectories: A Suggested Interpretation of the Determinants and Directions of Technical Change.' Research Policy 11 (3): 147–62.

Drexler, K. Eric. 2019. Reframing Superintelligence. FHI Technical Report, https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf

Dyson, George. 2012. Turing's Cathedral: The Origins of the Digital Universe. Penguin UK.

Elsevier. 2018. 'Artificial Intelligence: How Knowledge Is Created, Transferred, and Used.' Elsevier.

Eubanks, Virginia. 2018. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press.

Ford, Martin. 2015. Rise of the Robots: Technology and the Threat of a Jobless Future. Basic Books.

Furman, Jason, and Robert Seamans. 2018. 'AI and the Economy.' ID 3186591. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3186591

Garud, Raghu, and Peter KarnÃže. 2001. 'Path Creation as a Process of Mindful Deviation.' Path Dependence and Creation 138.

Gerlach, Martin, Tiago P. Peixoto, and Eduardo G. Altmann. 2018. 'A Network Approach to Topic Models.' Science Advances 4 (7): eaaq1360.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. Deep Learning. MIT press.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. 'Generative Adversarial Nets.' In Advances in Neural Information Processing Systems, 2672–80.

Index, A. I. n.d. 'AI Index.' 2018. https://aiindex.org

2017. "The Artificial Intelligence Index: 2017 Annual Report." http://cdn.aiindex.org/2017-report.pdf

Intellectual Property Office. n.d. 'Artificial Intelligence: A Worldwide Overview of AI Patents and Patenting by the UK AI Sector.' Intellectual Property Office: Newport. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/817610/Artificial_Intelligence_-_A_worldwide_overview_of_AI_patents.pdf

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. 'The Global Landscape of AI Ethics Guidelines.' Nature Machine Intelligence, September, 1–11.

Kattel, Rainer, and Mariana Mazzucato. 2018. Mission-Oriented Innovation Policy and Dynamic Capabilities in the Public Sector. Oxford University Press.

Klinger, J., J. Mateos-Garcia, and K. Stathoulopoulos. 2018. 'Deep Learning, Deep Change? Mapping the Development of the Artificial Intelligence General Purpose Technology.' arXiv Preprint arXiv:1808.06355.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. 'Imagenet Classification with Deep Convolutional Neural Networks.' In Advances in Neural Information Processing Systems, 1097–1105.

Kuhn, Thomas S. 2012. The Structure of Scientific Revolutions. University of Chicago press.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. 'Deep Learning.' Nature 521 (7553): 436–44.

Mann, Katja, and Lukas PÃŒttmann. 2017. 'Benign Effects of Automation: New Evidence from Patent Texts.'

Marcus, Gary. 2018. 'Deep Learning: A Critical Appraisal.' arXiv Preprint arXiv:1801.00631.

Marcus, Gary, and Ernest Davis. 2019. Rebooting AI: Building Artificial Intelligence We Can Trust. Pantheon.

Markoff, John. 2016. Machines of Loving Grace: The Quest for Common Ground between Humans and Robots. HarperCollins Publishers.

Mateos-Garcia, Juan C. 2018. 'The Complex Economics of Artificial Intelligence.' Available at SSRN 3294552.

Mazzucato, Mariana. 2015. The Entrepreneurial State: Debunking Public vs. Private Sector Myths. Anthem Press.

2018. 'Mission-Oriented Innovation Policies: Challenges and Opportunities.' Industrial and Corporate Change 27 (5): 803–15.

McAfee, Andrew, and Erik Brynjolfsson. 2017. Machine, Platform, Crowd: Harnessing Our Digital Future. WW Norton & Company.

Mikolov, Tomas, Wen-Tau Yih, and Geoffrey Zweig. 2013. 'Linguistic Regularities in Continuous Space Word Representations.' In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 746–51. Atlanta, Georgia: Association for Computational Linguistics.

Miotto, Riccardo, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T. Dudley. 2017. 'Deep Learning for Healthcare: Review, Opportunities and Challenges.' Briefings in Bioinformatics 19 (6): 1236–46.

Myers West, Sarah, Meredith Whittaker, and Kate Crawford. 2019. 'Discriminating Systems: Gender, Race, and Power in AI.' New York: AI Now Institute. https://ainowinstitute.org/discriminatingsystems.pdf

Nesta. 2019. 'Innovation Mapping Now.' London: Institution. https://media.nesta.org.uk/documents/Innovation-Mapping-Now-March-2019.pdf

Noble, Safiya Umoja. 2018. Algorithms of Oppression: How Search Engines Reinforce Racism. nyu Press.

Peng, Roger D. 2011. 'Reproducible Research in Computational Science.' Science 334 (6060): 1226–27.

Prediger, Lukas. 2017. 'On the Importance of Monitoring and Directing Progress in AI." AI Matters. https://doi.org/10.1145/3137574.3137583

Restrepo, Pascual, and Daron Acemoglu. 2018. "The Wrong Kind of AI?'

Rolnick, David, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, and Anna Waldman-Brown. 2019. 'Tackling Climate Change with Machine Learning.' arXiv Preprint arXiv:1906.05433.

Rosenberg, Nathan, and Rosenberg Nathan. 1982. Inside the Black Box: Technology and Economics. Cambridge University Press.

1994. Exploring the Black Box: Technology, Economics, and History. Cambridge University Press.

Russell, Stuart. 2019. Human Compatible: AI and the Problem of Control. Penguin UK.

Stathoulopoulos, Konstantinos, and Juan C. Mateos-Garcia. 2019. 'Gender Diversity in AI Research.' Available at SSRN 3428240.

Stirling, Andy. 2009. 'Direction, Distribution and Diversity! Pluralising Progress in Innovation, Sustainability and Development.'

2014. 'Towards Innovation Democracy? Participation, Responsibility and Precaution in Innovation Governance.' ID 2743136. Rochester, NY: Social Science Research Network. http://papers.ssrn.com/abstract=2743136

Teece, David J. 2008. 'Dosi's Technological Paradigms and Trajectories: Insights for Economics and Management.' Industrial and Corporate Change 17 (3): 507–12.

Topol, Eric J. 2019. 'High-Performance Medicine: The Convergence of Human and Artificial Intelligence.' Nature Medicine 25 (1): 44.

Trajtenberg, Manuel. 2018. 'AI as the next GPT: A Political-Economy Perspective.' National Bureau of Economic Research.

Voulodimos, Athanasios, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. 2018. 'Deep Learning for Computer Vision: A Brief Review.' Computational Intelligence and Neuroscience 2018.

Young, Tom, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. 'Recent Trends in Deep Learning Based Natural Language Processing.' IEEE Computational Intelligence Magazine 13 (3): 55–75.

Zuboff, Shoshana. 2019. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. Profile Books.

Endnotes

1. See for example this open database of global AI initiatives: https://www.nesta.org.uk/data-visualisation-and-interactive/mapping-ai-governance
2. https://github.com/nestauk/arxiv_ai
3. https://arxlive.org/
4. https://arxiv.org
5. See https://openai.com/progress/#papers and https://deepmind.com/research
6. https://jack-clark.net/
7. https://www.grid.ac/
8. This approach has been inspired by an analysis of about structural change in the biofuel industry (Parraguez [forthcoming]).
9. When we exclude China from the analysis, the variances in change rates between AI and arXiv overall become quite similar, but the geography of SotA topics remains more volatile, suggesting that the geography of modern AI methods is being more strongly disrupted than older topics for AI research.

About Nesta

Nesta is a global innovation foundation. We back new ideas to tackle the big challenges of our time.

We use our knowledge, networks, funding and skills – working in partnership with others, including governments, businesses and charities. We are a UK charity but work all over the world, supported by a financial endowment.

To find out more visit **www.nesta.org.uk**

If you'd like this publication in an alternative format such as Braille, large print, please contact us at: **information@nesta.org.uk**

nesta

58 Victoria Embankment
London EC4Y 0DS

+44 (0)20 7438 2500
information@nesta.org.uk
@nesta_uk
www.facebook.com/nesta.uk
www.nesta.org.uk