Journal of Evolutionary Economics
https://doi.org/10.1007/s00191-022-00804-4

REGULAR ARTICLE

Firms' influence on the evolution of published knowledge when a science-related technology emerges: the case of artificial intelligence

Su Jung Jee[1,2,3] · **So Young Sohn**[2]

Accepted: 23 November 2022

Abstract
Firms with the assets complementary to Artificial Intelligence (AI) have actively conducted AI research and selectively published their results since AI has resurged around 2006. Focusing on the recent AI development, we investigate how and to what extent firms' deep engagement in the publication of emerging science-related technology can influence the evolution of published knowledge. Using bibliometric analyses applied to the papers in major AI conferences and journals, we find that papers with at least one author affiliated to a firm, and particularly papers with only firm-affiliated author(s), have had higher influence on the formation of published knowledge trajectory than other papers. In addition, papers from firm and non-firm (university and public research institution) collaborations show higher novelty and conventionality than other papers. These findings deepen our understanding of the role of firms in the evolution of emerging science-related technology.

Keywords Science-related technology · Firms' publishing · Evolution of published knowledge · University-industry collaboration · Artificial intelligence

JEL codes O34 · O36

✉ Su Jung Jee
s.j.jee@bradford.ac.uk

[1] Faculty of Management, Law and Social Sciences, University of Bradford, Richmond Road, Bradford BD7 1DP, UK

[2] Department of Industrial Engineering, Yonsei University, 134 Shinchon-dong, Seoul 120-749, Republic of Korea

[3] Institute for New Economic Thinking at the Oxford Martin School, University of Oxford, Oxford, UK

Published online: 07 December 2022

1 Introduction

When a science-related technology emerges,[1] firms that aim to develop this technology invest in the relevant research. Such investments are long-term and high-risk; therefore, they are typically made by firms with strong market position (Rosenberg 1990). A considerable proportion of these firms' research outcomes are accumulated in the form of firm-specific tacit knowledge (Nelson and Winter 1982), and some outcomes are patented for future appropriation (Teece 1986). At the same time, some outcomes are selectively published in the form of research papers (Hicks 1995; Simeth and Raffo 2013; Grassano et al. 2019).

However, publishing by firms is a seemingly counterintuitive behavior in that firms reveal their research outcomes to the public without direct monetary rewards. As firms do not have enough incentives to produce public knowledge with their own money (Arrow 1972), the primary contributors to published knowledge have been universities that are largely supported by the government (Mowery and Rosenberg 1999). Nonetheless, firms in some emerging sectors publish their research outcomes when they expect that they can ultimately benefit by doing so. Although the overall intensity of firm science and publications has decreased since the 1980s (Arora et al. 2018),[2] firms in some sectors still publish actively (Grassano et al. 2019).

When firms engage deeply in research publications in emerging fields, a question arises concerning the influence of the private sector's engagement in published knowledge development. If firms possess assets that are critical for developing and commercializing an emerging technology, their research papers will likely attract the attention of researchers intending to contribute to the knowledge frontier (Dasgupta and David 1987) or economic development (Etzkowitz 2002). If the attraction is significantly strong, the development of published knowledge will be naturally focused around the firms' interests, facilitating progress in the corresponding field and shaping the direction of knowledge evolution. The recent advancement of artificial intelligence (AI) is a relevant case in which firms that have published a substantial number of papers possess key complementary assets required to advance and apply emerging technologies (Anthes 2017; Agrawal et al. 2018; Mansell and Steinmueller 2020). Although it is challenging to intentionally alter the rate and direction of the evolution of an emerging technology, papers published by firms in fields, such as recent AI, can considerably influence the evolution of published knowledge.

Such an impact can be of considerable significance in influencing policies related to science and technology with respect to government support, particularly when firm papers are of high quality. One of the reasons that we can expect

[1] The importance of understanding science when developing technology has increased throughout the twentieth century (Narin et al. 1997; Fleming and Sorenson 2004), although the relationship between science and technology is mutual (Kline and Rosenberg 1986) and the degree of interaction varies hugely across fields (Pavitt 1987). To better represent the mutual relationship between science and technology rather than the one-sided reliance of technology on science, this study uses the term favored by Freeman (1997), "science-related technology" instead of "science-based technology."

[2] Following Arora et al. (2018), we define "firm papers" (or "firm publications") as research articles in which at least one of the authors is affiliated to a firm.

high quality from firm papers in the field of AI is the widespread research collaboration between private and public organizations. Firm researchers in the field of AI have increasingly collaborated with researchers in universities (and public research institutions) because of the expected potential synergy of the collaboration (Gibney 2016). Firms can decrease their long-term investment risk by effectively gaining access to the scientific knowledge base of university researchers, while university researchers can acquire the opportunity to conduct better research by leveraging the firms' resources. In this vein, firms' engagement in research publications may have high-quality impacts on the published knowledge development of the field in the long run. If this is the case, we can derive important policy implications regarding the direction of public support for relevant innovations.

However, to the best of our knowledge, relatively little scholarly attention has been paid to empirically examining how and to what extent firms' publications in emerging science-related technologies can influence the evolution of knowledge that is built on research papers (Perkmann and Walsh 2009). Moreover, empirical studies barely shed light on the features of papers produced by firm and non-firm (i.e., university or public research institutions) collaborations. Focusing on the recent development of AI, this study bridges these gaps in the literature with the following research questions.

RQ1) To what extent do firms' publications (relative to other publications) influence the formation of published knowledge trajectory relevant to emerging science-related technology?
RQ2) Are there any qualitative differences between publications by firm and non-firm collaborations and other publications on emerging science-related technology in terms of their a) novelty and b) conventionality?

Referring to Uzzi et al. (2013), papers satisfying the conditions of both high novelty and conventionality are likely to be highly important ones in the long run. Therefore, the second research question reflects the long-term quality of published papers, focusing on the qualitative characteristics. To answer the research questions, this study controls for the previous track record of author(s) and affiliation(s) associated with each publication, as well as paper-specific characteristics. We employ statistical models, along with data-mining approaches such as community detection methods, main path analysis, and dynamic topic modeling on the data obtained from AI-related conference and journal papers. The findings of this study are expected to enhance our understanding of firms' publishing on emerging science-related technologies and thereby provide a more detailed picture of the role of firm and non-firm actors in the evolution of emerging science-related technology.

The rest of this paper is structured as follows. In Section 2, we provide a background of the literature and rationalize why the recent development of AI fits our context. Section 3 describes the research methodologies and data collection, while Section 4 discusses the findings. Lastly, Section 5 presents the study's contributions and conclusions.

2 Conceptual background

2.1 Firms' investment in an emerging science-related technology

In the post-war period of the twentieth century, research on high-risk science-related technologies was largely conducted by the corporate laboratories of technological giants such as AT&T, Du Pont, GE, and Xerox (Pisano 2010). These capital-intensive organizations invested heavily in internal research and concentrated their efforts on developing technologies that were expected to create huge long-term value despite the great uncertainties involved. The expected benefits from these risky investments include first-mover advantages such as learning experiences, buyer-switching costs, and barriers to the entry of other firms (Rosenberg 1990). Even when a firm does not benefit much from the first-mover advantage, the absorptive capacity obtained from such an investment helps the firm effectively monitor and understand research conducted by external actors (Cohen and Levinthal 1990).

Differing from this conventional pattern, there have been some cases in which small firms invest in research concerning high-risk technologies. When biotechnology emerged in the 1980s, the major actors that conducted innovation research were young entrepreneurial firms, particularly academic spin-offs (Rothaermel and Thursby 2007). This was possible because compared with research outcomes from other fields, those from biotechnology, in general, are readily appropriable. There were a number of venture capitalists, wealthy individuals, and large firms that financed the research of small firms. Similar patterns have been observed in emerging sectors such as nanotechnology and energy (Henderson 1993; Christensen and Rosenbloom 1995; Arora et al. 2018).

With the resurgence of AI starting around 2006, large firms have again been the major business actors engaged in research (Gibney 2016), as they detect AI's huge market potential—as an enabling technology (Teece 2018) or even as a general-purpose technology (Klinger et al. 2018)—applicable to various domains. Moreover, the complementary assets necessary to proceed with AI research—big data and computational power—are concentrated in firms, not in universities and public research institutions. The business models that derive value from AI are also likely to be firm-specific because an algorithm performs better when it has a narrow focus (Norton 2016). Such circumstances have incentivized firms to perform research on the emerging technology of AI (Waters 2015).

2.2 Evolutionary perspective on firms' publishing

Since the knowledge itself is not inherently private or public, the decision to make the research findings public depends on the strategic choices of firms. Firms strategically decide which parts of their research findings should be kept private or made public, and which methods of knowledge protection or disclosure should be employed for future appropriation. Firms can apply patents to protect their knowledge legally or choose to keep their knowledge as a trade secret if it is difficult to obtain a patent for the knowledge or the risk of releasing the knowledge through a

patent document is higher than the expected return. By contrast, firms can choose to publish their research outcomes purposefully through channels such as journals or conference proceedings.

Among firm-level actions that are used to manage the boundary between private and public knowledge, research publication is one of several *selective revealing* strategies through which firms intentionally make knowledge spillovers (Alexy et al. 2013). This firm behavior of revealing knowledge without direct monetary rewards is counterintuitive when we consider the free riding of external actors on the revealed knowledge (Arrow 1972). Due to this counterintuitive nature, many scholars have discussed the motivations underlying firms' publishing, such as to attract talented researchers by providing them an opportunity to get published (Simeth and Raffo 2013), to maintain links with academia (Rosenberg 1990), or to build the necessary technical reputation to exchange valuable knowledge with academia (Hicks 1995).

One of the main reasons why firms expect their publishing to serve key purposes, such as maintaining technical reputation and academic relationships, is because the published knowledge not only delivers written knowledge to the public but also indicates the unpublished aspects of firm resources (Hicks 1995). These resources can include complementary assets required to commercialize emerging technologies, such as infrastructure, data, and tacit knowledge and skills. Researchers in academic engineering may be interested in solving problems of industrial significance (Vincenti 1990). Institutional changes after the Bayh–Dole Act of 1980 and the subsequent founding of technology transfer offices in universities further encourage university researchers to contribute directly to economic and social development, which is known as the third mission of universities (Etzkowitz 2002). Therefore, papers published by firms that possess resources critical to an emerging technology's commercialization are likely to draw researchers' attention to these firms.

Thus, it is expected that when a firm that possesses assets complementary to an emerging technology actively engages in publishing, the influence of its papers on the evolution of published knowledge are considerable. Similar observations have been made in that the average impact of papers published by leading firms is outstanding in fields such as biotechnology, physical science, electrical engineering, and recently, AI (Hicks 1995; Narin et al. 1997; Hartmann and Henkel 2019; McKelvey and Rake 2020; Baruffaldi and Poege 2020). This is an interesting aspect considering that the overall evolution of published knowledge is directed by the collected interests of individual researchers mostly affiliated with universities that are largely financed by the government. Given the conventional understanding that research publications are mostly driven by governmental support (Mowery and Rosenberg 1999), the active engagement of firms in research and subsequent publishing leads to an important policy question: where to direct governmental support?

According to the literature on *technological systems* for innovation, the components constituting a technological system interact with each other, revealing a variety of problems and sources of knowledge needed to innovate (Carlsson and Stankiewicz 1991). Hughes (1987) emphasized that when technologies become increasingly embedded in a social structure, they stimulate *reverse salients* that fall behind the overall evolution of the system. Dealing with these sluggish components in the production and diffusion of a new technology requires the involvement of

different types of institutions, including private and public actors (Antonelli 2001). The engagement of various types of institutions helps maintain diversity in technological systems, which is necessary to cope with potential obstacles in innovation.

From this viewpoint, firms' engagement in research and publishing in an emerging field implies that the government might need to direct its support towards areas complementary and nonredundant to the firms' investment. When leading firms invest heavily in the early stages of a field's knowledge development, the field can lose its diversity as it can be locked into the direction pursued by the leading firms. Although optimists expect that AI will soon be widely applied, it has been proven that it is not straightforward to move from generic AI models to broad applications owing to factors falling behind, from technical to non-technical ones, in each application domain (Mansell and Steinmueller 2020). Careful maneuvers in the direction of governmental support can increase the diversity of components in the overall technological system, ultimately broadening the scope for selection among competing variants and coping with potential barriers.

To extend the above discussions, this study investigates the influence of firm papers on the overall direction of published knowledge evolution, which can be characterized by the significance of the impact on the formation of a knowledge trajectory or the direction of advancement within a technological paradigm (Dosi 1982). Our approach goes beyond previous observations that leading firms' publishing can make an outstanding impact in that we evaluate whether firm papers can even lead the direction of knowledge evolution in an emerging field of AI.

2.3 Collaborative knowledge production

While large corporate labs usually conducted in-house research during the post-war period, since the 1980s, there has been an increasing trend of sharing the risks of producing new knowledge among heterogeneous organizations including firms, universities, and public research institutions. This trend is in line with the strong attention given to the open innovation literature (Chesbrough 2003) and the third mission of universities (i.e., contribution to economic and social development; Etzkowitz 2003) over the past few decades. The transition to inter-organizational and transdisciplinary research efforts was also noted by Gibbons et al. (1994) as basic properties of Mode 2 knowledge production, which is a representative way of producing new knowledge in modern society.

Within this trend, the number of university-industry (U–I) collaborations has risen significantly. Firms have enjoyed various benefits from this type of relationship. In particular, theoretical depth necessary to advance science-related technologies often must be obtained from experts in universities. However, just as firms manage the boundary between private and public knowledge, so do university researchers. University researchers can keep some of their capabilities secret to maintain their competitive advantage in the academic community (Hilgartner and Brandt-Rauf 1998). Therefore, even when a firm recognizes the valuable capability of university researchers, it is not necessarily able to exploit that capability (Hicks 1995). U–I collaborations have been extensively employed to resolve this

issue, as collaborations allow firm researchers to gain formal access to university researchers' valuable scientific knowledge. In addition, if a trust-based relationship between a university and a firm evolves over time, other valuable information beyond the boundary of formal collaborative activities can flow between the firm and university researchers. By maintaining connections with universities, firm researchers can also participate in a broader network of academia, which sometimes unexpectedly brings useful information necessary to develop technology that the firm needs.

Meanwhile, since collaboration connotes reciprocity (Katz and Martin 1997), university researchers can also benefit from the U-I collaboration. When high-quality resources necessary to improve research productivity are concentrated in firms, the U-I collaboration can be a channel that allows university researchers access to these firm resources. The resources can be equipment, data, and technological know-how, to name a few. In addition, through the relationships, university researchers can better understand key industrial problems. Such an understanding can lead university researchers to pay more attention to real-world problems when they conduct research. Considering that university researchers have been encouraged to contribute to economic development over the past 40 years (Etzkowitz 2002), they can significantly benefit from awareness of industrial issues.

Moreover, such awareness can even lead to advancement of scientific knowledge, as application-focused research may yield an unintended scientific discovery. The distinction between applied and basic research is often meaningless in the realms of science-related technologies such as computer science (CS) and medicine (Rosenberg 1990). Stokes (1997) argued that the major contributors to the advancement of science-related technology are people who address real-world problems but do not lose sight of the contribution to scientific understanding, such as Louis Pasteur who tried to solve fermentation problems in the wine industry but additionally created the science of bacteriology. Other "star scientists," as defined by Zucker et al. (2002) have also performed this type of research. Hence, university researchers who are exposed to important industrial issues through the U–I collaborations are likely to be capable of performing salient research at the interface between science and technology.

Based on the reciprocal relationship between firms and university researchers, it can be said that U–I collaborations satisfy various necessary conditions that could lead to important research outcomes in the realms of emerging science-related technology. These conditions include the scientific knowledge base of university researchers, sufficient complementary resources provided by firms, and awareness of crucial industrial issues. Therefore, we expect that papers published by firm and non-firm collaborations are characterized by significantly higher novelty and conventionality, two qualitative characteristics related to the possibility of a paper gaining importance in the long run (Uzzi et al. 2013). High novelty means the extent to which the published knowledge includes novel ideas that can be obtained by allocating resources for enough exploration, while high conventionality characterizes to what extent a paper is well grounded in the body of academic discussions.

2.4 Firms' influence on the evolution of published knowledge: the case of AI

Since we delve into the phenomenon in a specific sector (i.e., AI), it is difficult to extend our results to all other fields. To mitigate this concern, we suggest boundary conditions to extend the results of this study. This section summarizes the re-emergence of AI and describes some of its general features as a science-related technology and a type of Mode 2 knowledge. Based on the features, we derive a set of hypotheses.

Although knowledge production related to AI has only relatively recently begun to grow rapidly, the beginning of the field dates back to the mid-1950s. From the mid-1950s to the mid-1970s, optimism about humans achieving the ultimate goal of creating intelligent machines was high. However, in the face of unresolved failure in the underlying theoretical model defining AI, funding for AI temporarily stopped (Boden 2016). From the early 1980s, the field started to regain attention because of firms' adoption of AI-based software called "expert systems," which provide domain-specific solutions to firm managers by relying on the codified knowledge of experts. However, the technical improvement in expert systems relative to their extremely high price was unsatisfactory, and most investors and governments again decided that AI funding should pause from around the late 1980s until the early 2000s (Nilsson 2009). Although the turning point in the public's perception of the field occurred in 1997, when IBM's Deep Blue beat the world chess champion, the field started to regain broad attention with the realization of Moore's law (Anthes 2017). Since the mid-2000s, there has been a resurgence in AI as the amount of data, computational power, and statistical machine learning tools have improved significantly (Anthes 2017). In particular, the methodological breakthrough by Geoffrey Hinton and his colleagues in 2006 suggested a creative way to deepen the structure of neural network models (i.e., deep-learning) (Hinton et al. 2006). The resurgence of AI also represents a kind of "new science-based technology," a term that indicates that some technologies emerge through the synergy between older scientific knowledge and rapidly advancing information technology (IT) (Koumpis and Pavitt 1999). Since IT capacity is a key factor supporting the such R&D process, incumbent firms in many fields with superior IT capacities advance the frontier knowledge of an emerging technology despite the high risk.

Besides the re-emerging feature of AI, knowledge production in this field shows characteristics that are in accordance with those of science-related technology or Mode 2 knowledge. First, the field of AI may lack an agreed definition, but broadly, it indicates activities devoted to making intelligent machines (Simon 1996; Nilsson 2009). Such a description shows that the final goal of AI is oriented towards solving practical problems, which is expected to result in useful applications. AI has even been regarded as a type of enabling technology for various industrial purposes including manufacturing, medical equipment, geography, automatic vehicles, and internet services (Teece 2018). Such a nature implies that the problem definitions of AI are specified from real-world issues, which is consistent with the basic property of Mode 2 knowledge that *"knowledge [is] produced in the context of application"* (Gibbons et al. 1994, p. 3).

Although several disciplines are related, AI addresses some phenomena surrounding computers and is a part of computer science (Simon 1995). However, computer science has long been considered difficult to classify into any particular category (natural science, applied science, and technological R&D; Rosenberg and Nelson 1994). The field not only develops technology to achieve a goal (Bolander 2019) but also produces theoretical knowledge in relevant areas (Boden 2016). Therefore, much of the science and technology content has become indistinguishable in this field. Even within the computer science community, some researchers argue that the field is science, whereas others argue that it is closer to technology (Denning 2005). This corresponds with one of the properties of Mode 2 knowledge production: "Conventional terms—such as applied science, technological research, or research and development—are inadequate" (Gibbons et al. 1994, p. 2).

The approach to conducting AI research is divided into two main groups: one based on theory and logic and the other described as an empirical art that involves learning through apprenticeship and cases rather than rigorous theory (Nilsson 2009). Although recent progress in AI is more associated with the latter categorization (Boden 2016; Parnas 2017; Hutson 2018), which requires big data and computing power, frontier knowledge in the field is still produced by efforts involving both aspects. Empirical AI researchers describe their research method as heuristic search, which is based on rules of thumb generated through experimentation (Parnas 2017). They repeat experiments until they obtain a more intelligent machine by changing its design. In many cases, researchers do not clearly understand why one machine performs better than another. Such a property is consistent with that of engineering and is referred to as a blind search process of trial and error that recombines various technological components until one obtains successful results (Vincenti 1990; Fleming and Sorenson 2004). Along with its rather messy nature, AI researchers need help from science and mathematics to design better machines with reduced uncertainties when searching for new designs. In addition, many researchers still heavily rely on theory rather than trial and error, again connoting the science-related nature of AI technology. Owing to its characteristics, the field requires professional experts not only from computer science but also from cognitive science, psychology, statistics, and mathematics, among others (Boden 2016). Indeed, it is difficult to adequately handle the problems in this field by relying on the knowledge base of a single organization. Heterogeneity in terms of both scientific disciplines and tacit experience must be coordinated to adequately address problems. Such a property of knowledge production corresponds with one of the important dimensions of Mode 2 knowledge production, namely, "*transdisciplinarity*" (Gibbons et al. 1994, p. 4).

Because of the transdisciplinary nature of the field, R&D collaborations between firms and universities, and even between competing firms, have been actively pursued. From the perspective of firms, such collaborative efforts help them share the high risk of performing frontier R&D as well as mitigating the difficulties in solving complex problems. In the field of AI, there are several forms of U-I collaboration. For example, university and firm researchers often work together on a firm's project, which is one of the most common forms of U-I collaboration. In addition, prominent professors in this field have taken the role of advising on and even leading firms' AI research, which is another form of U-I collaboration in this area. Furthermore, a number of academic AI talents have shifted from academia to industry since the

early 2010s (Gibney 2016). After Andrew Ng, a pioneering researcher in the field, joined Google from Stanford University, many academic experts followed in the same direction. This recent phenomenon is consistent with the *"heterogeneity and organisational diversity"* of Mode 2 knowledge production (Gibbons et al. 1994, p. 6).

Lastly, another important dimension of knowledge production in AI is the need for computational power in training and testing the designed machine. Advances in AI knowledge since around 2006 owe more to the rapidly increased capacity of hardware than to any other factor (Anthes 2017). This is consistent with another feature of Mode 2 knowledge production, described as follows: *"The feature contributing to innovation under Mode 2 conditions is the role that computers and especially computational modeling have come to play, … [and these] can be used to meet a wide variety of uses and of building more sophisticated techniques and instruments that will enhance the design principle and its range of application"* (Gibbons et al. 1994, p. 19).

As Mode 2 knowledge is usually produced in the context of application, we can expect firms performing AI research to play a leading role in suggesting research agendas where they have relevant complementary assets for application. Firms can devise novel technological solutions that are likely to be linked to commercialization, although the suggested ideas can be premature at an early stage. In particular, AI firms' collaboration with researchers in universities or public research institutions is expected to create knowledge that is valuable in the long run. This is because they can adopt more transdisciplinary approaches based on understanding from both scientific knowledge and problems defined in the context of application. Given the nature of knowledge that firms can contribute to this emerging area, we expect that the knowledge evolution of the field will naturally converge towards a direction preferred by firms. This implies that private investments can shape the knowledge trajectory within the emerging technological paradigm with the emergence of science-related technologies.

Hypothesis 1. When science-related technologies emerge, firms' publications influence the formation of published knowledge trajectory more than other publications.
Hypothesis 2. When science-related technologies emerge, publications by firm and non-firm collaborations show higher a) novelty and b) conventionality than other publications.

3 Methodology

3.1 Data

To answer the research questions, we use information on the papers in AI domain. In computer science, top-tier conferences are ranked higher than most journals, excluding a few highly ranked journals (Vardi 2009; Freyne et al. 2010).[3] The peer

[3] When computer science emerged as an independent discipline in the 1980s, both conference proceedings and journal papers were considered important. Since the late 1990s, the role of leading conferences has become more dominant than that of most journals, except for the few leading ones (Vardi 2009; Freyne et al. 2010).

review process in conferences is much faster than that in journals, accelerating the progress of knowledge diffusion in the field of AI. We select target venues, including conferences and journals, and collect information about publications in the target venues from the SCOPUS database, which is a bibliographic database that provides comprehensive information about both journal and conference publications. The top 10 journals and top 10 conferences[4] are listed by referring to Guide2Research,[5] a portal providing the rank of authors, journals, and conferences related to sub-fields of computer science.

Given the general-purpose nature of AI, studies in various domains can apply AI algorithms (WIPO 2019). Although the application of AI algorithms is important, setting the outlet boundaries in all potential fields of study makes it difficult to maintain the knowledge quality of the studies used in our analysis. Therefore, we consider papers published in highly reputed outlets that focus heavily on novel AI algorithms. Among the various sub-fields of computer science, we select a category titled "Machine Learning, Data Mining, and Artificial Intelligence." Table 1 presents a list of highly reputed journals and conferences. The list of selected outlets covers the major outlets mentioned in recent relevant studies, such as Hartmann and Henkel (2019). Based on the final list of journals and conferences, we collect information on the articles published during 2006–2016. As noted in the introduction, "firm paper" indicates a research article in which at least one of the authors is affiliated to a firm (Arora et al. 2018). In this study, firms publishing research include those firms that (i) published at least one article in the field during 2006–2016 and (ii) are listed as a sponsor of the listed conferences (i.e. firms satisfying both (i) and (ii)). We obtain 56,981 papers consisting of 6848 firm papers and 50,133 others (non-firm papers) that include the papers from universities and public research institutions. Among the 6848 firm papers, 5419 are published by firm and non-firm collaboration, while the rest are published by only firm researchers. 120 firms are related to the collected firm papers.[6] The list of firms covers key industry players that appeared in a recent report by the WIPO (2019).

[4] Numerous studies that investigate the phenomenon in certain academic fields have focused on the leading venues because of their representativeness and content quality (e.g., Liu et al. 2019). In addition, Baruffaldi and Poege (2020) noted that firms' publishing and sponsoring in the field of computer science is concentrated in the highly-ranked venues. Recently, Hartmann and Henkel (2019) adopted a similar approach, which focuses on the leading venues in the field of AI.

[5] http://www.guide2research.com/about-us. Ranking of the venues is based on the metric named Impact Score, which reflects both the quantity of contributing prominent scientists and the h-index evaluated from the papers published by top scientists in the last 3 years.

[6] In Appendix A, we report supplementary statistics that help understand our data. The technology field which most of the selected firms primarily focus on is computer technology; this is followed by IT methods for management, transportation, and digital communication (see Fig. 2). Major firms include Microsoft, Google, IBM, Yahoo, Siemens, Toyota, and Intel (see Fig. 3). In addition, Fig. 4 shows that the proportion of firm papers increased due to the increasing proportion of firm and non-firm collaborations.

Table 1 Major conferences and journals

Conferences	IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
	Neural Information Processing Systems (NIPS)
	International Conference on Machine Learning (ICML)
	IEEE International Conference on Computer Vision (ICCV)
	International Conference on Knowledge Discovery and Data Mining (SIGKDD)
	Meeting of the Association for Computational Linguistics (ACL)
	ACM SIGMOD International Conference on Management of Data
	Conference on Empirical Methods in Natural Language Processing (EMNLP)
	AAAI Conference on Artificial Intelligence (AAAI)
	IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Journals	IEEE Transactions on Evolutionary Computation
	IEEE Transactions on Pattern Analysis and Machine Intelligence
	IEEE Transactions on Fuzzy Systems
	IEEE Computational Intelligence Magazine
	International Journal of Neural Systems
	Information Fusion
	Automatica
	Neural Networks
	Journal of Machine Learning Research
	Information Sciences

3.2 Variables and methods

3.2.1 Dependent variables and models

The unit of analyses used to answer our research questions is an individual paper. The dependent variable of interest in H1 is each paper's influence on the formation of published knowledge trajectories. For H2, the dependent variables should reflect the content of the knowledge in each paper with respect to novelty (H2a) and conventionality (H2b).

Influence on the formation of the published knowledge trajectory To examine the influence of each paper on the formation of the published knowledge trajectory (H1), we adopt a measure that reflects the minimum distance of each paper to the knowledge trajectories. A shorter distance represents larger influence of each paper on the knowledge trajectory formation. Boden (2016) noted that state-of-the-art AI includes an extraordinarily wide range of methodologies and therefore cannot be unified by any single core technique. Hence, we first divide the full citation network constructed from the collected AI papers into several groups to detect more detailed trajectories by sub-area. Using a well-established community detection technique (Blondel et al. 2008)[7] as implemented in the Gephi program, we find 22

[7] In Blondel et al.'s (2008) method, modularity is locally maximized by allowing the move of each node assigned to a distinct community to the community of its neighbor(s). Then, the communities are aggregated to form a new network of communities. This operation is repeated until maximum modularity is achieved.

communities that occupy more than 1% of the full citation network. The proportions of the five largest communities among all the communities chosen are as follows: 12.56%, 7.77%, 6.4%, 6.32%, and 5.44%. Modularity is 0.737, which is larger than a widely accepted threshold (i.e., 0.3) for judging whether communities are successfully grouped (Clauset et al. 2004). The identified communities can be understood as major academic groups focusing on different sub-domains within AI.[8]

To identify the knowledge trajectory from each community, we employ the Search Path Node Pair (SPNP) measure, a connectivity measure widely used to detect the main paths in large citation networks (Hummon and Dereian 1989). This represents the key paths of knowledge flow in a network by allocating weights to each link based on how many times the particular link bridges pairs of nodes in the network.[9] Several prior studies have used this algorithm to identify critical knowledge trajectories from papers or patent citation networks (Verspagen 2007; Fontana et al. 2009; Martinelli and Nomaler 2014). We detect the main path for each community chosen.[10] For example, Fig. 1 presents the distinguished trajectory within the largest community along with firm papers lying on the trajectory. We compute how far each paper is from the trajectory (i.e., each paper's minimum distance to one of the papers lying on the distinguished main paths) by regarding a direct citation linkage between two papers as the unit distance. For instance, if paper k can reach one of the trajectory papers by passing through at least two other papers, the minimum distance of paper k to the trajectory is 3. Because the distance value is infinite in the case where a paper does not have any forward citations, we find and set the maximum distance d_m in which the number of papers without any forward citations becomes closest to the number of papers with distance d_m when we regard distances greater than or equal to d_m as d_m. The d_m satisfying this condition is 9 in our data. We use OLS regression to examine whether the firm papers' distance to the knowledge trajectory is significantly shorter than that of other papers, which would imply a significant influence of firms' publications (relative to other publications) on the formation of published knowledge trajectory.

Novelty and conventionality of published knowledge We employ the suggestion in Uzzi et al. (2013), when measuring both the novelty (H2a) and conventionality (H2b) of knowledge represented in a paper. The authors suggested the z-score, which evaluates how often a particular combination of two distinct knowledge bases has been observed, compared with the expected degree in a certain period. The combination of knowledge bases is defined as a pair of different journals in the references of each paper, presuming that a journal addresses similar knowledge. When the knowledge combination appears less often than expected by chance, the

[8] Subsequent topic analysis (see the second part of 3.2.1.) to understand communities shows that each community represents subfield related to AI (e.g. the first community represents 'robotics (12.56%)', the second community represents 'computer vision (7.77%)', and the third community represents 'optimization (6.4%)', etc). Contact the first author for detailed keyword information of communities.

[9] Further intuition of SPNP is suggested in Fig. 5 of Appendix B.

[10] As with previous studies (e.g., Fontana et al. 2009), we use the Pajek program to calculate the SPNP and detect the main path for each community.

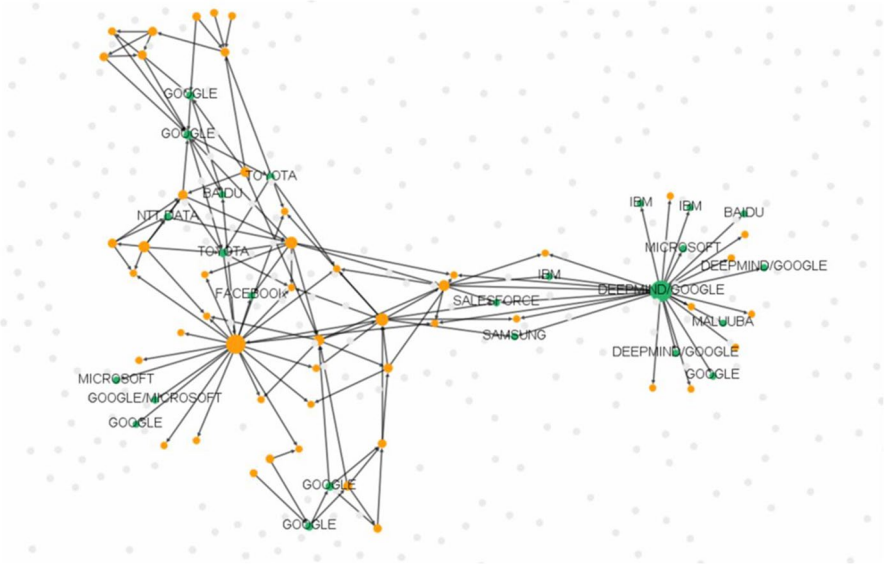

Fig. 1 Papers on the trajectory of the largest community. Note: Trajectory papers including at least one firm-affiliated author are in green and the other papers are in orange

combination is interpreted as a more unusual and atypical one. Several z-scores can be assigned to each paper because a paper usually includes numerous references. Uzzi et al. (2013) proposed that two summary statistics, namely, the median of the z-scores and left 10th percentile of the z-scores allocated to each paper, characterize conventionality and novelty, respectively.

However, applying this measure to our research context is somewhat problematic under the basic assumption that a journal (or a conference) is a distinct knowledge base. As previously mentioned, the boundary of AI is extremely broad and several journals and conferences try to accommodate this variety. Considering that a single journal or conference often addresses a broad range of AI-related research topics,[11] it is difficult to define a venue as the homogeneous unit of knowledge. Therefore, we adjust the measure used in Uzzi et al. (2013) by considering a "research topic" instead of a "journal" as a unit of the knowledge base, following Kaplan and Vakili (2015), which used topic modeling to specify the concept of knowledge recombinations from patent data. This way of adopting z-scores is also in line with Kim et al. (2016), which applied z-scores to patent data by regarding a single patent as a combination of different technology classes instead of using reference information.

To do this, we employ a text-mining technique called dynamic topic modeling (DTM) (Blei and Lafferty 2006). Based on a collection of documents, this algorithm captures not only the per-document topic proportion but also the evolution

[11] For example, the conference program of AAAI 2016 shows a huge variety of topics presented in a single venue (https://www.aaai.org/Conferences/AAAI/2016/aaai16program.pdf).

of detailed keywords composing each topic. The latter part is a differentiated aspect of the DTM from other static topic models. This is useful in our research context because today's AI evolves rapidly with generating new keywords although most of the broad research themes are maintained.[12] DTM is conducted on the 56,981 articles collected, with the number of topics mentioned by scholars set to 100 to allow a feasible interpretation of each topic (Blei and Lafferty 2007; Chang et al. 2009; Hall et al. 2008). We obtain the per-paper topic proportion vectors, which consist of some outstandingly high values showing the paper's strong relevance to some topics. However, most of the other values are close to zero, denoting a lack of relevance between the paper and most other topics. The topics assigned with significantly large proportions can be regarded as the key knowledge components of each paper. For instance, if a vector allocated to a paper has a value of 0.4., 0.2, and 0.3 for three topics (α, β, and γ) and 0.001 for the other 97 topics, one can consider three possible topic combinations (i.e., $\alpha\beta$, $\alpha\gamma$, and $\beta\gamma$) in this paper.

Based on the DTM results, we compute the z-score for topics α and β in year t as follows:

$$z_{\alpha\beta t} = \frac{o_{\alpha\beta t} - \mu_{\alpha\beta}}{\sigma_{\alpha\beta}},$$

where $o_{\alpha\beta t}$ is the observed number of papers belonging to α and β in year t, $\mu_{\alpha\beta}$ is the expected number of papers belonging to α and β in year t based on the papers published in the previous year (i.e., *t-1*), and $\sigma_{\alpha\beta}$ is the corresponding standard deviation. A one-year time lag is employed to reflect the rapidly changing research trend of emerging AI. Following Kim et al. (2016) and Uzzi et al. (2013), median and minimum z-scores are used as the conventionality and novelty of each topic pair, respectively. For both conventionality and novelty, the OLS regression model is applied because the dependent variables are normally distributed. The regression is conducted using the papers published during 2007–2016 (each year is *t*), with each of those papers' conventionality and novelty computed based on papers published in the previous year (i.e., 2006–2015; each year is *t-1*). In addition, we drop papers assigned a single topic because z-scores can be calculated only when at least one topic pair exists. Of the 56,981 papers collected, 50,050 papers remain for the regression analysis.

3.2.2 Independent variables and controls

Independent variables The main independent variable is a dummy variable indicating whether a paper is a firm paper (1) or not (0) (*Firm Paper1 (Firm Involved or Not)*). In addition, we further consider a categorical variable with three levels, differentiating among publications by only non-firm researchers (including those in

[12] To apply DTM, we use Python's Gensim (https://radimrehurek.com/gensim/) library. Appendix C provides further clarity on why this study adopts DTM.

universities or public research institutions) (0), both firm and non-firm researchers (1), and only firm researchers (2) (*Firm Paper2 (Firm Only, Non-firm Only, or Firm/Non-firm Collaboration)*).

Controls We include several variables to control for other factors that may influence each paper's impact on the published knowledge evolution. Garfield (1979) underlined that the average number of references per paper reflects the potential impact of papers in a given field. Hence, we control for the number of references in each paper (*Number of References*). The year of publication is included to control for the increasing popularity of AI during its recent period of development, and the expected positive relationship between the age of paper and distance to the trajectory (*Publication Year* (*t*)). Within the various sub-areas of AI, some research topics garner more attention in a period than other topics (Boden 2016), which may influence the impact of each paper. Therefore, we control for the size of the academic groups with which each paper is associated. We use the community detection results to consider the number of papers in each community as the size of the academic group having interest in a specific sub-area of AI (*Size of Community*). A dummy variable on whether the paper is published in a conference or journal is used as a control because computer science tends to prefer conferences as a channel for publishing high-quality papers, though some disagreement exists (Vardi 2009) (*Conference Journal Dummy*).

The number of authors is controlled for to reflect controversial evidence on whether multi-author publications are likely to have a higher impact than single-author publications (Gazni and Didegah 2011; Didegah and Thelwall 2013; *Number of Authors*). Regarding author-related characteristics, the reputation of authors is also a crucial factor that determines the impact of publications, besides their quality (Bornmann et al. 2012). Merton (1968) similarly mentioned the Matthew effect in science, which is the tendency to give credit to well-known scholars. Moreover, firms that invest in AI try to collaborate with highly reputed researchers (Gibney 2016). To capture this influence, we control for the authors' average performance that relates to each paper in the two following complementary ways. First is the average number of the authors' recent publications in computer science (*Authors' Average Number of CS Publications* (*t*-4 ~ *t*)), which reflects the quantity of the authors' research output. Second is the average of the authors' *h*-index, which denotes the number of publications with a citation number $\geq h$ (Hirsch 2005; *Authors' Average h-index at t*).[13] The *h*-index reflects not only the quantity of the author's research outputs but also their academic impact.

In addition, we control for the number of affiliations per paper because heterogeneity in author affiliations positively correlates with the impact of publications (Franceschet and Costantini 2010; *Number of Affiliations*). Moreover, the institutional characteristics of affiliations can shape researchers' publishing behavior, potentially influencing the impact of each paper. For instance, since firm-affiliated researchers are less pressured to publish papers than those who are

[13] "A scientist has index *h* if *h* of his or her N_p papers have at least *h* citations each and the other (N_p- *h*) papers have *h* citations each" (Hirsch 2005, p. 16569).

university-affiliated, firm researchers can selectively publish only part of their research outcomes. In addition, firm researchers should often follow the firm-level approach on strategic publishing regarding where, when, and what to publish, leading to a potential bias in firm publications and its subsequent impact (e.g., Li et al. 2015). Therefore, we control for the average academic performance of affiliations related to each paper using the average number of their recent publications in the CS field (*Affiliations' Average Number of CS Publications* (t-4 ~ t)) and the average of their h-index (*Affiliations' Average h-index at t*).

To compute for the described control variables related to authors and author affiliations, a range of publication information broader than the collected sample of AI papers should be considered. We additionally employ the Microsoft Academic Graph (MAG; Sinha et al. 2015),[14] which is a widely used database that includes publication records, citation relationships, fields of study,[15] and author or affiliation-specific identification codes.[16] Based on this database, we compute the author and affiliation-related control variables introduced above.[17]

Lastly, some sub-fields of AI have been particularly affected by the recent deep-learning revolution. Although there are more relevant areas, computer vision and natural language processing (NLP) are representative cases of these sub-fields (LeCun et al. 2015). These sub-fields have strongly benefited from firms' resources (big data and computational facilities). Considering firms' significant investment in recent AI, we can conjecture that the impact of papers in such sub-fields may have been higher than the papers in other fields. Therefore, we control for whether the paper is related to vision or NLP (*Firm-resource Sensitive Field*). This variable is obtained by using fields of study information in MAG data.[18]

4 Findings

Table 2 presents basic statistics and correlations. Two sub-tables are reported because we lost some observations in the process of computing novelty and conventionality, as described above. The upper and lower sub-tables correspond to the first (H1) and second (H2a and H2b) hypotheses, respectively. Although some observations are lost when we test H2, the two sub-tables show that the correlation and basic statistics are consistent overall. The averages of the number

[14] We use the MAG version updated in March 2020.

[15] See Table 8 in Appendix E for the field of study keywords we use to extract CS-related publications.

[16] Based on the identification codes of MAG, we find that 14,132 out of 101,187 authors related to our sample papers have co-affiliations (i.e., about 13%). If the multiple affiliations of an author include a firm, we regard the author as a firm-affiliated author and the author's paper, a firm paper.

[17] We compute the author- and affiliation-related control variables within the boundary of papers published at the Web of Science listed journals (https://mjl.clarivate.com/home) and major CS conferences (http://www.guide2research.com/about-us).

[18] See Table 9 in Appendix E for the field of study keywords we use to extract computer vision and NLP-related publications.

Table 2 Descriptive statistics and correlations

	Mean	S.D.	Min	Max									
*Level of analysis: a paper													
H1 ($n=56{,}981$)													
Dependent variable													
H1) Distance to the knowledge trajectory (influence on the formation of the published knowledge trajectory)	8.02	2.47	0	9									
Independent variables													
1) Firm Paper1 (*Firm Involved or Not*) (1 if including firm(s); 0 otherwise)	6848 (1), 50,133 (0)												
2) Firm Paper2 (*Firm Only, Non-firm Only, or Firm/Non-firm Collaboration*) (2 if firm only; 1 if firm and non-firm collaboration; 0 otherwise)	1419 (2), 5429 (1), 50,133 (0)												
Control variables					3)	4)	5)	6)	7)	8)	9)	10)	11)
3) Number of References	27.55	16.65	0	349	1.00								
4) Publication Year (t)	2011.56	3.12	2006	2016	0.24	1.00							
5) Number of Authors	3.24	1.53	1	69	0.04	0.13	1.00						
6) Authors' Average Number of CS Publications (t-4 ~ t)	11.74	10.95	0	144	0.14	0.15	0.07	1.00					
7) Authors' Average h-index at t	6.99	5.24	0	73	0.16	0.23	0.01	0.60	1.00				
8) Number of Affiliations	1.52	0.74	1	19	0.12	0.14	0.41	0.16	0.17	1.00			
9) Affiliations' Average Number of CS Publications (t-4 ~ t)	1288.68	1098.05	0	7822	0.04	0.26	0.09	0.30	0.29	0.03	1.00		
10) Affiliations' Average h-index at t	183.96	122.74	0	862	0.01	0.24	0.02	0.19	0.35	−0.03	0.59	1.00	
11) Size of Community	3.72	3.74	0	12.56	0.07	0.14	0.01	0.13	0.12	0.02	0.16	0.13	1.00
12) Conference/Journal (1 if a conference paper; 0 otherwise (a journal paper))	34,748 (1), 18,233 (0)												
13) Firm-resource Sensitive Field (1 if related to vision or NLP; 0 otherwise)	18,896 (1), 38,085 (0)												
H2 ($n=50{,}050$)													

Table 2 (continued)

	Mean	S.D.	Min	Max									
Dependent variables													
H2a) z-min (novelty)	−2.97	5.16	−20.03	114.61									
H2b) z-median (conventionality)	2.24	4.60	−16.65	118.91									
Independent variables													
1) Firm Paper1 (*Firm Involved or Not*) (1 if including firm(s); 0 otherwise)	6214 (1), 43,836 (0)												
2) Firm Paper2 (*Firm Only, Non-firm Only, or Firm/Non-firm Collaboration*) (2 if firm only; 1 if firm and non-firm collaboration; 0 otherwise)	1268 (2), 4946 (1), 43,836 (0)												
Control variables					3)	4)	5)	6)	7)	8)	9)	10)	11)
3) Number of References	28.46	16.33	1	321	1.00								
4) Publication Year (*t*)	2011.92	2.86	2007	2016	0.23	1.00							
5) Number of Authors	3.27	1.53	1	69	0.03	0.12	1.00						
6) Authors' Average Number of CS Publications (*t*-4 ~ *t*)	12.00	10.98	0	144	0.15	0.13	0.08	1.00					
7) Authors' Average h-index at *t*	7.12	5.21	0	73	0.16	0.21	0.02	0.60	1.00				
8) Number of Affiliations	1.53	0.75	1	19	0.12	0.13	0.41	0.16	0.18	1.00			
9) Affiliations' Average Number of CS Publications (*t*-4 ~ *t*)	1329.03	1117.95	0	7822	0.03	0.24	0.08	0.29	0.30	0.02	1.00		
10) Affiliations' Average h-index at *t*	188.42	125.22	0	862	−0.01	0.23	0.02	0.18	0.35	−0.04	0.59	1.00	
11) Size of Community	3.94	3.75	0	12.56	0.03	0.11	0.00	0.14	0.12	0.02	0.16	0.12	1.00
12) Conference/Journal (1 if a conference paper; 0 otherwise (a journal paper))	33,955 (1), 16,095 (0)												
13) Firm-resource Sensitive Field (1 if related to vision or NLP; 0 otherwise)	17,120 (1), 32,930 (0)												

of authors and author affiliations in a single paper are 3.24 and 1.52, respectively. On average, authors publish about 11 CS papers within five years from the publication year t (i.e., t-4 ~ t), and their average h-index at year t is 6.99. On average, affiliations associated with a paper produce about 1288 CS papers during the period t-4 ~ t, and their average h-index at year t is 183.96. The minimum value of the authors' (and affiliations') average number of CS publications can be zero even though they have published at least one paper in the selected list of AI journals and conferences. This is because there are few AI papers not involved in the CS field categories that we employ from the MAG database (see Appendix E), owing to the transdisciplinary nature of AI. The *Firm-resource Sensitive Field* dummy shows that about one-third of the collected papers are related to vision or NLP.

Table 3 compares the following three paper groups—firm only, firm and non-firm collaboration, and non-firm only—in terms of the author and affiliation-related variables. This table shows that the firm and non-firm collaboration group consists of more papers that were produced by high performing researchers (in terms of both publishing quantity and h-index) than the other two groups. This suggests that firms selectively collaborate with outstanding researchers in universities. The table also shows that the average number of publications (in all fields) by affiliations associated with a single paper is highest in the non-firm paper group, which is an intuitive statistic because this group mostly consists of universities. By contrast, the number of CS publications is shown to be higher in the firm-related paper groups than in the non-firm paper group. This implies that compared with universities, a firm investing in basic research of emerging science-related technology can, on average, publish even more papers in that field, which is in line with Hicks's (1995) observation. In addition, the average h-index of affiliations is shown to be highest in the firm and non-firm collaboration paper group, which reconfirms that firms selectively collaborate with highly reputed affiliations that include outstanding researchers.

Table 4 reports the results of the regression analyses for H1 (OLS models; DV is the distance to the knowledge trajectory). Model 1 shows the OLS results with only the control variables. This baseline model reveals that the coefficients of the control variables are significant in the directions that we expected. Model 2 adds *Firm Paper1 (Firm Involved or Not)* and shows that the coefficient of this variable is significantly negative. That is, firm papers tend to have shorter distances to the knowledge trajectory (about 0.49 shorter than other papers when a direct citation distance equals 1), indicating their higher influence on the formation of the published knowledge trajectory compared to other papers. Model 3 adds *Firm Paper2 (Firm Only, Non-firm Only, or Firm/Non-firm Collaboration)* and shows that firm-only papers (2) had the shortest distance to the knowledge trajectory (negative and significant; about 0.74 shorter than non-firm only (0) papers when a direct citation distance equals 1), which confirms the group's significantly higher influence on the knowledge trajectory formation than other papers. The 95% confidence intervals on the firm and non-firm collaboration (1) and firm-only (2) publications run from −0.49 to −0.35 and − 0.86 to −0.62, respectively, confirming the significantly higher influence of the latter than he former on the formation

Table 3 The averages of the groups' control variables related to authors and affiliations

Variables	Paper groups		
	Firm only ($n=1419$)	Firm and non-firm collaboration ($n=5429$)	Non-firm ($n=50,133$)
Number of Authors	3.17	4.04	3.16
Authors' Average Number of Publications (in all fields) (t-4 ~ t)	11.90	19.13	15.87
Authors' Average Number of CS Publications (t-4 ~ t)	10.19	16.28	11.29
Authors' Average h-index at t	6.18	8.58	6.83
Number of Affiliations	1.13	2.17	1.45
Affiliations' Average Number of Publications (in all fields) (t-4 ~ t)	5038.43	10,848.91	12,269.17
Affiliations' Average Number of CS Publications (t-4 ~ t)	1822.81	1865.42	1211.10
Affiliations' Average h-index at t	129.72	189.84	184.85

t is the publication year. The *Authors'* (*Affiliations'*) *Average Number of Publications (in all fields)* is not used in our main model because of its high correlation with the *Authors'* (*Affiliations'*) *Average Number of CS Publications*

Table 4 Regression results (H1)

	H1) OLS ($n = 56{,}981$)		
	DV: Distance to the knowledge trajectory		
	1) Model 1	1) Model 2	1) Model 3
Firm Paper1 (*Firm Involved or Not*) (1 if including firm(s))		−0.4948**	
		(0.0317)	
Firm Paper2 (*Firm Only, Non-firm Only, or Firm/Non-firm Collaboration*)			
(1 if firm and non-firm collaboration)			−0.4243**
			(0.0353)
(2 if firm only)			−0.7452**
			(0.0634)
Number of References	−0.0102**	−0.0102**	−0.0102**
	(0.0006)	(0.0006)	(0.0006)
Publication Year (t)	0.2009**	0.1987**	0.1991**
	(0.0034)	(0.0034)	(0.0034)
Number of Authors	0.0062	0.0104	0.0101†
	(0.0071)	(0.0070)	(0.0070)
Authors' Average Number of CS Publications (t-4 ~ t)	−0.0111**	−0.0113**	−0.0116**
	(0.0011)	(0.0011)	(0.0011)
Authors' Average h-index at t	−0.0433**	−0.0417**	−0.0415**
	(0.0025)	(0.0025)	(0.0025)
Number of Affiliations	0.0057	0.0489**	0.0367**
	(0.0148)	(0.0150)	(0.0152)
Affiliations' Average Number of CS Publications (t-4 ~ t)	−0.0001**	−0.0000**	−0.0000*
	(0.0000)	(0.0000)	(0.0000)
Affiliations' Average h-index at t	−0.0012**	−0.0015**	−0.0015**
	(0.0001)	(0.0001)	(0.0001)
Size of Community	−0.1452**	−0.1420**	−0.1418**
	(0.0027)	(0.0027)	(0.0027)
Conference/Journal (1 if a conference paper)	−0.1175**	−0.0735**	−0.0699**
	(0.0246)	(0.0247)	(0.0247)
Firm-resource Sensitive Field (1 if related to vision or NLP)	−0.2871**	−0.2882**	−0.2899**
	(0.0221)	(0.0220)	(0.0220)
Constant	−394.56**	−390.16**	−390.88**
	(6.8362)	(6.8277)	(6.8284)
Adjusted R-square	0.1295	0.1331	0.1334

† p<0.1, * p<0.05, **p<0.01. Standard errors are in parentheses

Descriptions of the dummy variables are as follows: 1. Firm Paper1 (*Firm Involved or Not*): 1 if including firm(s) and 0 otherwise; 2. Firm Paper2 (*Firm Only, Non-firm Only, or Firm/Non-firm Collaboration*): 2 if firm only, 1 if firm and non-firm collaboration, and 0 otherwise; 3. Conference/Journal: 1 if a conference paper and 0 otherwise (a journal paper); and 4. Firm-resource Sensitive Field: 1 if related to vision or NLP and 0 otherwise

of knowledge trajectory. Overall, Table 4 shows the outstanding impact of papers that include firm-affiliated authors—particularly those authored by firm researchers only—on the trajectory formation of published knowledge. For a robustness check (see Table 6 in Appendix D), we provide the results of the log-normal regression (see Models 4–6 in Table 6), which is widely used to model the distance. In addition, because a considerable proportion of papers is disconnected or extremely far from the trajectory, we also suggest the results of the models that exclude papers without citations in the five years since its publication (see Models 7–9 in Table 6) and those excluding papers that are not involved in any of the major communities (i.e., community size is 0) (see Models 10–12 in Table 6).

Table 5 reports the regression results of H2a (the OLS models for DV, z-min (novelty)) and H2b (the OLS models for DV, z-med (conventionality)). The first set of three columns in Table 5 reports the H2a-related results. Model 1 contains only control variables, and Models 2 and 3 add *Firm Paper1 (Firm Involved or Not)* and *Firm Paper2 (Firm Only, Non-firm Only, or Firm/Non-firm Collaboration)*, respectively. Model 2 shows that firm papers have significantly higher novelty (i.e., lower value of the minimum z-score) than the non-firm papers. In particular, Model 3 shows that papers produced by firm and non-firm collaboration (1) have significantly higher novelty than non-firm papers (0). The second set of three columns in Table 5 reports H2b-related results. Model 1 includes only control variables, and Models 2 and 3 add *Firm Paper1 (Firm Involved or Not)* and *Firm Paper2 (Firm Only, Non-firm Only, or Firm/Non-firm Collaboration)*, respectively. Models 2 shows that firm papers have significantly higher conventionality (i.e., higher value of the median z-score), than non-firm papers. Based on Model 3, we find that the 95% confidence intervals on firm and non-firm collaboration (1) and firm-only (2) publications run from −0.02 to 0.26 and − 0.03 to 0.48, respectively. This indicates that there is no significant difference between these two groups in terms of conventionality. Since the ranges of z-min and z-med are fairly large (−20.03 to 114.61 and − 16.65 to 118.91, respectively; see Table 2), we conduct robustness check after removing extreme z-values from the original data. Table 7 in Appendix D shows the robust results after winsorization.

In summary, Table 5 interestingly shows that during AI's period of emergence, firm and non-firm collaboration papers satisfied both conditions of higher conventionality and novelty than non-firm papers. According to observations in Uzzi et al. (2013), publications that satisfy the conditions of both outstanding novelty and conventionality are likely to have high impact in the long run. In this regard, our findings imply that during the recent development of AI, collaborations between firms and universities have produced new knowledge that is likely to be critical in the long run.

5 Discussion and conclusion

This study investigated the influence of firms' research publication on the evolution of published knowledge over the period of the resurgence of a science-related technology, specifically, AI. The empirical results suggest that papers with firm-affiliated

Table 5 Regression results (H2a and H2b)

	H2a) OLS ($n=50{,}050$)			H2b) OLS ($n=50{,}050$)		
	DV: z-min (novelty)			DV: z-median (conventionality)		
	2a) Model 1	2a) Model 2	2a) Model 3	2b) Model 1	2b) Model 2	2b) Model 3
Firm Paper1 (*Firm Involved or Not*) (1 if including firm(s))		−0.1900**			0.1440*	
		(0.0747)			(0.0663)	
Firm Paper2 (*Firm Only, Non-firm Only, or Firm/Non-firm Collaboration*)						
(1 if firm and non-firm collaboration)			−0.2193**			0.1206†
			(0.0829)			(0.0736)
(2 if firm only)			−0.0841			0.2285*
			(0.1498)			(0.1330)
Number of References	−0.0062**	−0.0062**	−0.0062**	0.0088**	0.0088**	0.0088**
	(0.0015)	(0.0015)	(0.0015)	(0.0013)	(0.0013)	(0.0013)
Publication Year (t)	0.1535**	0.1526**	0.1524**	0.0110†	0.0117†	0.0116†
	(0.0087)	(0.0087)	(0.0087)	(0.0077)	(0.0077)	(0.0077)
Number of Authors	−0.0657**	−0.0641**	−0.0640**	−0.0361**	−0.0374**	−0.0373†
	(0.0168)	(0.0168)	(0.0168)	(0.0149)	(0.0149)	(0.0149)
Authors' Average Number of CS Publications (t-4 ~ t)	−0.0225**	−0.0226**	−0.0225**	0.0100**	0.0100**	0.0101**
	(0.0027)	(0.0027)	(0.0027)	(0.0024)	(0.0024)	(0.0024)
Authors' Average h-index at t	0.0546**	0.0553**	0.0552**	0.0441**	0.0436**	0.0435**
	(0.0059)	(0.0059)	(0.0059)	(0.0052)	(0.0052)	(0.0052)
Number of Affiliations	0.1334**	0.1508**	0.1561**	0.1188**	0.1056**	0.1098**
	(0.0350)	(0.0356)	(0.0362)	(0.0310)	(0.0316)	(0.0321)
Affiliations' Average Number of CS Publications (t-4 ~ t)	0.0000†	0.0000*	0.0000**	0.0000**	0.0000**	0.0000**
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)

Table 5 (continued)

	H2a) OLS ($n=50{,}050$)			H2b) OLS ($n=50{,}050$)		
	DV: z-min (novelty)			DV: z-median (conventionality)		
	2a) Model 1	2a) Model 2	2a) Model 3	2b) Model 1	2b) Model 2	2b) Model 3
Affiliations' Average h-index at *t*	−0.0008**	−0.0009**	−0.0009**	−0.0005**	−0.0004**	−0.0004*
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Size of Community	0.0399**	0.0410**	0.0409**	0.0074†	0.0066	0.0065
	(0.0063)	(0.0064)	(0.0064)	(0.0056)	(0.0056)	(0.0056)
Conference/Journal (1 if a conference paper)	0.1012*	0.1196*	0.1181*	0.6449**	0.6310**	0.6298**
	(0.0595)	(0.0599)	(0.0600)	(0.0528)	(0.0532)	(0.0532)
Firm-resource Sensitive Field (1 if related to vision or NLP)	−0.6453**	−0.6465**	−0.6457**	0.7514**	0.7523**	0.7529**
	(0.0521)	(0.0521)	(0.0520)	(0.0463)	(0.0463)	(0.0463)
Constant	−311.79**	−309.99**	−309.63**	−21.53†	−22.88†	−22.60†
	(17.4871)	(17.5003)	(17.5060)	(15.5229)	(15.5350)	(15.5401)
Adjusted R-square	0.0151	0.0152	0.0152	0.0197	0.0198	0.0198

† $p<0.1$, * $p<0.05$, ** $p<0.01$. Standard errors are in parentheses

Notes: This table presents OLS model results for the z-min (H2a) and z-median (H2b)

Descriptions of the dummy variables are as follows: 1. Firm Paper1 (*Firm Involved or Not*): 1 if including firm(s) and 0 otherwise; 2. Firm Paper2 (*Firm Only, Non-firm Only, or Firm/Non-firm Collaboration*): 2 if firm only, 1 if firm and non-firm collaboration, and 0 otherwise; 3. Conference/Journal: 1 if a conference paper and 0 otherwise (a journal paper); and 4. Firm-resource Sensitive Field: 1 if related to vision or NLP and 0 otherwise

authors, particularly those with only firm-affiliated author(s), have had a significantly higher influence on the direction of published knowledge evolution than other types of papers evaluated in this study. In addition, we show that collaborations between firms and non-firms have played a pivotal role in producing more conventional and novel knowledge. This finding suggests that the collaborated papers are likely to include potentially salient knowledge in the long run.

The results of this study are largely attributable to the distinctive characteristics of AI; therefore, we extend them in consideration of the boundary conditions outlined in Section 2.4. In addition to the technological nature of the study's subject, we emphasize on the industrial context in which the assets necessary to develop and commercialize AI are concentrated in firms. We expect that the findings of this study would be extended to sectors in which the circumstances of the technology itself and the distribution of complementary assets are similar to those of AI. Despite constraints in generalizing the findings of this study, we, nevertheless, link our results to findings in the literature to explain our contributions.

Our findings imply that under high technological uncertainty, researchers can be tempted to focus their efforts on the research areas in which firms have published. The finding extends an argument of Hicks (1995) that firm papers signal the presence of unpublished tacit knowledge and resources of the respective firms. We showed that when firms' unpublished assets can be used to contribute to the innovation and knowledge frontier of an emerging science-related technology, their publishing can affect the type of knowledge that researchers produce (i.e., where to focus their research efforts).

In terms of the evolution of science-related technology, our findings suggest a case in which published knowledge is significantly driven by private investments instead of public ones. This is an interesting but less studied phenomenon, requiring further investigation to understand its implications. While the support of private money has led to the evolution of published knowledge, we find that both firms and non-firm entities have contributed to knowledge evolution through their own strengths.

Our results have implications for science and technology policies. First, while our main interpretations focus on non-firm researchers who collaborate with firm researchers, the basic statistics in Table 2 show that most papers are still produced solely by non-firm researchers. This implies that many researchers in universities and public research institutions have addressed topics less attended by firms. The diversity of knowledge maintained by non-firm researchers is necessary to create another breakthrough in the future (Nieto and Santamaría 2007; Uzzi et al. 2013). In this vein, to sustain diversity in research topics, which will trigger another breakthrough in the long run, greater public funding for AI research should be allocated to research topics in which firms are less interested in.

Second, innovation can be achieved only when the components relevant to technology commercialization are successfully coordinated. Although the private sector invests in and leads AI research in which they have complementary assets, there may be hurdles firms cannot easily handle in the long process of innovation. For example, although firms interested in autonomous vehicles invest significantly in and impact computer vision and robotics research, infrastructure and ethical issues remain

critical challenges to successful commercialization and adoption of autonomous vehicles. This concern is relevant to the concept of "reverse salients" (Hughes 1987), which indicates factors deterring innovation within a system, including technological, institutional, market, infrastructure, and legal aspects (Bijker et al. 1989; Mulder and Knot 2001; Takeishi and Lee 2005; Murmann and Frenken 2006). To make the firms' investment in knowledge production more meaningful, the government should help mitigate potential obstacles to innovation that private actors cannot handle easily.

The implications of the role of public funding are relevant to recent evidence of China's emergence in the field of AI. In line with our results, Lundvall and Rikap (2022) suggested the concept of "Corporate Innovation System" in the field of AI, underlining the dominant role of tech giants in leading the direction of the field evolution. In particular, the authors emphasize the co-evolution between the national and corporate innovation systems in China, which offers a new window of opportunity for the country. China's government has provided a stepping stone for the country to catch up in AI by funding AI research of top universities and public research institutions and investing in AI-related infrastructure, such as markets for data and digital services. The case of China shows the importance of the government's role in supporting targeted areas that can facilitate innovation, given that firms lead knowledge development in the field.

From the perspective of firms, the results can be interpreted to mean that firms engaged in publishing have had opportunities to drive knowledge evolution in the direction they chose to publish—opportunities that imply such firms' research investments have gained legitimacy. A desirable scenario for firms investing in AI research is that the emerging knowledge evolves in a direction that enables them to leverage their resources and capabilities efficiently. Therefore, the value of gaining legitimacy for a firm's research investment in an uncertain technology is significant.

Some of the limitations of this study suggest future research directions. The first limitation of this study is the uncertain scope of generalizing the results. Although we highlight the scope covered by this study, future studies are needed to investigate whether and how firms publishing on areas other than AI can (differently or similarly) influence the evolution of published knowledge. Second, while this study sheds light on the implications of firms' publications from the policy perspective, another area that must be studied is when firms can benefit from publishing. Although prior studies have shown some of the motivational aspects of publishing for firms (Rosenberg 1990; Hicks 1995; Polidoro and Toh 2011), the consequences of publishing remain an unanswered question. Accordingly, future studies should explore firms' benefits from publishing from such aspects as, say, knowledge spillovers and incentivizing firm-affiliated researchers. Third, although we used categorical variables to distinguish firm papers, future studies can use more fine-grained measures reflecting the exact share of firm-affiliated authors in paper to get further understandings of firms' engagement in research publication. Lastly, reflecting on the results of the present study, further consideration of firms that forge deep links with universities is needed. We show that firms not only provide significant funding to university researchers but also help build the main research trajectories for some emerging science-related technologies. This provides evidence of the bidirectional impact between industry and academia. Future studies are needed to investigate the details of such bidirectional linkages in more depth.

Appendix A

Note: The primary technology field where each firm focuses on is determined by its patents. We use the PATSTAT database, which provides worldwide patent information obtained from global patent offices. The PATSTAT database includes 35 technology fields of patents, which are referred from the World Intellectual Property Organization (http://www.wipo.int/export/sites/www/ipstats/en/statistics/patents/xls/ipc_technology.xls). A firm is regarded as focusing mainly on a specific technology field for which it has applied the largest proportion of its patents in the period 2006–2016

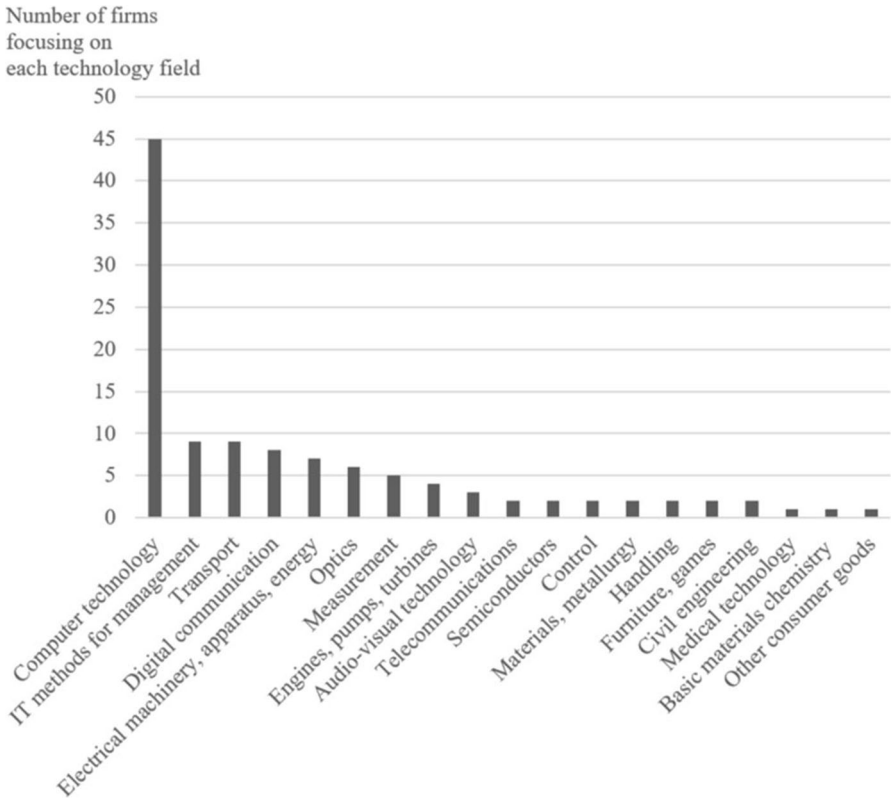

Fig. 2 Number of firms focusing on each technology field

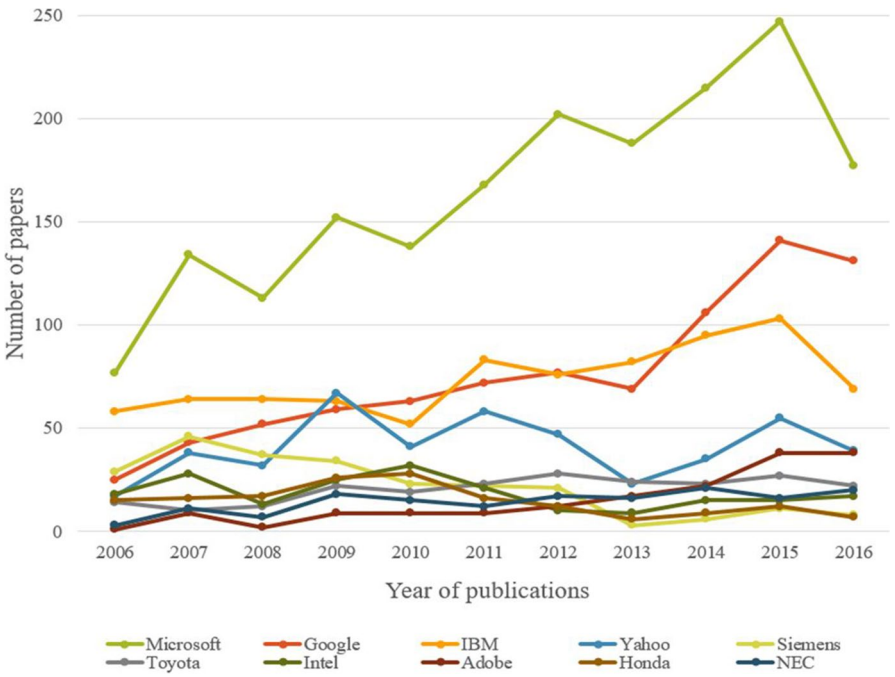

Fig. 3 Number of papers published in the selected conferences and journals by major firms

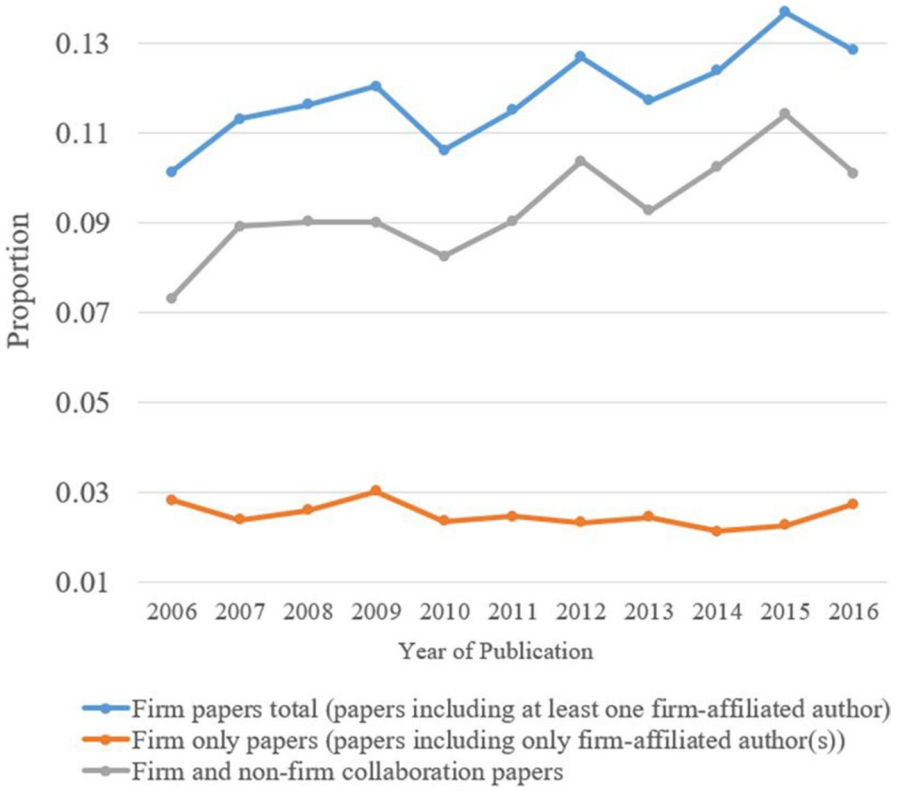

Fig. 4 Proportion of total firm papers and its subgroups (i.e., firm-only papers and firm and non-firm collaboration papers)

Appendix B

Note: In this figure, each node is a paper, and each link connects from the cited to the citing papers. For example, the SPNP value of the link D–E is 24 because there are 24 (4×6) paths beginning from A, B, C, and D (A, B, and C are D's ancestors) and ending at E, G, H, I, J, and K (G, H, I, J, and K are E's descendants).

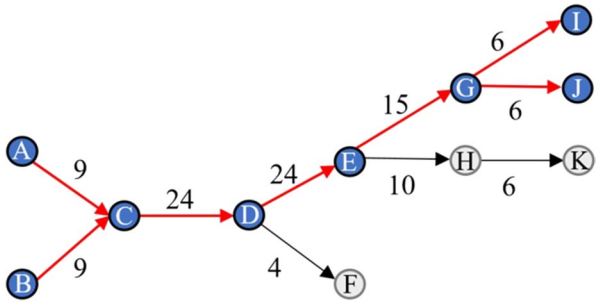

Fig. 5 An example of the main paths (red lines) based on the Search Path Node Pair values

Appendix C

DTM (Blei and Lafferty 2006) is a generative model for analyzing the evolution of topics of documents over time. Like other static topic models (e.g., Latent Dirichlet Allocation), a document is regarded as a mixture of unobserved topics and a topic defines a multinomial distribution over words. A topic is drawn from the mixture and a word is drawn from the multinomial distribution corresponding to that topic. However, in the DTM, documents are organized into time slices and the documents in each time slice are modeled with a *k*-topic model. The detected topics evolve from the last time slice's topics, generating a chain-like topic evolution (see Fig. 6).

Note: This figure illustrates the evolution of one of the AI-related topics used in our main analysis. We can observe that the priority of words related to this topic changes over time (e.g., the priority of "Kinematics" increases over time, and "KinematicModel" first appears in the top ten in 2011), while dominant words are maintained overall across time. Considering these characteristics, using DTM allows us to reflect better the rapidly evolving nature of recent AI when we compute z-score based on yearly topic combinations

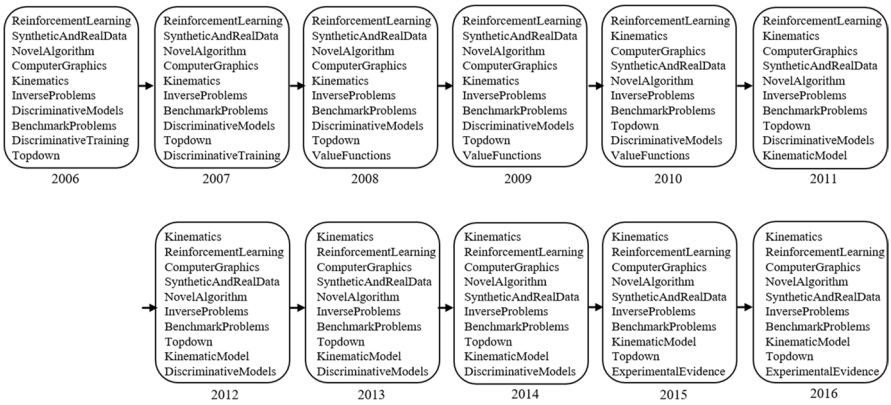

Fig. 6 An example of topic evolution (the top 10 words from the inferred posterior distribution are reported in descending order).

Appendix D. Robustness checks

Table 6 Additional regression for robustness check (H1)

	H1) Log-normal regression ($n=56{,}981$)			H1) 5-year citation >0 ($n=21{,}926$)			H1) Community size >0 ($n=37{,}496$)		
	DV: Ln (Distance)			DV: Distance			DV: Distance		
	1) Model 4	1) Model 5	1) Model 6	1) Model 7	1) Model 8	1) Model 9	1) Model 10	1) Model 11	1) Model 12
Firm Paper1 (*Firm Involved or Not*)		−0.1697**			−0.6652**			−0.4840**	
(1 if including firm(s))		(0.0125)			(0.064)			(0.0421)	
Firm Paper2 (*Firm Only, Non-firm Only, or Firm/Non-firm Collaboration*)									
(1 if firm and non-firm collaboration)			−0.1418**			−0.4880**			−0.3731**
			(0.0139)			(0.0703)			(0.0465)
(2 if firm only)			−0.2688**			−1.3377**			−0.8964**
			(0.0251)			(0.1285)			(0.0846)
Number of References	−0.0041**	−0.0004**	−0.0040**	−0.0166**	−0.0166**	−0.0165**	−0.0127**	−0.0129**	−0.0128**
	(0.0002)	(0.0002)	(0.0002)	(0.0015)	(0.0014)	(0.0014)	(0.0009)	(0.0009)	(0.0009)
Publication Year (*t*)	0.0435**	0.0427**	0.0429**	0.3350**	0.3318**	0.3330**	0.3433**	0.3304**	0.3413**
	(0.0013)	(0.0013)	(0.0013)	(0.0085)	(0.0085)	(0.0085)	(0.0051)	(0.0051)	(0.0051)
Number of Authors	−0.0054*	−0.0040†	−0.0041†	−0.0381*	0.0459**	0.0466	−0.0120	0.0170**	0.0170
	(0.0028)	(0.0028)	(0.0028)	(0.0168)	(0.0168)	(0.0168)	(0.0103)	(0.0103)	(0.0103)
Authors' Average Number of CS Publications (t-4 ~ t)	−0.0027**	−0.0028**	−0.0029**	−0.0123**	−0.0128**	−0.0139**	−0.0039*	−0.0045**	−0.0050**
	(0.0004)	(0.0004)	(0.0004)	(0.0024)	(0.0024)	(0.0024)	(0.0015)	(0.0015)	(0.0015)
Authors' Average h-index at t	−0.0151**	−0.0146**	−0.0145**	−0.0682**	−0.0658**	−0.0647**	−0.0607**	−0.0589**	−0.0585**
	(0.0009)	(0.0009)	(0.0009)	(0.0054)	(0.0054)	(0.0054)	(0.0035)	(0.0035)	(0.0035)

Table 6 (continued)

	H1) Log-normal regression ($n=56{,}981$)			H1) 5-year citation >0 ($n=21{,}926$)			H1) Community size >0 ($n=37{,}496$)		
	DV: Ln (Distance)			DV: Distance			DV: Distance		
	1) Model 4	1) Model 5	1) Model 6	1) Model 7	1) Model 8	1) Model 9	1) Model 10	1) Model 11	1) Model 12
Number of Affiliations	0.0043	0.0192**	0.0143*	0.0292	0.0476†	0.0037	0.0044	0.0468†	0.0228
	(0.0058)	(0.0059)	(0.0060)	(0.0334)	(0.0342)	(0.0349)	(0.0211)	(0.0216)	(0.0220)
Affiliations' Average Number of CS Publications (t-4 ~ t)	−0.0000**	−0.0000*	0.0000*	−0.0001**	−0.0001**	0.0000*	−0.0000**	−0.0000**	−0.0000*
	(0.0000)	(0.0000)	(0.0000)	(0.0334)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Affiliations' Average h-index at t	−0.0003**	−0.0004**	−0.0004**	−0.0016**	−0.0022**	−0.0023**	−0.0013**	−0.0016**	−0.0017**
	(0.0000)	(0.0000)	(0.0000)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.0001)
Size of Community	−0.0329**	−0.0318**	−0.0317**	−0.1175**	−0.1160**	−0.1151**	−0.0375**	−0.0375**	−0.0370**
	(0.0011)	(0.0010)	(0.0010)	(0.0062)	(0.0062)	(0.0061)	(0.0001)	(0.0048)	(0.0044)
Conference/Journal	−0.0553**	−0.0403**	−0.0388**	−0.3068**	−0.2446**	−0.2275**	−0.3208**	−0.2657**	−0.2576**
(1 if a conference paper)	(0.0097)	(0.0097)	(0.0097)	(0.057)	(0.0572)	(0.0572)	(0.0372)	(0.0374)	(0.0375)
Firm-resource Sensitive Field	−0.0876**	−0.0880**	−0.0887**	−0.5663**	−0.5641**	−0.5697**	−0.3301**	−0.3324**	−0.3355**
(1 if related to vision or NLP)	(0.0087)	(0.0087)	(0.0087)	(0.0469)	(0.0468)	(0.0468)	(0.031)	(0.0309)	(0.0309)
Constant	−85.1112**	−83.5991**	−83.8870**	−664.39**	−658.13**	−660.39**	−681.57**	−675.68**	−677.43**
	(2.7033)	(2.7014)	(2.7016)	(17.2376)	(17.2062)	(17.1964)	(10.3709)	(10.3656)	(10.3660)
Adjusted R-square	0.0594	0.0624	0.0627	0.1116	0.1159	0.1174	0.1272	0.1302	0.1309

† $p<0.1$, * $p<0.05$, ** $p<0.01$. Standard errors are in parentheses

The OLS results for the distance to the knowledge trajectory [1) Models 1–3 in Table 4] are robust to the log-normal models [1) Models 4–6 in Table 6], the models excluding papers in which five year citation is zero [1) Models 7–9 in Table 6] and the models excluding papers in which the community size is zero [1) Models 10–12 in Table 6]. In Models 4–6, before taking the log-transformation, distance values with zero are replaced with a small positive value, 0.01

Table 7 Additional regression analyses for robustness check (H2a and H2b) – Using data after winsorizing outliers

	H2a) OLS ($n = 49{,}550$)			H2b) OLS ($n = 49{,}550$)		
	DV: z-min (novelty)			DV: z-median (conventionality)		
	2a) Model 4	2a) Model 5	2a) Model 6	2b) Model 4	2b) Model 5	2b) Model 6
Firm Paper1 (*Firm Involved or Not*)		−0.0900†			0.1687**	
(1 if including firm(s))		(0.0616)			(0.0511)	
Firm Paper2 (*Firm Only, Non-firm Only, or Firm/Non-firm Collaboration*)						
(1 if firm and non-firm collaboration)			−0.1202*			0.1484**
			(0.0684)			(0.0567)
(2 if firm only)			0.0189			0.2418**
			(0.1235)			(0.1024)
Number of References	−0.0052**	−0.0052**	−0.0052**	0.0083**	0.0083**	0.0083**
	(0.0012)	(0.0012)	(0.0012)	(0.0010)	(0.0010)	(0.0010)
Publication Year (t)	0.1349**	0.1344**	0.1343**	0.0194**	0.0202**	0.0201**
	(0.0071)	(0.0071)	(0.0071)	(0.0059)	(0.0059)	(0.0059)
Number of Authors	−0.0823**	−0.0815**	−0.0814**	−0.0358**	−0.0373**	−0.0373**
	(0.0138)	(0.0138)	(0.0138)	(0.0115)	(0.0115)	(0.0115)
Authors' Average Number of CS Publications (t-4 ~ t)	−0.0157**	−0.0157**	−0.0156**	0.0120**	0.0121**	0.0122**
	(0.0022)	(0.0022)	(0.0022)	(0.0018)	(0.0018)	(0.0018)
Authors' Average h-index at t	0.0436**	0.0439**	0.0438**	0.0346**	0.0340**	0.0339**
	(0.0049)	(0.0049)	(0.0049)	(0.0040)	(0.0040)	(0.0040)
Number of Affiliations	0.1116**	0.1198**	0.1252**	0.0752**	0.0599**	0.0635**
	(0.0288)	(0.0293)	(0.0298)	(0.0239)	(0.0244)	(0.0248)
Affiliations' Average Number of CS Publications (t-4 ~ t)	0.0000	0.0000†	0.0000†	0.0000**	0.0000**	0.0000**
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)

Table 7 (continued)

	H2a) OLS ($n=49{,}550$)			H2b) OLS ($n=49{,}550$)		
	DV: z-min (novelty)			DV: z-median (conventionality)		
	2a) Model 4	2a) Model 5	2a) Model 6	2b) Model 4	2b) Model 5	2b) Model 6
Affiliations' Average h-index at *t*	−0.0008**	−0.0008**	−0.0008**	−0.0004**	−0.0003*	−0.0003*
	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.0001)
Size of Community	0.0464**	0.0469**	0.0468**	0.0189**	0.0179**	0.0179**
	(0.0052)	(0.0052)	(0.0052)	(0.0043)	(0.0043)	(0.0043)
Conference/Journal (1 if a conference paper)	−0.0682†	−0.0594	−0.0609	0.4671**	0.4508**	0.4497**
	(0.0490)	(0.0494)	(0.0494)	(0.0407)	(0.0410)	(0.0410)
Firm-resource Sensitive Field (1 if related to vision or NLP)	−0.6561**	−0.6567**	−0.6558**	0.5208**	0.5219**	0.5225**
	(0.0431)	(0.0431)	(0.0431)	(0.0357)	(0.0357)	(0.0357)
Constant	−274.19**	−273.34**	−272.97**	−38.25**	−39.85**	−39.59**
	(14.4267)	(14.4382)	(14.4429)	(11.9749)	(11.9836)	(11.9877)
Adjusted R-square	0.0195	0.0196	0.0196	0.0225	0.0227	0.0227

† $p<0.1$, * $p<0.05$, **$p<0.01$. Standard errors are in parentheses

The results are robust when the dependent variables (z-min in H2a and z-median in H2b) are winsorized at the 0.5 and 99.5 percentiles. After removing the outliers, the z-min and z-median range from −17.67 to 17.81 and −5.99 to 26. 04, respectively. These values differ from those derived from the original dataset: z-min and z-median, ranging from −20.03 to 114.61 and −16.65 to 118.91, respectively (see Table 2)

Appendix E

Table 8 The fields of study used to extract computer science papers from MAG

Fields of Study
in ('Computer science', 'Algorithm', 'Artificial intelligence', 'Computational science', 'Computer architecture', 'Computer engineering', 'Computer graphics', 'Computer hardware', 'Computer network', 'Computer security', 'Computer vision', 'Data mining', 'Data science', 'Database', 'Distributed computing', 'Embedded system', 'Human–computer interaction', 'Information retrieval', 'Internet privacy', 'Knowledge management', 'Library science', 'Machine learning', 'Multimedia', 'Natural language processing', 'Operating system', 'Parallel computing', 'Pattern recognition', 'Programming language', 'Real-time computing', 'Simulation', 'Software engineering', 'Speech recognition', 'Telecommunications', 'Theoretical computer science', 'World Wide Web')

MAG provides hierarchical information for the field of study. The fields suggested in this table include "Computer Science and its 34 child fields

Table 9 The fields of study used to extract computer vision- or NLP-related papers from MAG

Fields of Study
in ('Computer vision', 'Computer graphics', 'Natural language processing', 'Speech recognition')

The suggested fields are child fields of "Computer science"

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2020R1A2C2005026). The authors are grateful to Ben Martin, Paul Nightingale, and Ed Steinmueller at the Science Policy Research Unit, University of Sussex for their helpful comments on the early draft of this manuscript. This study is a revised version of the first author's doctoral thesis chapter supervised by the second author at Yonsei University.

Data availability The data that support the findings of this study are available from the corresponding author upon request.

Declarations

Conflicts of interest There is no potential conflicts of interest.

Informed consent This research does not involve human participants.

References

Agrawal A, Gans J, Goldfarb A (2018) Prediction machines: the simple economics of artificial intelligence. Harvard Business Press

Alexy O, George G, Salter AJ (2013) Cui bono? The selective revealing of knowledge and its implications for innovative activity. Acad Manag Rev 38(2):270–291
Anthes G (2017) Artificial intelligence poised to ride a new wave. Commun ACM 60(7):19–21
Antonelli C (2001) The microeconomics of technological systems. Oxford University Press, Oxford, England
Arora A, Belenzon S, Patacconi A (2018) The decline of science in corporate R&D. Strateg Manag J 39(1):3–32
Arrow KJ (1972) Economic welfare and the allocation of resources for invention. In: Readings in industrial economics. Palgrave, London, pp 219–236
Baruffaldi S, Poege F (2020) A firm scientific community: industry participation and knowledge diffusion. Max Planck Institute for Innovation & Competition Research Paper, pp 20–10
Bijker WE, Hughes TP, Pinch TJ (eds) (1989) The social construction of technological systems: new directions in the sociology and history of technology. MIT Press
Blei DM, Lafferty JD (2006) Dynamic topic models. In: Proceedings of the 23rd international conference on machine learning. ACM, pp 113–120
Blei DM, Lafferty JD (2007) A correlated topic model of science. Ann Appl Stat 1(2):634–642
Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):P10008
Boden MA (2016) AI: its nature and future. Oxford University Press
Bolander T (2019) What do we loose when machines take the decisions? J Manag Gov 23(4):849–867
Bornmann L, Schier H, Marx W, Daniel HD (2012) What factors determine citation counts of publications in chemistry besides their quality? J Informetr 6(1):11–18
Carlsson B, Stankiewicz R (1991) On the nature, function and composition of technological systems. J Evol Econ 1(2):93–118
Chang J, Boyd-Graber J, Wang C, Gerrish S, Blei D (2009) Reading tea leaves: how humans interpret topic models. In: Proceedings of neutral information processing systems. Vancouver, B.C., Canada. Available at: http://goo.gl/QzeEv9. Accessed 6 Dec 2022
Chesbrough H (2003) Open innovation – the new imperative for creating and profiting from technology. Harvard Business School Press, Boston
Christensen CM, Rosenbloom RS (1995) Explaining the attacker's advantage: technological paradigms, organizational dynamics, and the value network. Res Policy 24(2):233–257
Clauset A, Newman ME, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70(6):066111
Cohen WM, Levinthal DA (1990) Absorptive capacity: a new perspective on learning and innovation. Adm Sci Q 35(1):128–152
Dasgupta P, David PA (1987) Information disclosure and the economics of science and technology. In: Arrow and the ascent of modern economic theory. Palgrave Macmillan, London, pp 519–542
Denning PJ (2005) Is computer science? Commun ACM 48(4):27–31
Didegah F, Thelwall M (2013) Which factors help authors produce the highest impact research? Collaboration, journal and document properties. J Informetr 7(4):861–873
Dosi G (1982) Technological paradigms and technological trajectories. Res Policy 2(3):I47–I62
Etzkowitz H (2002) MIT and the rise of entrepreneurial science. Routledge, London
Etzkowitz H (2003) Research groups as 'quasi-firms': the invention of the entrepreneurial university. Res Policy 32(1):109–121
Fleming L, Sorenson O (2004) Science as a map in technological search. Strateg Manag J 25(8–9):909–928
Fontana R, Nuvolari A, Verspagen B (2009) Mapping technological trajectories as patent citation networks. An application to data communication standards. Econ Innov New Technol 18(4):311–336
Franceschet M, Costantini A (2010) The effect of scholar collaboration on impact and quality of academic papers. J Informetr 4(4):540–553
Freeman C (1997) The economics of industrial innovation, 3rd edn. Routledge
Freyne J, Coyle L, Smyth B, Cunningham P (2010) Relative status of journal and conference publications in computer science. Commun ACM 53(11):124–132
Garfield E (1979) Citation indexing. Its theory and application in science, technology and humanities. Wiley, New York
Gazni A, Didegah F (2011) Investigating different types of research collaboration and citation impact: a case study of Harvard University's publications. Scientometrics 87(2):251–265
Gibbons M, Limoges C, Nowotny H, Schwartzman S, Scott P, Trow M (1994) The New Production of Knowledge: the Dynamics of Science and Research in Contemporary Societies. Sage, London
Gibney E (2016) AI firms lure academics. Nature 532(7600):422–423

Grassano N, Camerani R, Rotolo D (2019) Do firms publish? A multi-sectoral analysis, DRUID19 Copenhagen Business School, Copenhagen, Denmark June 19-21, 2019
Hall D, Jurafsky D, Manning CD (2008) Studying the histories of ideas using topic models. In proceedings of the conference on empirical methods in natural language processing. Honolulu, Hawaii. Available at: http://dl.acm.org/citation.cfm?id=1613763. Accessed 15 April 2014
Hartmann P, Henkel J (2019) The rise of corporate science in AI: data as a strategic resource. Academy of Management Discoveries
Henderson RM (1993) Underinvestment and incompetence as responses to radical innovation: evidence from the photolithographic industry. RAND J Econ 24(2):248–270
Hicks D (1995) Published papers, tacit competencies and corporate management of the public/private character of knowledge. Ind Corp Chang 4(2):401–424
Hilgartner S, Brandt-Rauf S (1998) Controlling data and resources: access strategies in molecular genetics. In: David P, Steinmueller E (eds) Information technology and the productivity paradox. Harwood Academic Publishers, Newark
Hinton GE, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554
Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci 102(46):16569–16572
Hughes TP (1987) The evolution of large technological systems. In: Bijker WE, Hughes TP, Pinch T (eds) The social construction of technological systems: new directions in the sociology and history of technology. MIT Press, pp 51–82
Hummon NP, Dereian P (1989) Connectivity in a citation network: the development of DNA theory. Soc Networks 11(1):39–63
Hutson M (2018) Has artificial intelligence become alchemy? Science 360(6388):478
Kaplan S, Vakili K (2015) The double-edged sword of recombination in breakthrough innovation. Strateg Manag J 36(10):1435–1457
Katz JS, Martin BR (1997) What is research collaboration? Res Policy 26(1):1–18
Kim D, Cerigo DB, Jeong H, Youn H (2016) Technological novelty profile and invention's future impact. EPJ Data Science 5(1):8
Kline JK, Rosenberg N (1986) An overview of innovation. In: Landau R, Rosenburg N (eds) The positive sum strategy: harnessing technology for economic growth Washington DC
Klinger J, Mateos-Garcia J, Stathoulopoulos K (2018) Deep learning, deep change? Mapping the development of the Articial Intelligence General Purpose Technology, arXiv preprint arXiv:1808.06355
Koumpis K, Pavitt K (1999) Corporate activities in speech recognition and natural language: another" new science"-based technology. Int J Innov Manag 3(03):335–366
LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
Li Y, Youtie J, Shapira P (2015) Why do technology firms publish scientific papers? The strategic use of science by small and midsize enterprises in nanotechnology. J Technol Transf 40(6):1016–1033
Liu J, Tian J, Kong X, Lee I, Xia F (2019) Two decades of information systems: a bibliometric review. Scientometrics 118(2):617–643
Lundvall BÅ, Rikap C (2022) China's catching-up in artificial intelligence seen as a co-evolution of corporate and national innovation systems. Res Policy 51(1):104395
Mansell R, Steinmueller WE (2020) Advanced introduction to platform economics. Edward Elgar Publishing
Martinelli A, Nomaler Ö (2014) Measuring knowledge persistence: a genetic approach to patent citation networks. J Evol Econ 24(3):623–652
McKelvey M, Rake B (2020) Exploring scientific publications by firms: what are the roles of academic and corporate partners for publications in high reputation or high impact journals? Scientometrics 122(3):1323–1360
Merton RK (1968) The Matthew effect in science: the reward and communication systems of science are considered. Science 159(3810):56–63
Mowery DC, Rosenberg N (1999) Paths of innovation: technological change in 20th-century America. Cambridge University Press
Mulder K, Knot M (2001) PVC plastic: A history of systems development and entrenchment. Technol Soc 23(2):265–286
Murmann JP, Frenken K (2006) Toward a systematic framework for research on dominant designs, technological innovations, and industrial change. Res Policy 35(7):925–952
Narin F, Hamilton KS, Olivastro D (1997) The increasing linkage between US technology and public science. Res Policy 26(3):317–330

Nelson R, Winter SG (1982) An evolutionary theory of economic change. Harvard University Press, Cambridge
Nieto MJ, Santamaría L (2007) The importance of diverse collaborative networks for the novelty of product innovation. Technovation 27(6–7):367–377
Nilsson NJ (2009) The quest for artificial intelligence. Cambridge University Press
Norton S (2016) CIO explainer: what is artificial intelligence? wsj.com, July 18. http://blogs.wsj.com/cio/2016/07/18/cio-explainer-what-is-artificial-intelligence/. Accessed 6 Dec 2022
Parnas DL (2017) The real risks of artificial intelligence. Commun ACM 60(10):27–31
Pavitt K (1987) The objectives of technology policy. Sci Public Policy 14(4):182–188
Perkmann M, Walsh K (2009) The two faces of collaboration: impacts of university-industry relations on public research. Ind Corp Chang 18(6):1033–1065
Pisano GP (2010) The evolution of science-based business: innovating how we innovate. Ind Corp Chang 19(2):465–482
Polidoro F Jr, Toh PK (2011) Letting rivals come close or warding them off? The effects of substitution threat on imitation deterrence. Acad Manag J 54(2):369–392
Rosenberg N (1990) Why do firms do basic research (with their own money)? Res Policy 19(2):165–174
Rosenberg N, Nelson RR (1994) American universities and technical advance in industry. Res Policy 23(3):323–348
Rothaermel FT, Thursby M (2007) The nanotech versus the biotech revolution: sources of productivity in incumbent firm research. Res Policy 36(6):832–849
Simeth M, Raffo JD (2013) What makes companies pursue an open science strategy? Res Policy 42(9):1531–1543
Simon HA (1995) Artificial intelligence: an empirical science. Artif Intell 77(1):95–127
Simon HA (1996) The sciences of the artificial. MIT Press
Sinha A, Shen Z, Song Y, Ma H, Eide D, Hsu BJ, Wang K (2015) An overview of microsoft academic service (mas) and applications. In: Proceedings of the 24th international conference on world wide web, pp 243–246
Stokes DE (1997) Pasteur's quadrant—basic science and technological innovation. Brookings Institution Press, Washington, DC
Takeishi A, Lee KJ (2005) Mobile music business in Japan and Korea: copyright management institutions as a reverse salient. J Strateg Inf Syst 14(3):291–306
Teece DJ (1986) Profiting from technological innovation: implications for integration, collaboration, licensing and public policy. Res Policy 15(6):285–305
Teece DJ (2018) Profiting from innovation in the digital economy: enabling technologies, standards, and licensing models in the wireless world. Res Policy 47(8):1367–1387
Uzzi B, Mukherjee S, Stringer M, Jones B (2013) Atypical combinations and scientific impact. Science 342(6157):468–472
Vardi MY (2009) Conferences vs. journals in computing research. Commun ACM 52(5):5–5
Verspagen B (2007) Mapping technological trajectories as patent citation networks: a study on the history of fuel cell research. Adv Complex Syst 10(01):93–115
Vincenti WG (1990) What engineers know and how they know it: analytical studies from Aeronatical history. Johns Hopkins studies in the history of technology. The Johns Hopkins University Press, Baltimore, MD
Waters R (2015) Investor rush to artificial intelligence is real deal. Financial Times. San Francisco. Retrieved from http://www.ft.com/cms/s/2/019b3702-92a2-11e4-a1fd-00144feabdc0.html#axzz48ZGxiCut. Accessed 6 Dec 2022
WIPO (2019) WIPO Technology Trends 2019 – Artificial Intelligence. URL: https://www.wipo.int/publications/en/details.jsp?id=4386. Accessed 6 Dec 2022
Zucker LG, Darby MR, Armstrong J (2002) Commercializing knowledge: university science, knowledge capture, and firm performance in biotechnology. Manag Sci 48(1):138–153