CHAPTER 1:

Research & Development

Artificial Intelligence
Index Report 2021

CHAPTER 1:

Chapter Preview

ACCESS THE PUBLIC DATA

Overview

The report opens with an overview of the research and development (R&D) efforts in artificial intelligence (AI) because R&D is fundamental to AI progress. Since the technology first captured the imagination of computer scientists and mathematicians in the 1950s, AI has grown into a major research discipline with significant commercial applications. The number of AI publications has increased dramatically in the past 20 years. The rise of AI conferences and preprint archives has expanded the dissemination of research and scholarly communications. Major powers, including China, the European Union, and the United States, are racing to invest in AI research. The R&D chapter aims to capture the progress in this increasingly complex and competitive field.

This chapter begins by examining AI publications—from peer-reviewed journal articles to conference papers and patents, including the citation impact of each, using data from the Elsevier/Scopus and Microsoft Academic Graph (MAG) databases, as well as data from the arXiv paper preprint repository and Nesta. It examines contributions to AI R&D from major AI entities and geographic regions and considers how those contributions are shaping the field. The second and third sections discuss R&D activities at major AI conferences and on GitHub.

CHAPTER HIGHLIGHTS

- The number of AI journal publications grew by 34.5% from 2019 to 2020—a much higher percentage growth than from 2018 to 2019 (19.6%).

- In every major country and region, the highest proportion of peer-reviewed AI papers comes from academic institutions. But the second most important originators are different: In the United States, corporate-affiliated research represents 19.2% of the total publications, whereas government is the second most important in China (15.6%) and the European Union (17.2%).

- In 2020, and for the first time, China surpassed the United States in the share of AI journal citations in the world, having briefly overtaken the United States in the overall number of AI journal publications in 2004 and then retaken the lead in 2017. However, the United States has consistently (and significantly) more cited AI conference papers than China over the last decade.

- In response to COVID-19, most major AI conferences took place virtually and registered a significant increase in attendance as a result. The number of attendees across nine conferences almost doubled in 2020.

- In just the last six years, the number of AI-related publications on arXiv grew by more than sixfold, from 5,478 in 2015 to 34,736 in 2020.

- AI publications represented 3.8% of all peer-reviewed scientific publications worldwide in 2019, up from 1.3% in 2011.

AI publications include peer-reviewed publications, journal articles, conference papers, and patents. To track trends among these publications and to assess the state of AI R&D activities around the world, the following datasets were used: the Elsevier/Scopus database for peer-reviewed publications; the Microsoft Academic Graph (MAG) database for all journals, conference papers, and patent publications; and arXiv and Nesta data for electronic preprints.

1.1 PUBLICATIONS

PEER-REVIEWED AI PUBLICATIONS

This section presents data from the Scopus database by Elsevier. Scopus contains 70 million peer-reviewed research items curated from more than 5,000 international publishers. The 2019 version of the data shown below is derived from an entirely new set of publications, so figures of all peer-reviewed AI publications differ from those in previous years' AI Index reports. Due to changes in the methodology for indexing publications, the accuracy of the dataset increased from 80% to 84% (see the Appendix for more details).

Overview

Figure 1.1.1a shows the number of peer-reviewed AI publications, and Figure 1.1.1b shows the share of those among all peer-reviewed publications in the world. The total number of publications grew by nearly 12 times between 2000 and 2019. Over the same period, the percentage of peer-reviewed publications increased from 0.82% of all publications in 2000 to 3.8% in 2019.

By Region[1]

Among the total number of peer-reviewed AI publications in the world, East Asia & Pacific has held the largest share since 2004, followed by Europe & Central Asia, and North America (Figure 1.1.2). Between 2009 and 2019, South Asia and sub-Saharan Africa experienced the highest growth in terms of the number of peer-reviewed AI publications, increasing by eight- and sevenfold, respectively.

NUMBER of PEER-REVIEWED AI PUBLICATIONS, 2000-19
Source: Elsevier/Scopus, 2020 | Chart: 2021 AI Index Report

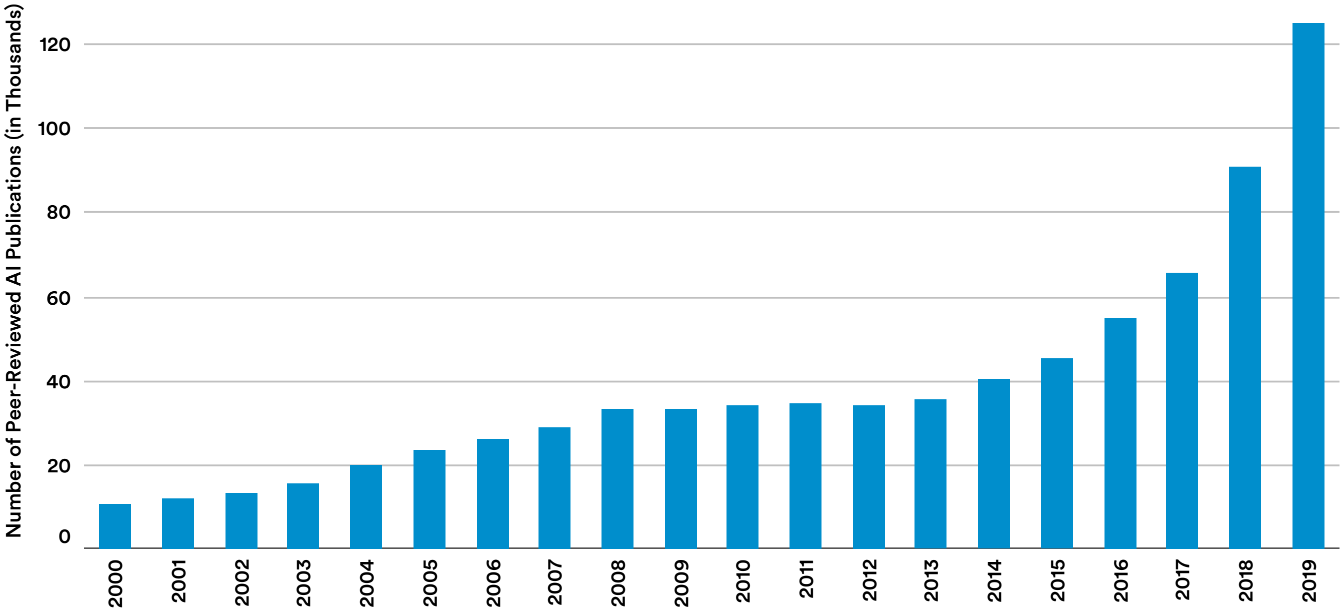

Figure 1.1.1a

[1] Regions in this chapter are classified according to the World Bank analytical grouping.

PEER-REVIEWED AI PUBLICATIONS (% of TOTAL), 2000-19

Source: Elsevier/Scopus, 2020 | Chart: 2021 AI Index Report

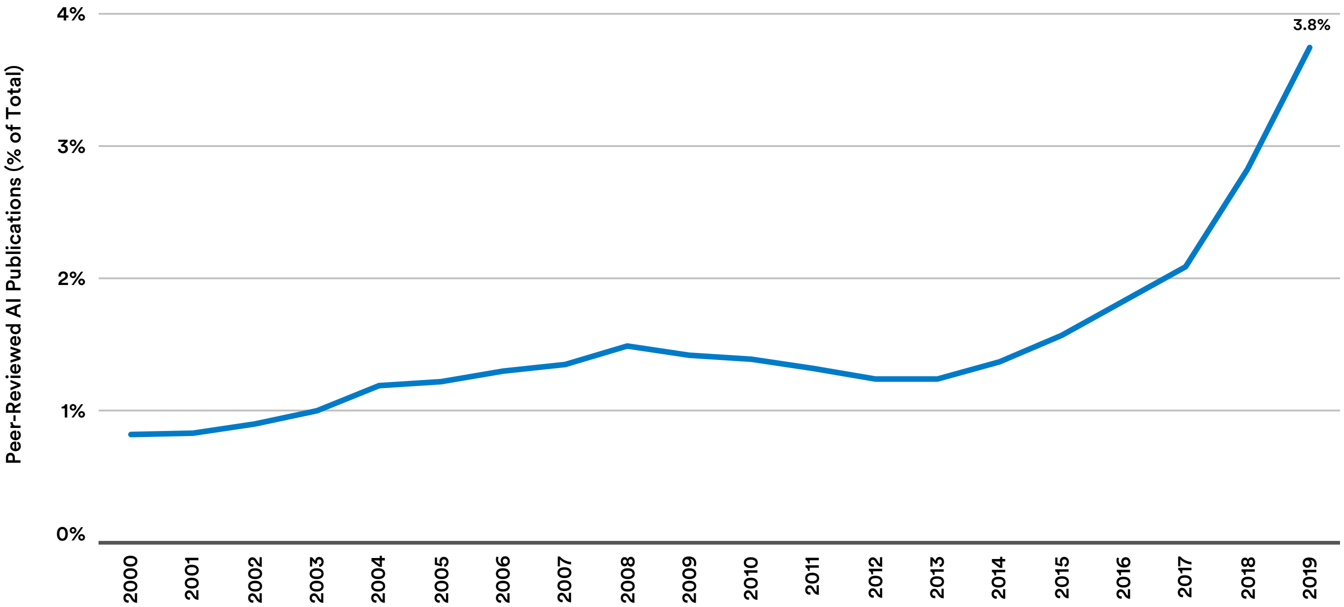

Figure 1.1.1b

PEER-REVIEWED AI PUBLICATIONS (% of TOTAL) by REGION, 2000-19

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

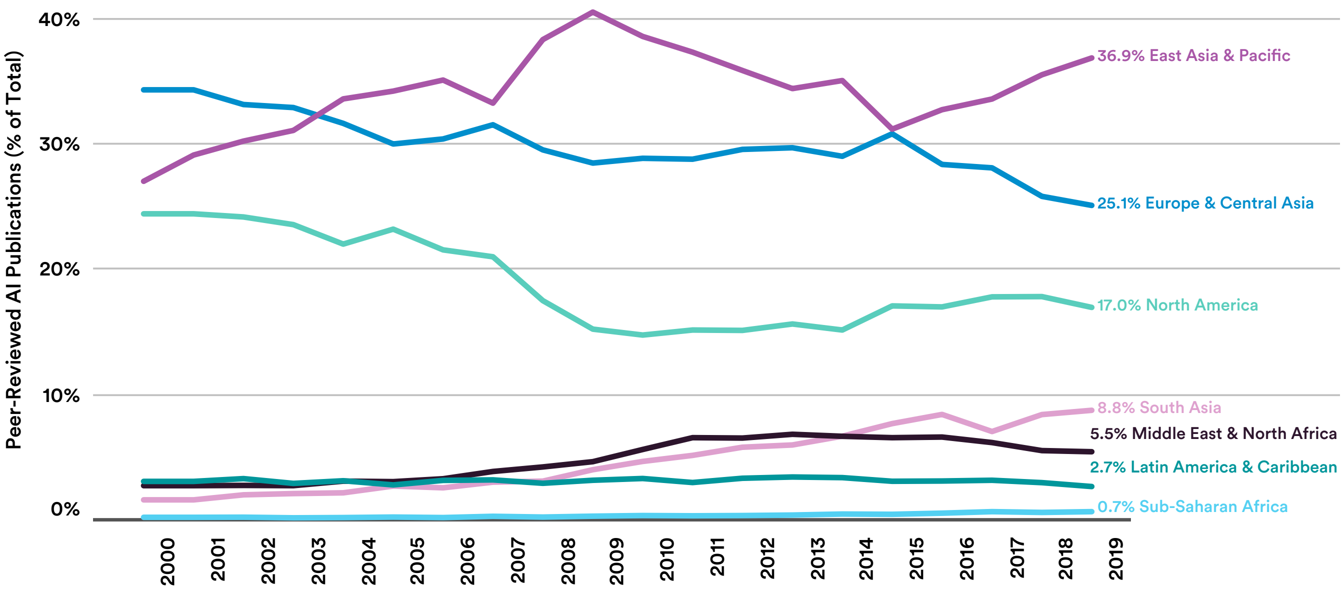

Figure 1.1.2

By Geographic Area

To compare the activity among the world's major AI players, this section shows trends of peer-reviewed AI publications coming out of China, the European Union, and the United States. As of 2019, China led in the share of peer-reviewed AI publications in the world, after overtaking the European Union in 2017 (Figure 1.1.3). It published 3.5 times more peer-reviewed AI papers in 2019 than it did in 2014—while the European Union published just 2 times more papers and the United States 2.75 times more over the same period.

PEER-REVIEWED AI PUBLICATIONS (% of WORLD TOTAL) by GEOGRAPHIC AREA, 2000-19

Source: Elsevier/Scopus, 2020 | Chart: 2021 AI Index Report

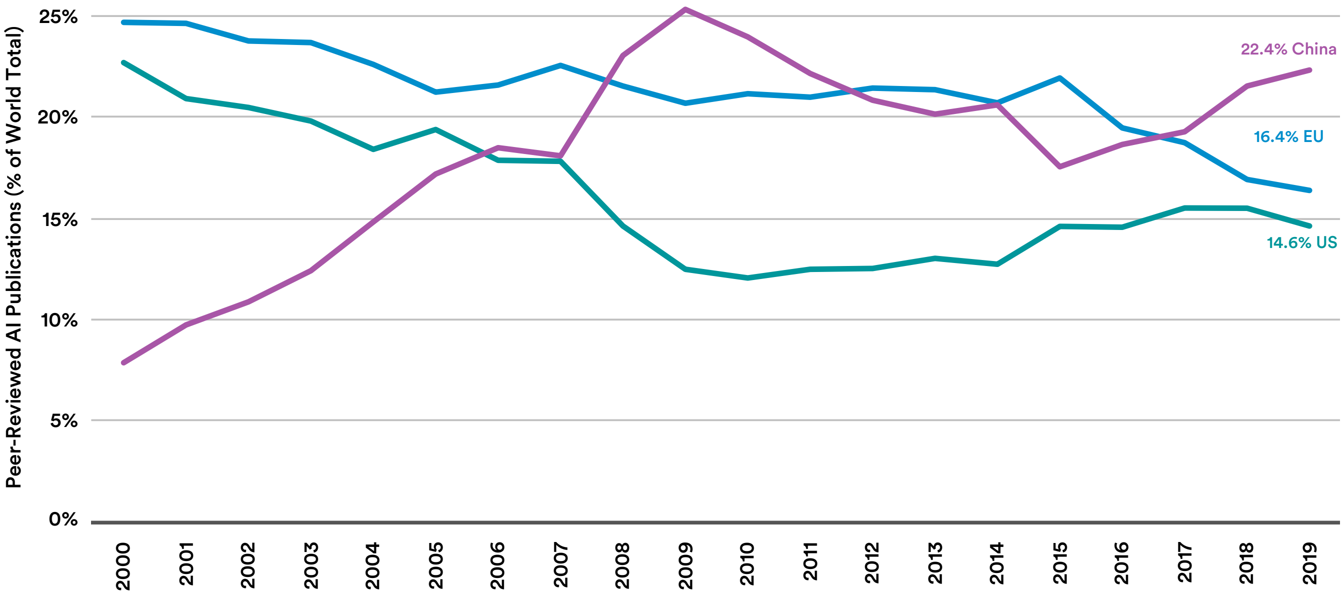

Figure 1.1.3

By Institutional Affiliation

The following charts show the number of peer-reviewed AI publications affiliated with corporate, government, medical, and other institutions in China (Figure 1.1.4a), the European Union (Figure 1.1.4b), and the United States (Figure 1.1.4c).[2] In 2019, roughly 95.4% of overall peer-reviewed AI publications in China were affiliated with the academic field, compared with 81.9% in the European Union and 89.6% in the United States. Those affiliation categories are not mutually exclusive, as some authors could be affiliated with more than one type of institution.

The data suggests that, excluding academia, government institutions—more than those in other categories—consistently contribute the highest percentage of peer-reviewed AI publications in both China and the European Union (15.6% and 17.2 %, respectively, in 2019), while in the United States, the highest portion is corporate-affiliated (19.2%).

NUMBER of PEER-REVIEWED AI PUBLICATIONS in CHINA by INSTITUTIONAL AFFILIATION, 2000-19

Source: Elsevier/Scopus, 2020 | Chart: 2021 AI Index Report

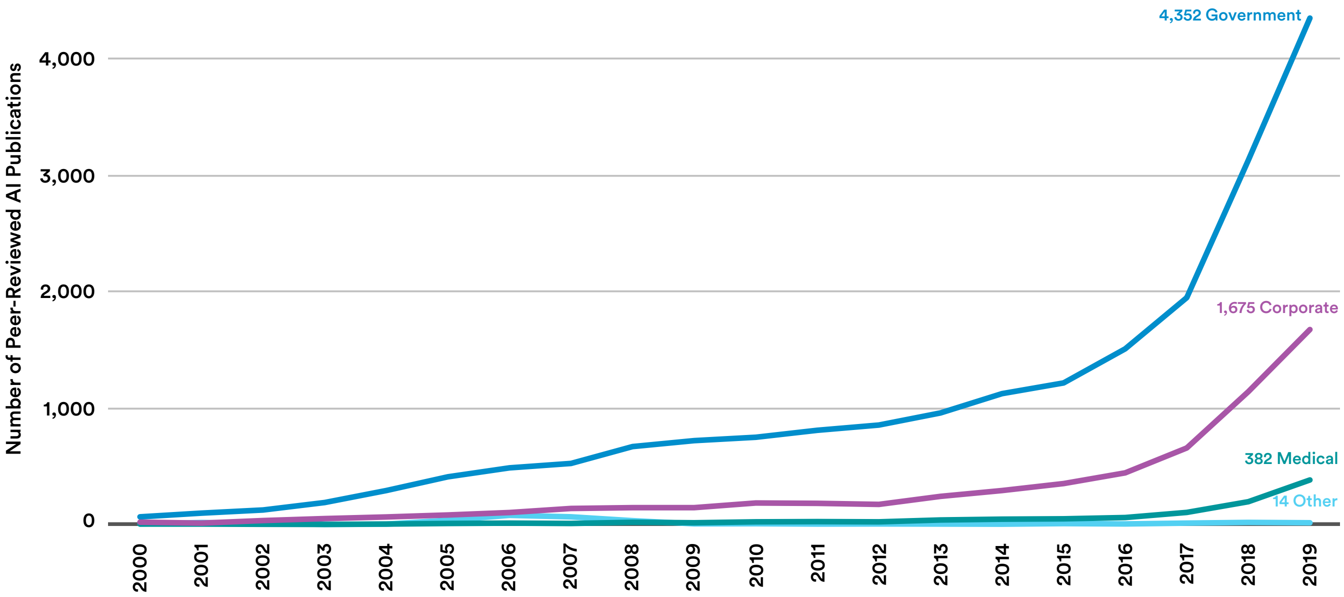

Figure 1.1.4a

[2] Across all three geographic areas, the number of papers affiliated with academia exceeds that of government-, corporate-, and medical-affiliated ones; therefore, the academia affiliation is not shown, as it would distort the graphs.

NUMBER of PEER-REVIEWED AI PUBLICATIONS in the EUROPEAN UNION by INSTITUTIONAL AFFILIATION, 2000-19

Source: Elsevier/Scopus, 2020 | Chart: 2021 AI Index Report

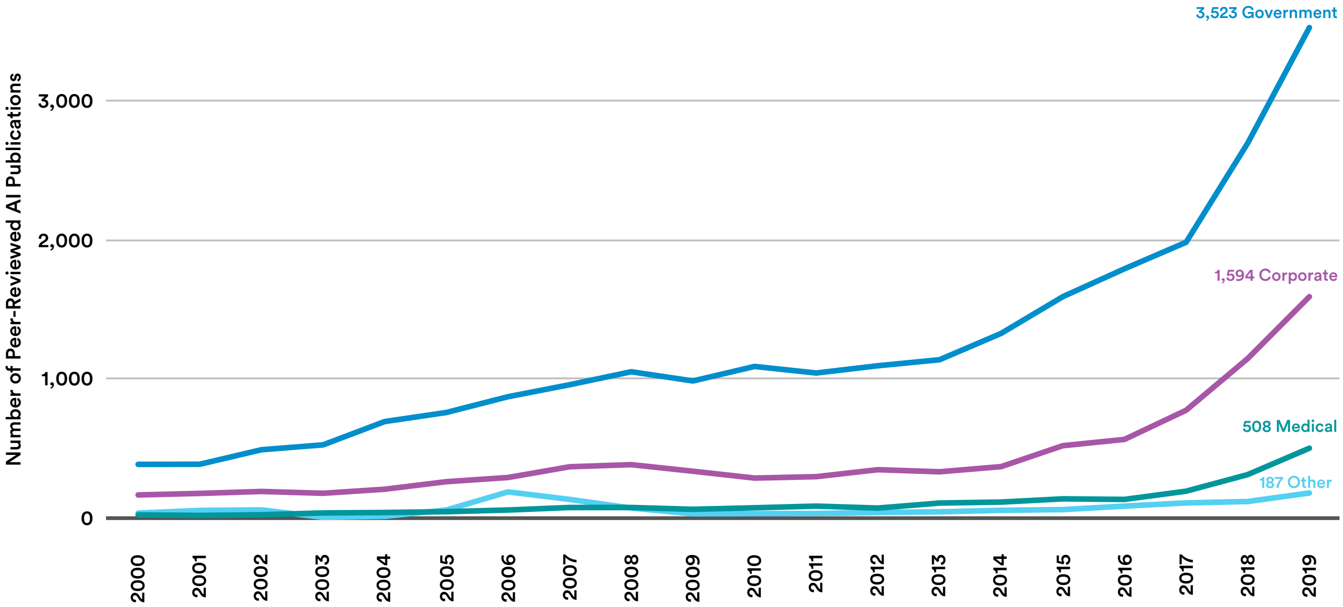

Figure 1.1.4b

NUMBER of PEER-REVIEWED AI PUBLICATIONS in the UNITED STATES by INSTITUTIONAL AFFILIATION, 2000-19

Source: Elsevier/Scopus, 2020 | Chart: 2021 AI Index Report

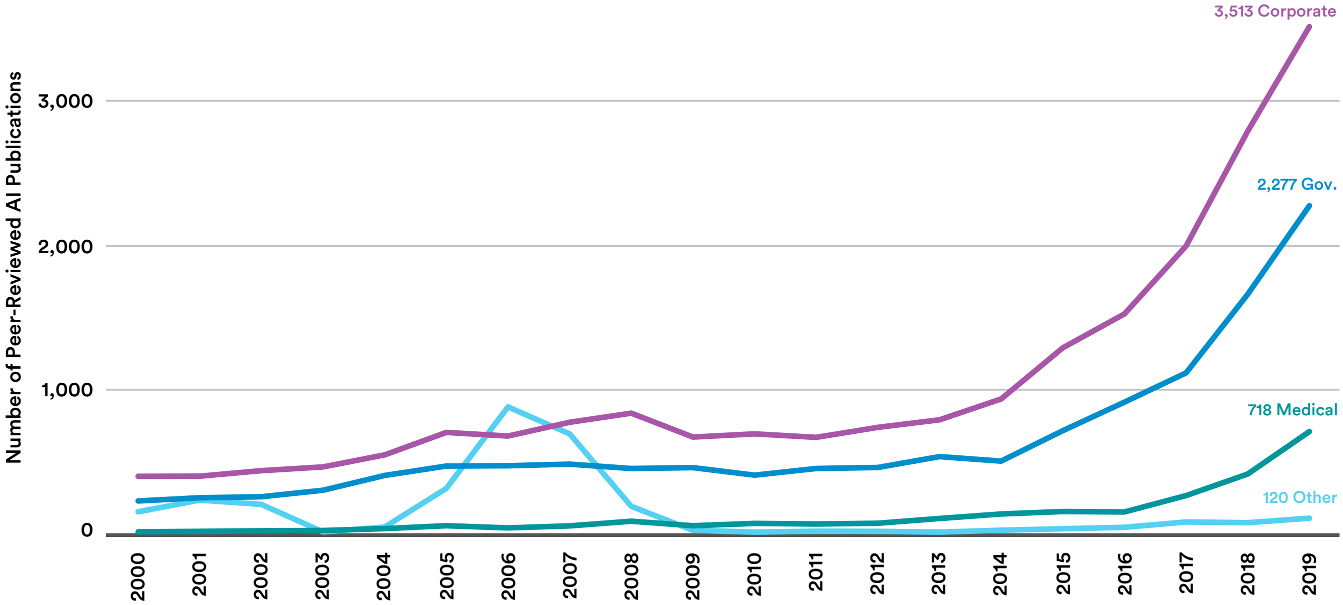

Figure 1.1.4c

Academic-Corporate Collaboration

Since the 1980s, the R&D collaboration between academia and industry in the United States has grown in importance and popularity, made visible by the proliferation of industry-university research centers as well as corporate contributions to university research. Figure 1.1.5 shows that between 2015 and 2019, the United States produced the highest number of hybrid academic-corporate, co-authored, peer-reviewed AI publications—more than double the amount in the European Union, which comes in second, followed by China in third place.

NUMBER of ACADEMIC-CORPORATE PEER-REVIEWED AI PUBLICATIONS by GEOGRAPHIC AREA, 2015-19 (SUM)

Source: Elsevier/Scopus, 2020 | Chart: 2021 AI Index Report

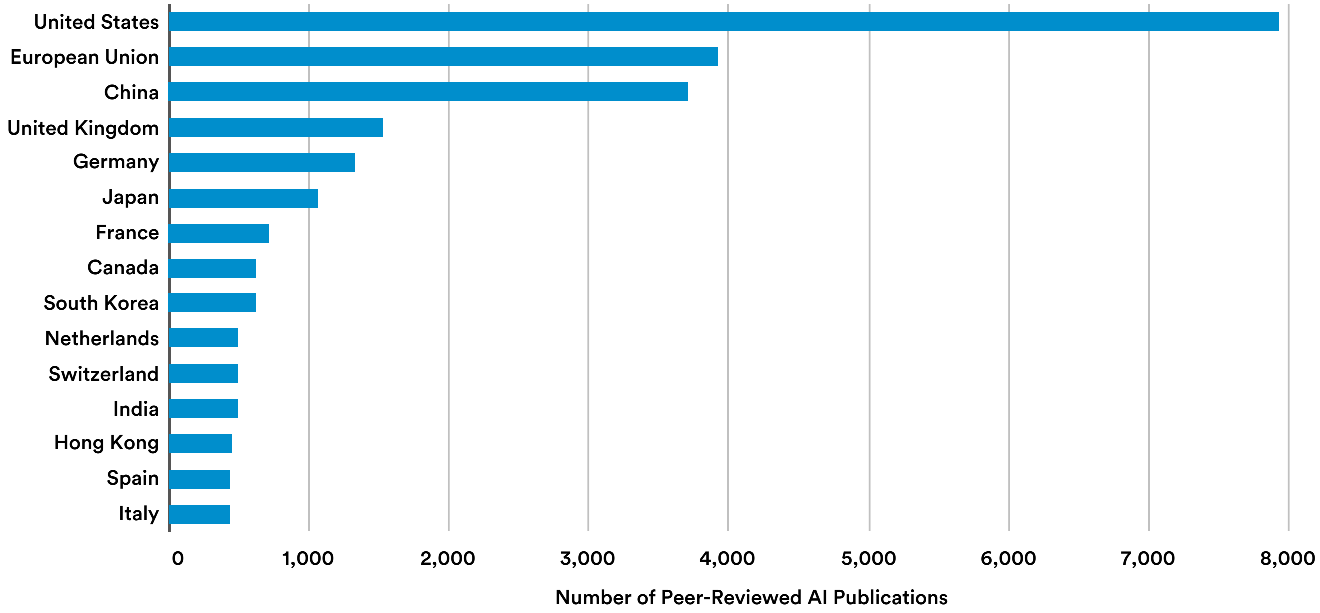

Figure 1.1.5

To assess how academic-corporate collaborations impact the Field-Weighted Citation Impact (FWCI) of AI publications from different geographic regions, see Figure 1.1.6. FWCI measures how the number of citations received by publications compares with the average number of citations received by other similar publications in the same year, discipline, and format (book, article, conference paper, etc.). A value of 1.0 represents the world average. More than or less than 1 means publications are cited more or less than expected, according to the world average. For example, an FWCI of 0.75 means 25% fewer citations than the world average.

The chart shows the FWCI for all peer-reviewed AI publications on the y-axis and the total number (on a log scale) of academic-corporate co-authored publications on the x-axis. To increase the signal-to-noise ratio of the FWCI metric, only countries that have more than 1,000 peer-reviewed AI publications in 2020 are included.

PEER-REVIEWED AI PUBLICATIONS' FIELD-WEIGHTED CITATION IMPACT and NUMBER of ACADEMIC-CORPORATE PEER-REVIEWED AI PUBLICATIONS, 2019

Source: Elsevier/Scopus, 2020 | Chart: 2021 AI Index Report

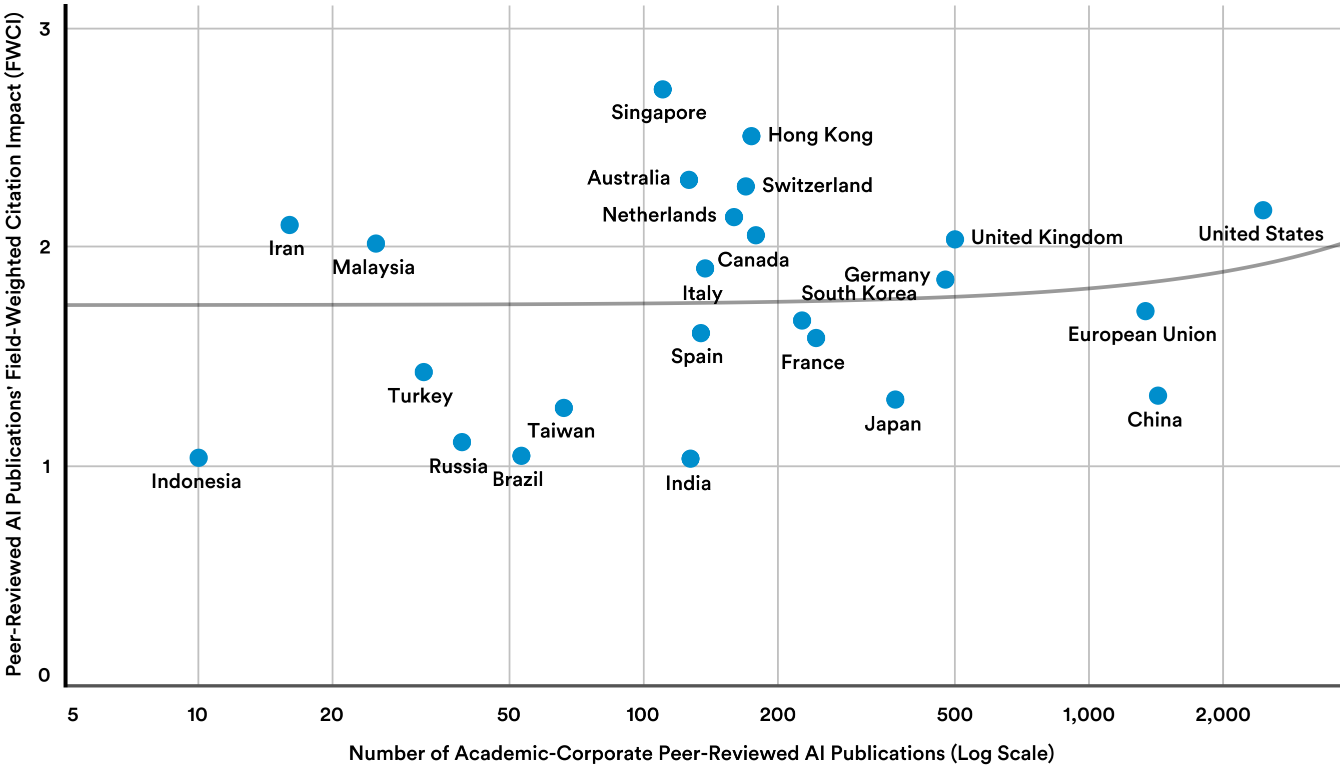

Figure 1.1.6

AI JOURNAL PUBLICATIONS

The next three sections chart the trends in the publication of AI journals, conference publications, and patents, as well as their respective citations that provide a signal for R&D impact, based on data from Microsoft Academic Graph. MAG[3] is a knowledge graph consisting of more than 225 million publications (at the end of November 2019).

Overview

Overall, the number of AI journal publications in 2020 is 5.4 times higher than it was in 2000 (Figure 1.1.7a). In 2020, the number of AI journal publications increased by 34.5% from 2019—a much higher percentage growth than from 2018 to 2019 (19.6%). Similarly, the share of AI journal publications among all publications in the world has jumped by 0.4 percentage points in 2020, higher than the average of 0.03 percentage points in the past five years (Figure 1.1.7b).

NUMBER of AI JOURNAL PUBLICATIONS, 2000-20
Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

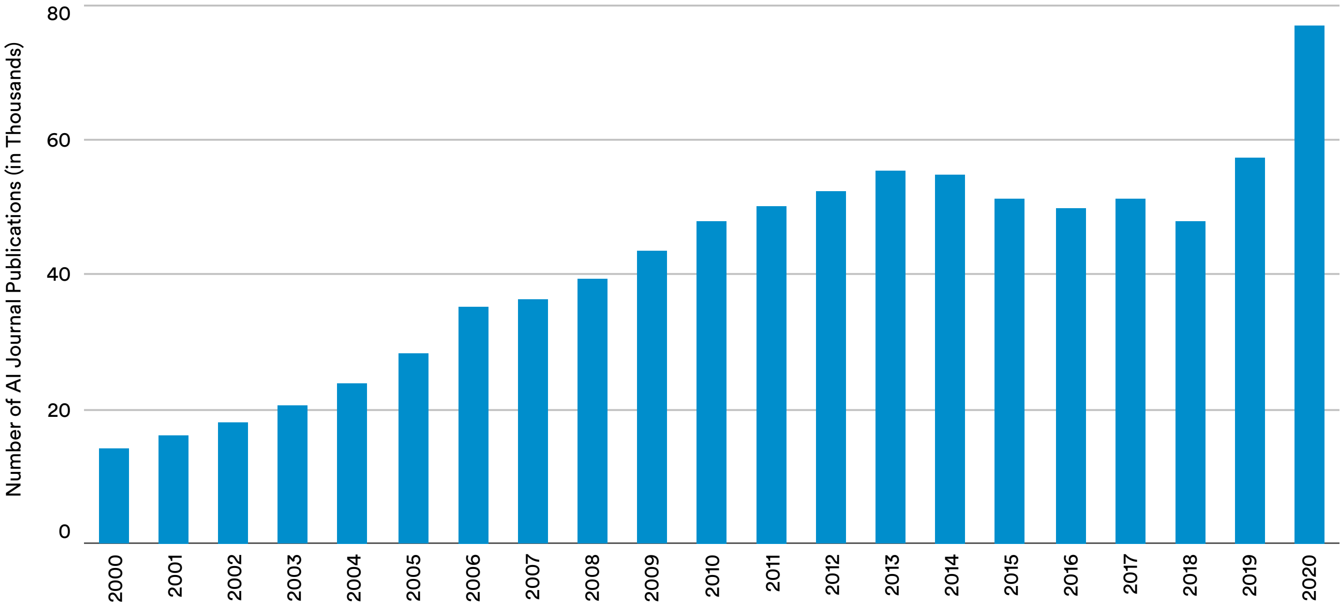

Figure 1.1.7a

AI JOURNAL PUBLICATIONS (% of ALL JOURNAL PUBLICATIONS), 2000-20
Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

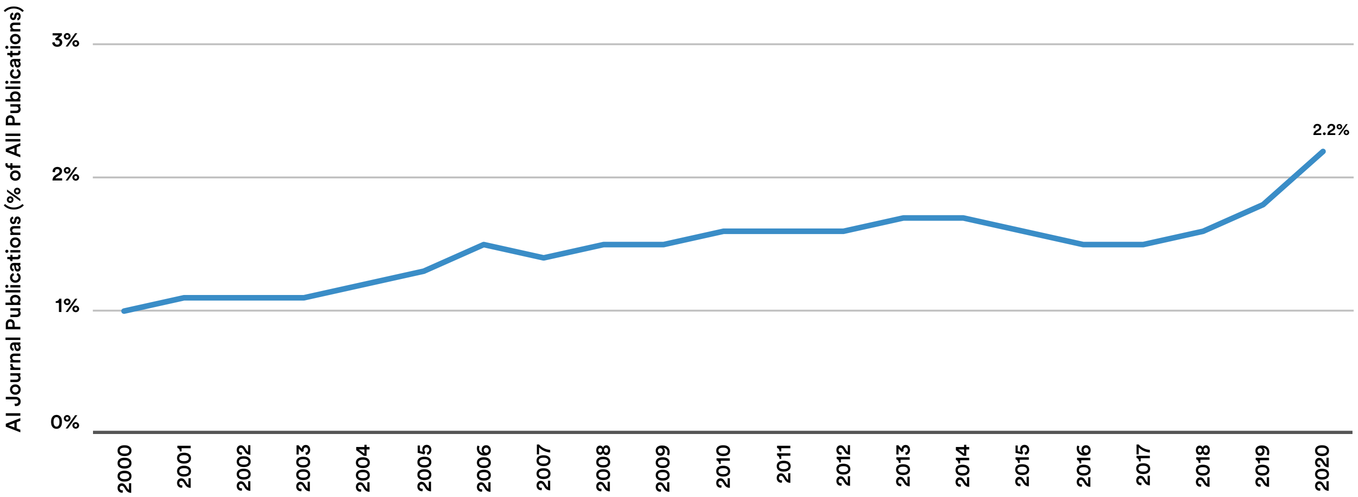

Figure 1.1.7b

[3] See "An Overview of Microsoft Academic Service (MAS) and Applications" and "A Review of Microsoft Academic Services for Science of Science Studies" for more details.

By Region

Figure 1.1.8 shows the share of AI journals—the dominant publication entity in terms of numbers in the MAG database—by region between 2000 and 2020. East Asia & Pacific, Europe & Central Asia, and North America are responsible for the majority of AI journal publications in the past 21 years, while the lead position among the three regions changes over time. In 2020, East Asia & Pacific held the highest share (26.7%), followed by Europe & Central Asia (13.3%) and North America (14.0%). Additionally, in the last 10 years, South Asia, and Middle East & North Africa saw the most significant growth, as the number of AI journal publications in those two regions grew six- and fourfold, respectively.

AI JOURNAL PUBLICATIONS (% of WORLD TOTAL) by REGION, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

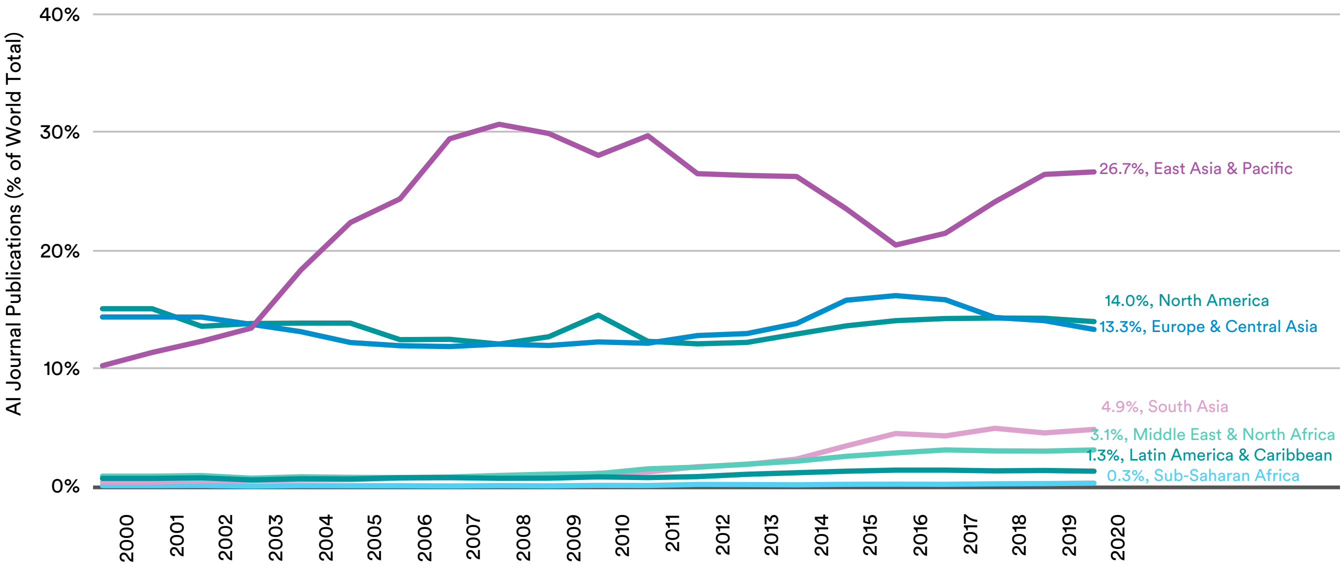

Figure 1.1.8

By Geographic Area

Figure 1.1.9 shows that among the three major AI powers, China has had the largest share of AI journal publications in the world since 2017, with 18.0% in 2020, followed by the United States (12.3%) and the European Union (8.6%).

AI JOURNAL PUBLICATIONS (% of WORLD TOTAL) by GEOGRAPHIC AREA, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

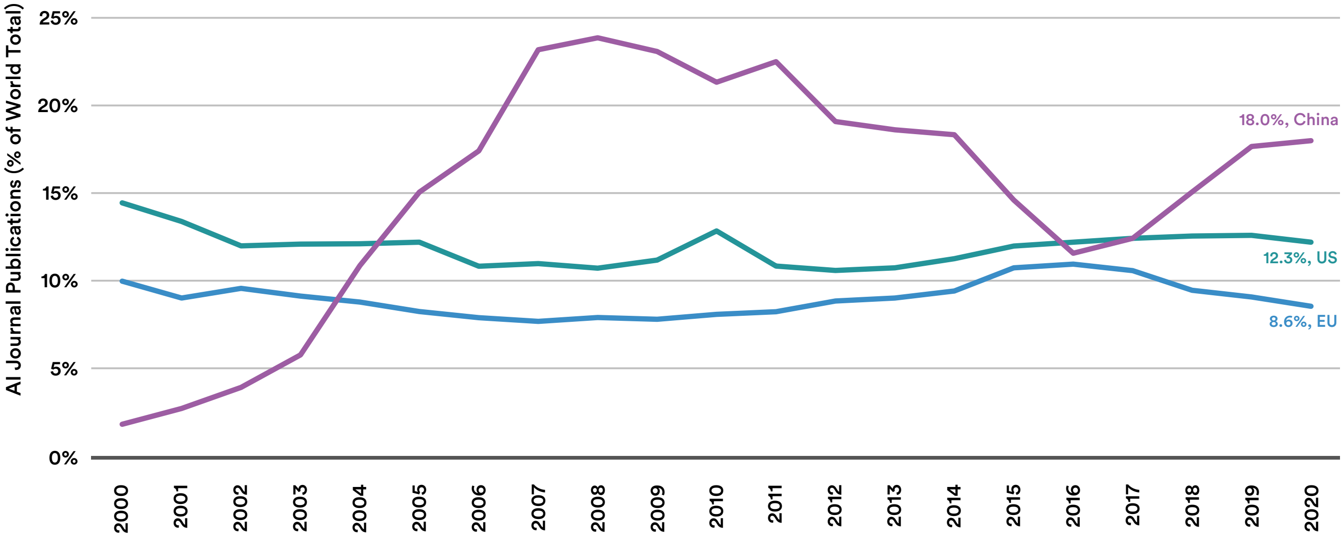

Figure 1.1.9

Citation

In terms of the highest share of AI journal citations, Figure 1.1.10 shows that China (20.7%) overtook the United States (19.8%) in 2020 for the first time, while the European Union continued to lose overall share.

AI JOURNAL CITATIONS (% of WORLD TOTAL) by GEOGRAPHIC AREA, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

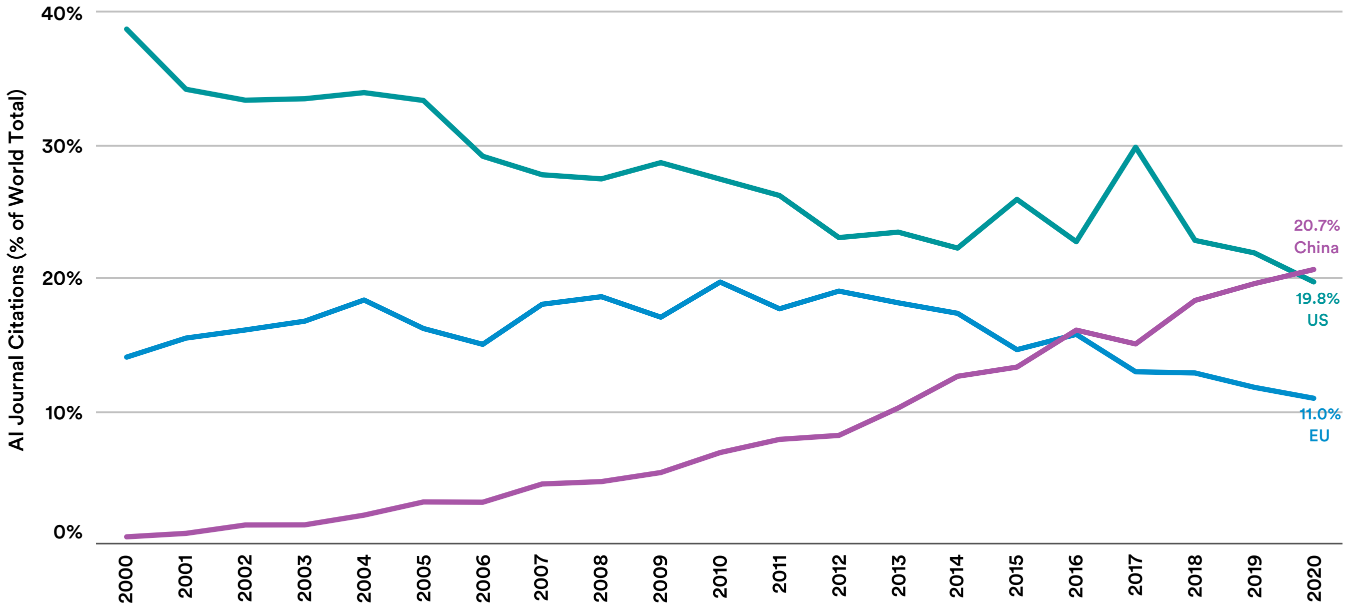

Figure 1.1.10

AI CONFERENCE PUBLICATIONS

Overview

Between 2000 and 2019, the number of AI conference publications increased fourfold, although the growth flattened out in the past ten years, with the number of publications in 2019 just 1.09 times higher than the number in 2010.[4]

NUMBER of AI CONFERENCE PUBLICATIONS, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

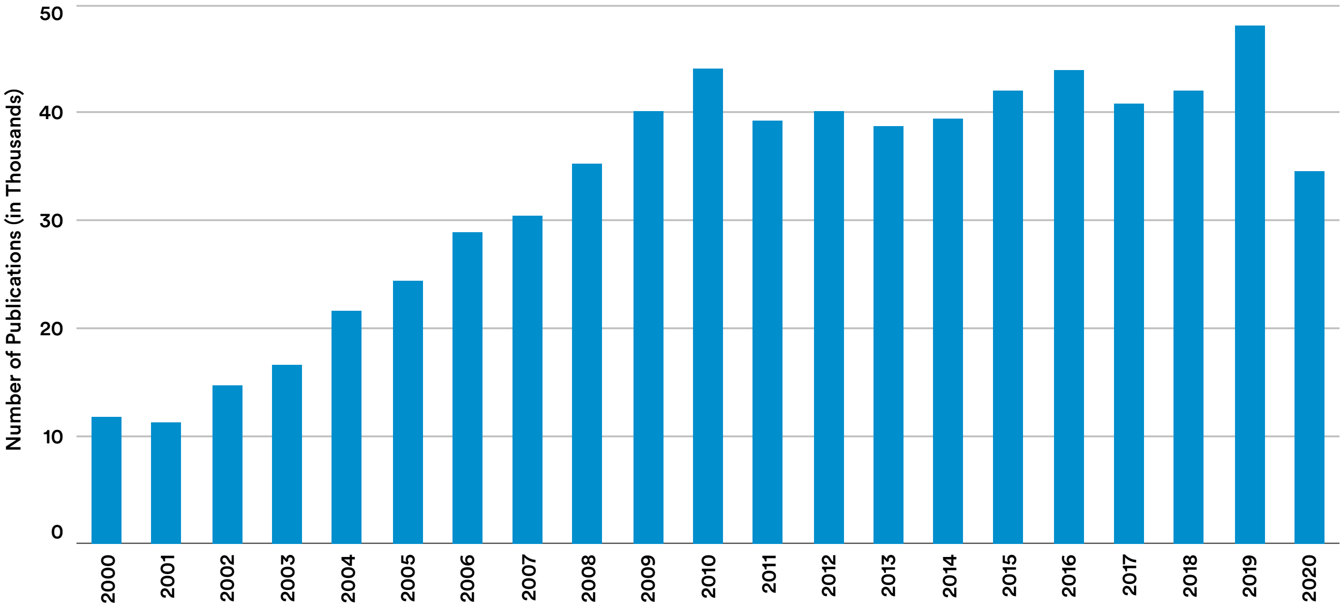

Figure 1.1.11a

AI CONFERENCE PUBLICATIONS (% of ALL CONFERENCE PUBLICATIONS), 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

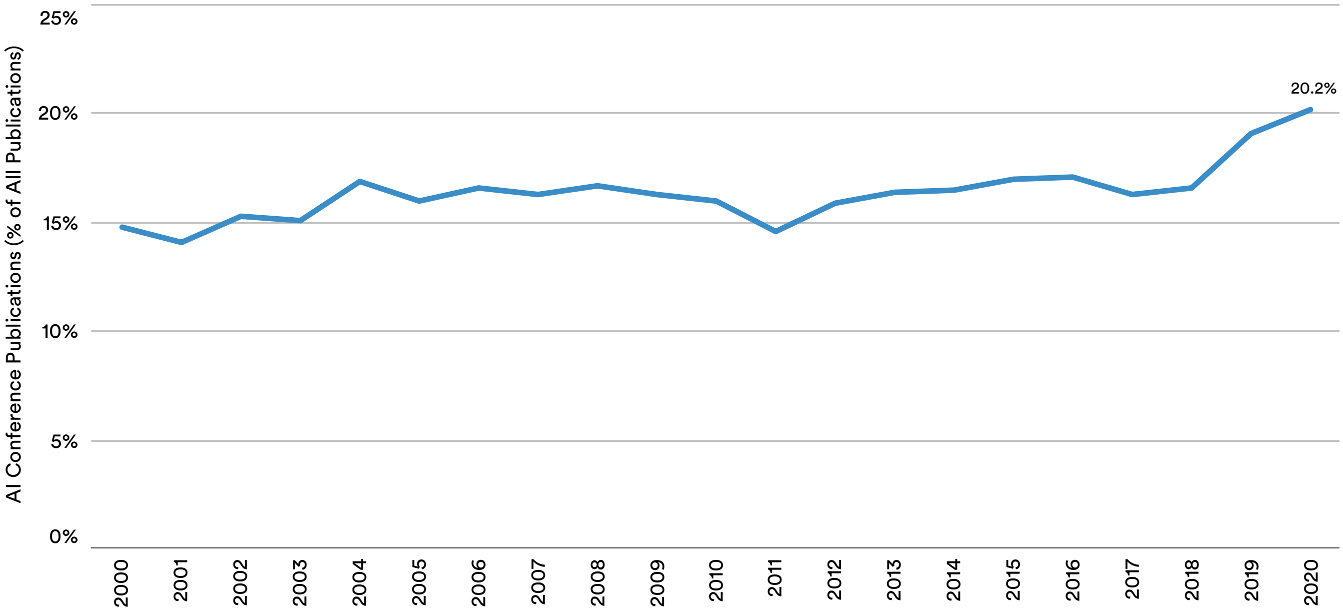

Figure 1.1.11b

[4] Note that conference data in 2020 on the MAG system is not yet complete. See the Appendix for details.

By Region

Figure 1.1.12 shows that, similar to the trends in AI journal publication, East Asia & Pacific, Europe & Central Asia, and North America are the world's dominant sources for AI conference publications. Specifically, East Asia & Pacific took the lead starting in 2004, accounting for more than 27% in 2020. North America overtook Europe & Central Asia to claim second place in 2018, accounting for 20.1%, followed by 21.7% in 2020.

AI CONFERENCE PUBLICATIONS (% of WORLD TOTAL) by REGION, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

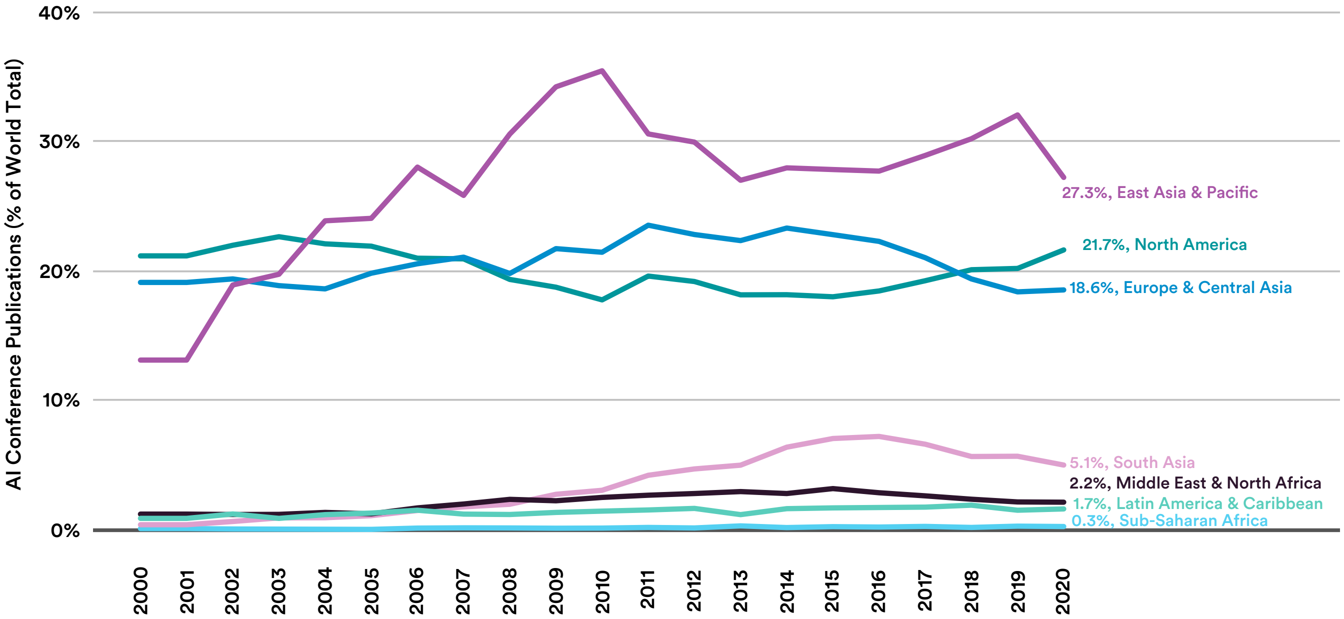

Figure 1.1.12

By Geographic Area

China overtook the United States in the share of AI conference publications in the world in 2019 (Figure 1.1.13). Its share has grown significantly since 2000. China's percentage of AI conference publications in 2019 is almost nine times higher than it was in 2000. The share of conference publications for the European Union peaked in 2011 and continues to decline.

AI CONFERENCE PUBLICATIONS (% of WORLD TOTAL) by GEOGRAPHIC AREA, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

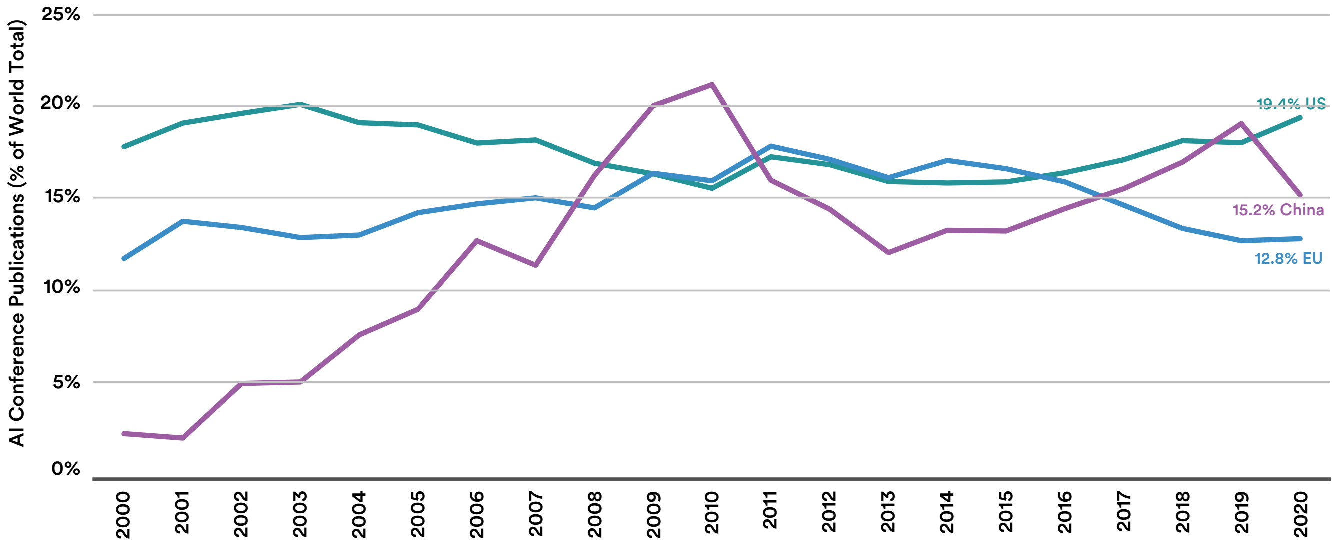

Figure 1.1.13

Citation

With respect to citations of AI conference publications, Figure 1.1.14 shows that the United States has held a dominant lead among the major powers over the past 21 years. The United States tops the list with 40.1% of overall citations in 2020, followed by China (11.8%) and the European Union (10.9%).

AI CONFERENCE CITATIONS (% of WORLD TOTAL) by GEOGRAPHIC AREA, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

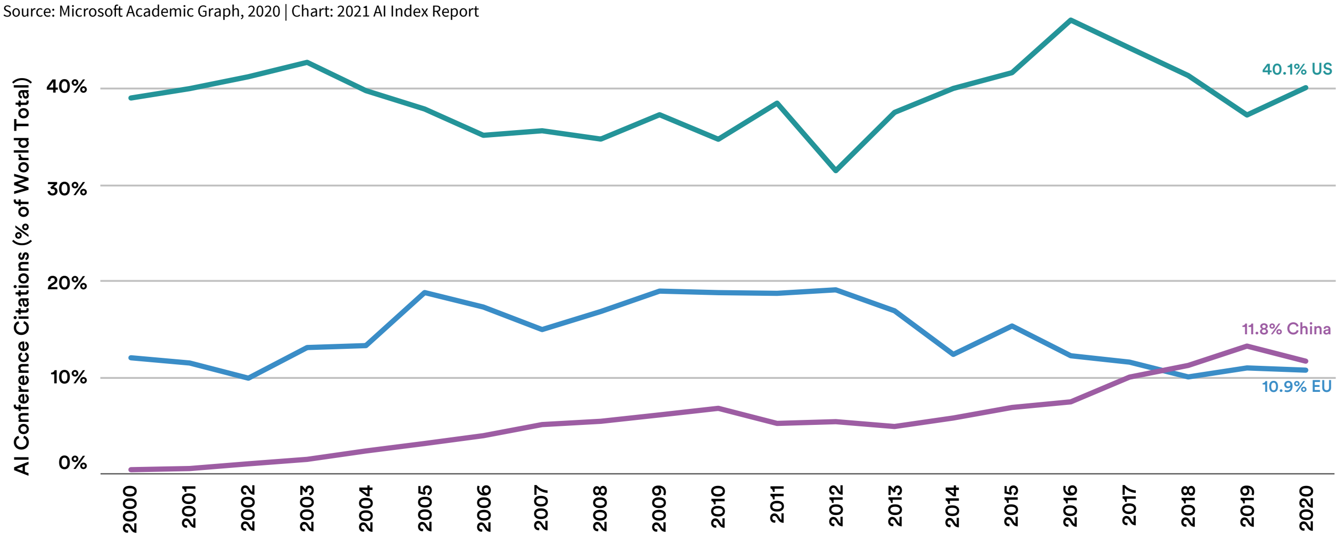

Figure 1.1.14

AI PATENTS

Overview

The total number of AI patents published in the world has been steadily increasing in the past two decades, growing from 21,806 in 2000 to more than 4.5 times that, or 101,876, in 2019 (Figure 1.1.15a). The share of AI patents published in the world exhibits a lesser increase, from around 2% in 2000 to 2.9% in 2020 (Figure 1.1.15b). The AI patent data is incomplete—only 8% of the dataset in 2020 includes a country or regional affiliation. There is reason to question the data on the share of AI patent publications by both region and geographic area, and it is therefore not included in the main report. See the Appendix for details.

NUMBER of AI PATENT PUBLICATIONS, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

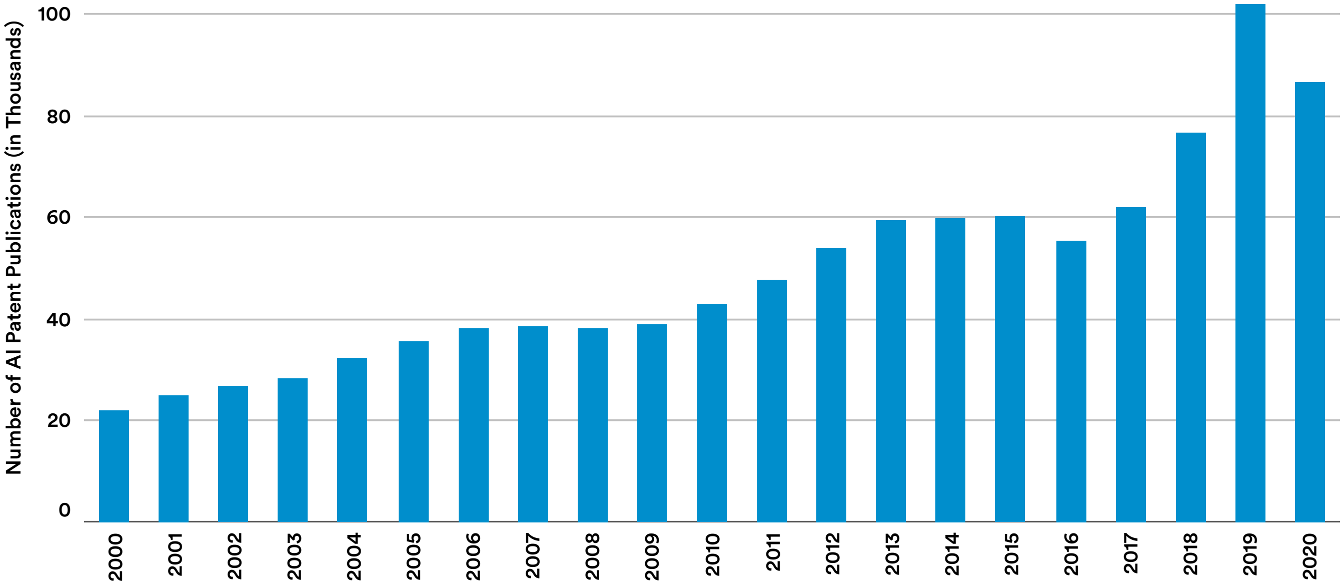

Figure 1.1.15a

AI PATENT PUBLICATIONS (% of ALL PATENT PUBLICATIONS), 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

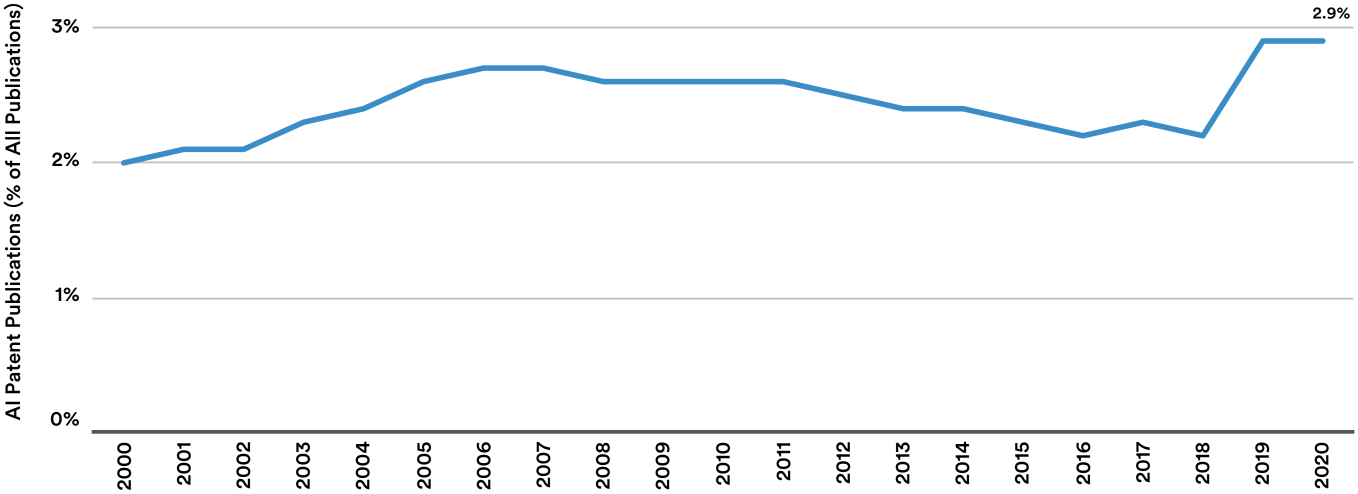

Figure 1.1.15b

ARXIV PUBLICATIONS

In addition to the traditional avenues for publishing academic papers (discussed above), AI researchers have embraced the practice of publishing their work (often pre–peer review) on arXiv, an online repository of electronic preprints. arXiv allows researchers to share their findings before submitting them to journals and conferences, which greatly accelerates the cycle of information discovery and dissemination. The number of AI-related publications in this section includes preprints on arXiv under cs.AI (artificial intelligence), cs.CL (computation and language), cs.CV (computer vision), cs.NE (neural and evolutionary computing), cs.RO (robotics), cs.LG (machine learning in computer science), and stat.ML (machine learning in statistics).

Overview

In just six years, the number of AI-related publications on arXiv grew more than sixfold, from 5,478 in 2015 to 34,736 in 2020 (Figure 1.1.16).

NUMBER of AI-RELATED PUBLICATIONS on ARXIV, 2015-20

Source: arXiv, 2020 | Chart: 2021 AI Index Report

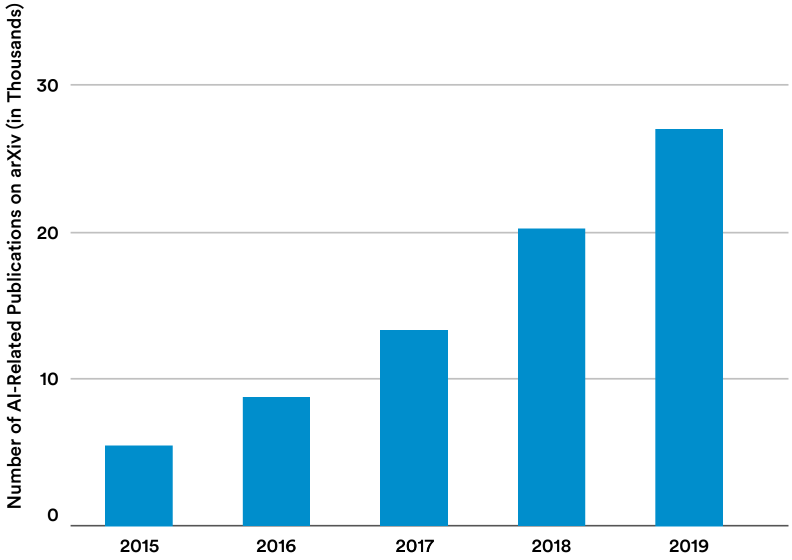

Figure 1.1.16

By Region

The analysis by region shows that while North America still holds the lead in the global share of arXiV AI-related publications, its share has been decreasing—from 41.6% in 2017 to 36.3% in 2020 (Figure 1.1.17). Meanwhile, the share of publications in East Asia & Pacific has grown steadily in the past five years—from 17.3% in 2015 to 26.5% in 2020.

ARXIV AI-RELATED PUBLICATIONS (% of WORLD TOTAL) by REGION, 2015-20

Source: arXiv, 2020 | Chart: 2021 AI Index Report

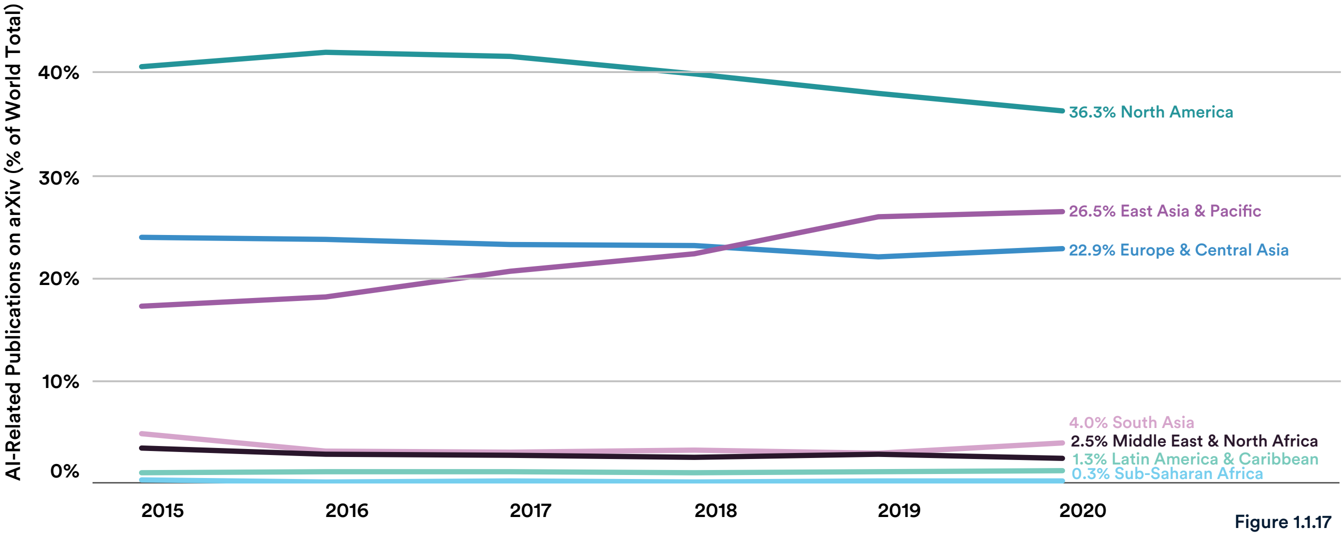

Figure 1.1.17

By Geographic Area

While the total number of AI-related publications on arXiv is increasing among the three major AI powers, China is catching up with the United States (Figure 1.1.18a and Figure 1.1.18b). The share of publication counts by the European Union, on the other hand, has remained largely unchanged.

NUMBER of AI-RELATED PUBLICATIONS on ARXIV by GEOGRAPHIC AREA, 2015-20

Source: arXiv, 2020 | Chart: 2021 AI Index Report

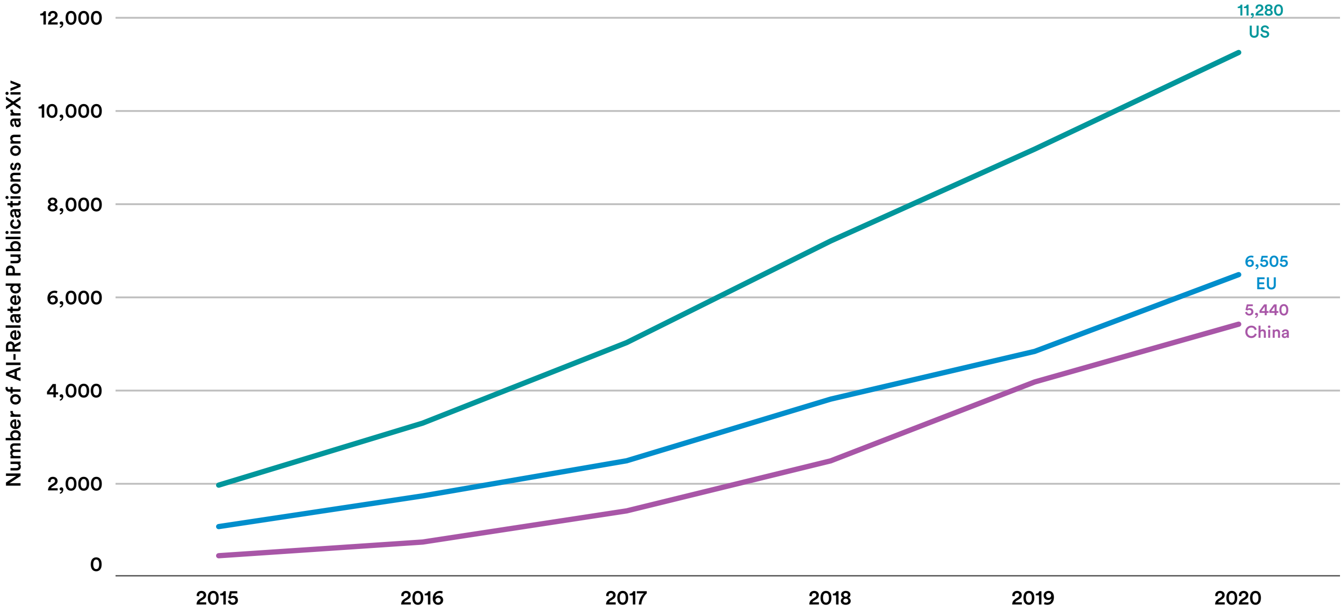

Figure 1.1.18a

ARXIV AI-RELATED PUBLICATIONS (% of WORLD TOTAL) by GEOGRAPHIC AREA, 2015-20

Source: arXiv, 2020 | Chart: 2021 AI Index Report

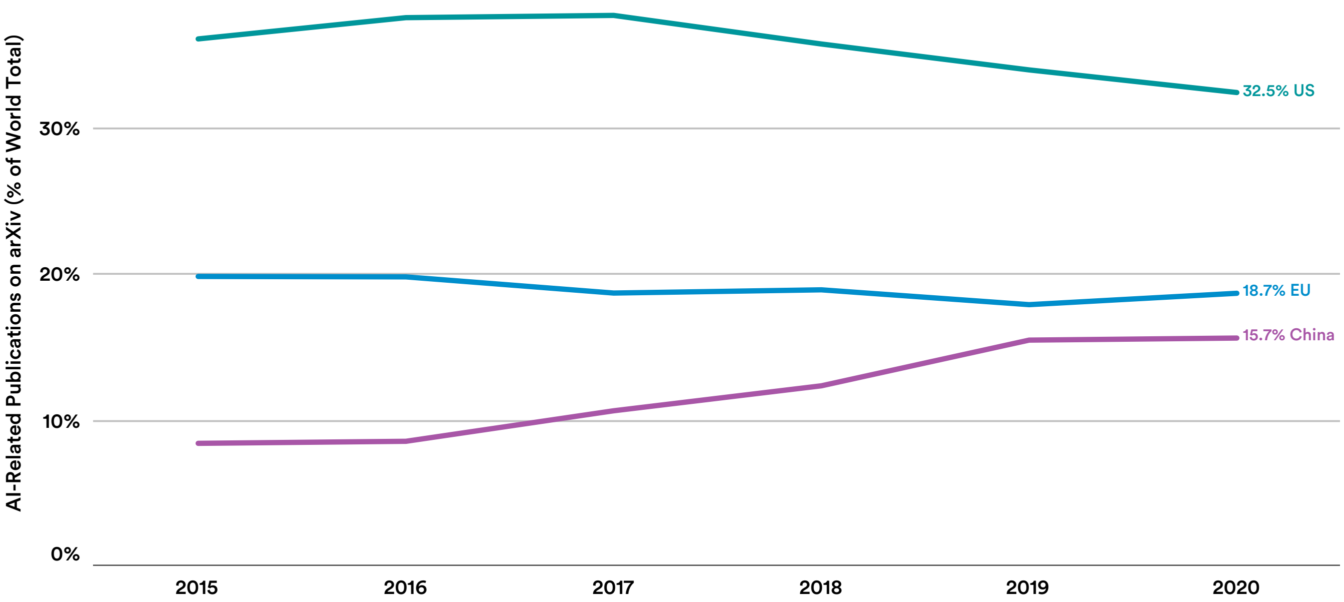

Figure 1.1.18b

By Field of Study

Among the six fields of study related to AI on arXiv, the number of publications in Robotics (cs.RO) and Machine Learning in computer science (cs.LG) have seen the fastest growth between 2015 and 2020, increasing by 11 times and 10 times respectively (Figure 1.1.19). In 2020, cs.LG and Computer Vision (cs.CV) lead in the overall number of publications, accounting for 32.0% and 31.7%, respectively, of all AI-related publications on arXiv. Between 2019 and 2020, the fastest-growing categories of the seven studied here were Computation and Language (cs.CL), by 35.4%, and cs.RO, by 35.8%.

Among the six fields of study related to AI on arXiv, the number of publications in Robotics (cs.RO) and Machine Learning in computer science (cs.LG) have seen the fastest growth between 2015 and 2020, increasing by 11 times and 10 times respectively.

NUMBER of AI-RELATED PUBLICATIONS on ARXIV by FIELD of STUDY 2015-20
Source: arXiv, 2020 | Chart: 2021 AI Index Report

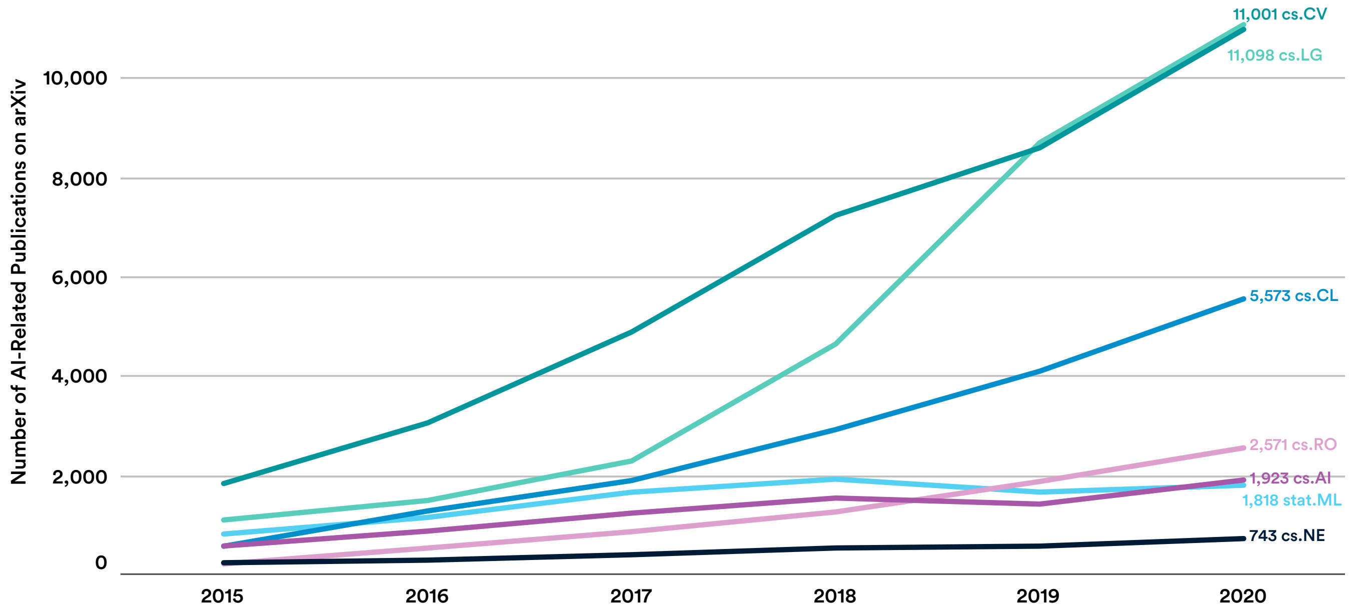

Figure 1.1.19

Deep Learning Papers on arXiv

With increased access to data and significant improvements in computing power, the field of deep learning (DL) is growing at breakneck speed. Researchers from Nesta used a topic modeling algorithm to identify the deep learning papers on arXiv by analyzing the abstract of arXiv papers under the Computer Science (CS) and Machine Learning in Statistics (state.ML) categories. Figure 1.1.20 suggests that in the last five years alone, the overall number of DL publications on arXiv grew almost sixfold.

NUMBER of DEEP LEARNING PUBLICATIONS on ARXIV, 2010-19
Source: arXiv/Nesta, 2020 | Chart: 2021 AI Index Report

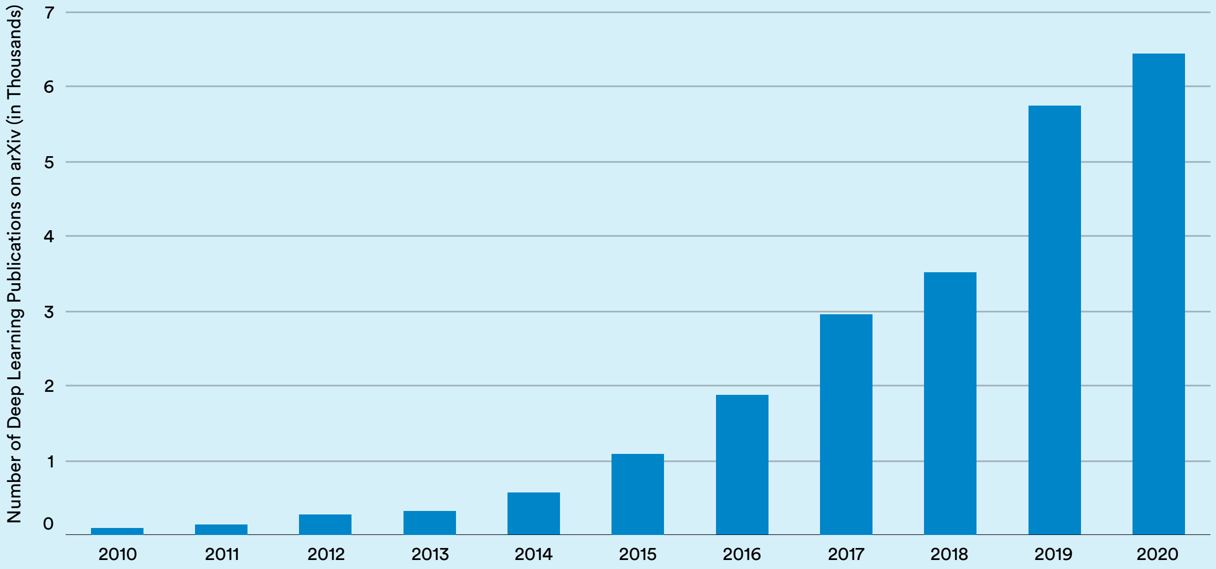

Figure 1.1.20

Conference attendance is an indication of broader industrial and academic interest in a scientific field. In the past 20 years, AI conferences have grown not only in size but also in number and prestige. This section presents data on the trends in attendance at and submissions to major AI conferences.

1.2 CONFERENCES

CONFERENCE ATTENDANCE

Last year saw a significant increase in participation levels at AI conferences, as most were offered through a virtual format. Only the 34th Association for the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence was held in person in February 2020. Conference organizers report that a virtual format allows for higher attendance of researchers from all over the world, though exact attendance numbers are difficult to measure.

Due to the atypical nature of 2020 conference attendance data, the 11 major AI conferences in 2019 have been split into two categories based on 2019 attendance data: large AI conferences with over 3,000 attendees and small AI conferences with fewer than 3,000 attendees. Figure 1.2.1 shows that in 2020, the total number of attendees across nine conferences almost doubled.[5] In particular, the International Conference on Intelligent Robots and Systems (IROS) extended the virtual conference to allow users to watch events for up to three months, which explains the high attendance count. Because the International Joint Conference on Artificial Intelligence (IJCAI) was held in 2019 and January 2021—but not in 2020—it does not appear on the charts.

Conference organizers report that a virtual format allows for higher attendance of researchers from all over the world, though exact attendance numbers are difficult to measure.

[5] For the AAMAS conference, the attendance in 2020 is based on the number of users on site reported by the platform that recorded the talks and managed the online conference; For the KR conference, the attendance in 2020 is based on the number of registrations; For the ICPAS conference, the attendance of 450 in 2020 is an estimate as some participants may have used anonymous Zoom accounts.

ATTENDANCE at LARGE AI CONFERENCES, 2010-20

Source: Conference Data | Chart: 2021 AI Index Report

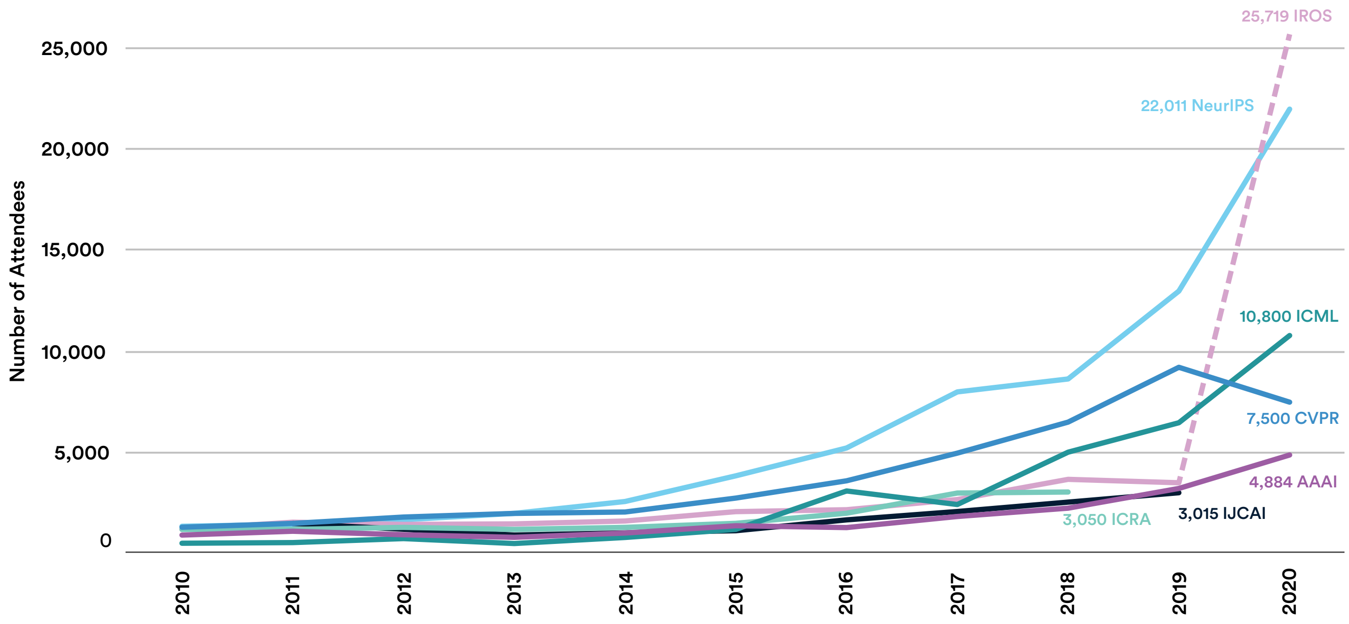

Figure 1.2.1

ATTENDANCE at SMALL AI CONFERENCES, 2010-20

Source: Conference Data | Chart: 2021 AI Index Report

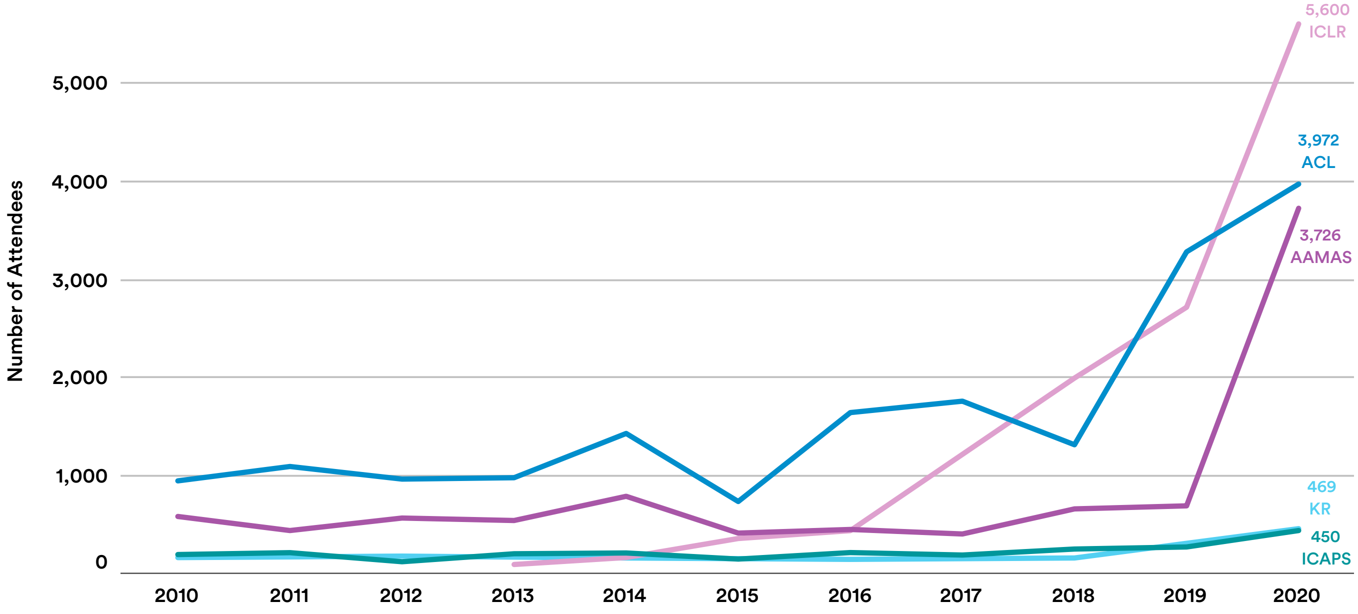

Figure 1.2.2

Corporate Representation at AI Research Conferences

Researchers from Virginia Tech and Ivey Business School, Western University found that large technology firms have increased participation in major AI conferences. In their paper, titled "The De-Democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research," the ressearchers use the share of papers affiliated with firms over time at AI conferences to illustrate the increased presence of firms in AI research. They argue that the unequal distribution of compute power in academia, which they refer to as the "compute divide," is adding to the inequality in the era of deep learning. Big tech firms tend to have more resources to design AI products, but they also tend to be less diverse than less elite or smaller institutions. This raises concerns about bias and fairness within AI. All 10 major AI conferences displayed in Figure 1.2.3 show an upward trend in corporate representation, which further extends the compute divide.

SHARE of FORTUNE GLOBAL 500 TECH-AFFILIATED PAPERS
Source: Ahmed & Wahed, 2020 | Chart: 2021 AI Index Report

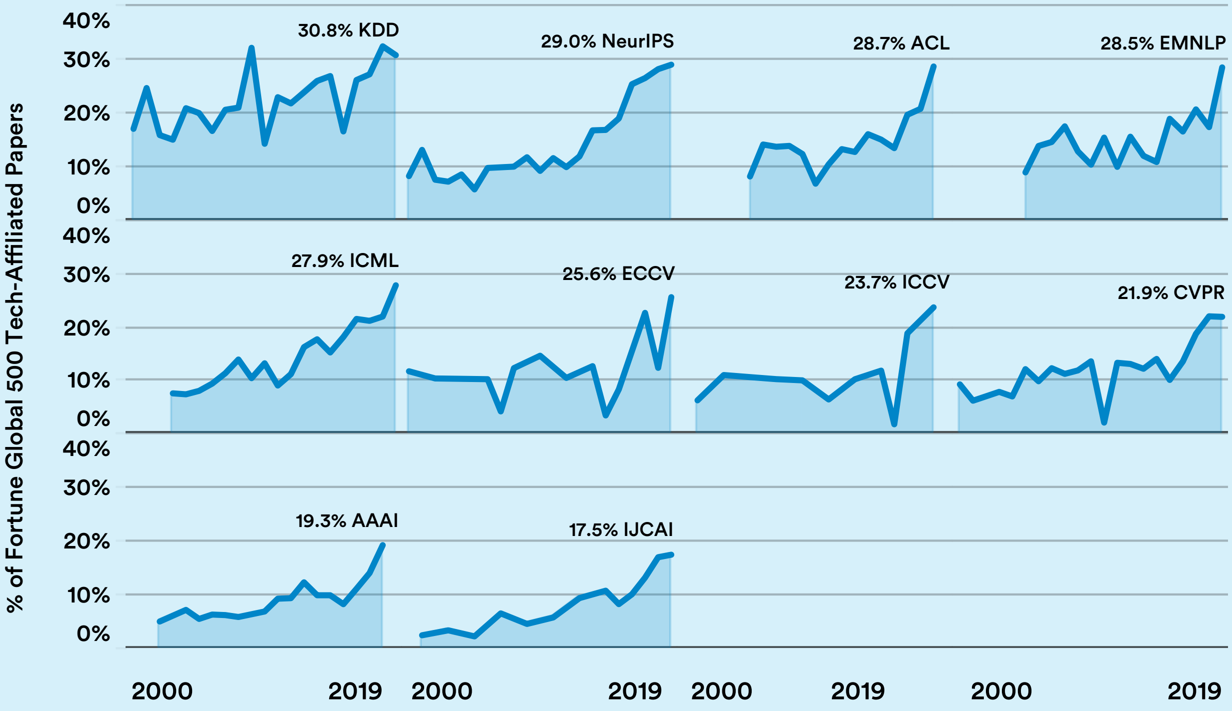

Figure 1.2.3

A software library is a collection of computer code that is used to create applications and products. Popular AI-specific software libraries—such as TensorFlow and PyTorch—help developers create their AI solutions quickly and efficiently. This section analyzes the popularity of software libraries through GitHub data.

1.3 AI OPEN-SOURCE SOFTWARE LIBRARIES

GITHUB STARS

GitHub is a code hosting platform that AI researchers and developers frequently use to upload, comment on, and download software. GitHub users can “star” a project to save it in their list, thereby expressing their interests and likes—similar to the “like’’ function on Twitter and other social media platforms. As AI researchers upload packages on GitHub that mention the use of an open-source library, the “star” function on GitHub can be used to measure the popularity of various AI programming open-source libraries.

Figure 1.3.1 suggests that TensorFlow (developed by Google and publicly released in 2017) is the most popular AI software library. The second most popular library in 2020 is Keras (also developed by Google and built on top of TensorFlow 2.0). Excluding TensorFlow, Figure 1.3.2 shows that PyTorch (created by Facebook) is another library that is becoming increasingly popular.

TensorFlow (developed by Google and publicly released in 2017) is the most popular AI software library. The second most popular library in 2020 is Keras (also developed by Google and built on top of TensorFlow 2.0).

NUMBER of GITHUB STARS by AI LIBRARY, 2014-20

Source: GitHub, 2020 | Chart: 2021 AI Index Report

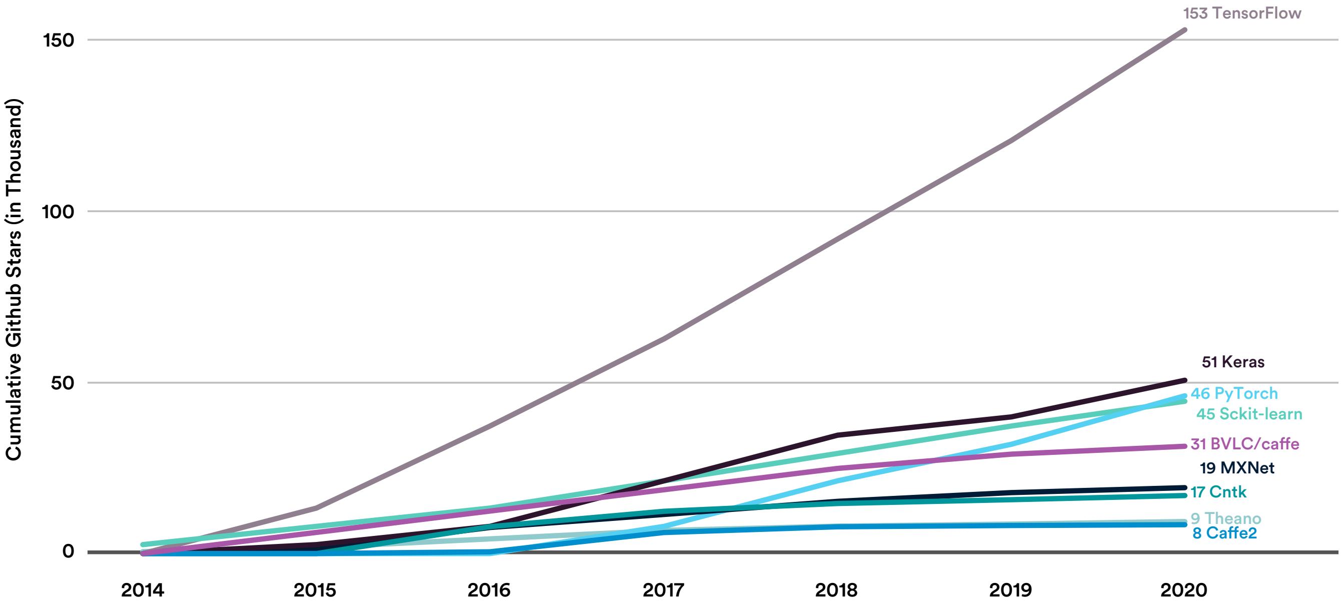

Figure 1.3.1

NUMBER of GITHUB STARS by AI LIBRARY (excluding TENSORFLOW), 2014-20

Source: GitHub, 2020 | Chart: 2021 AI Index Report

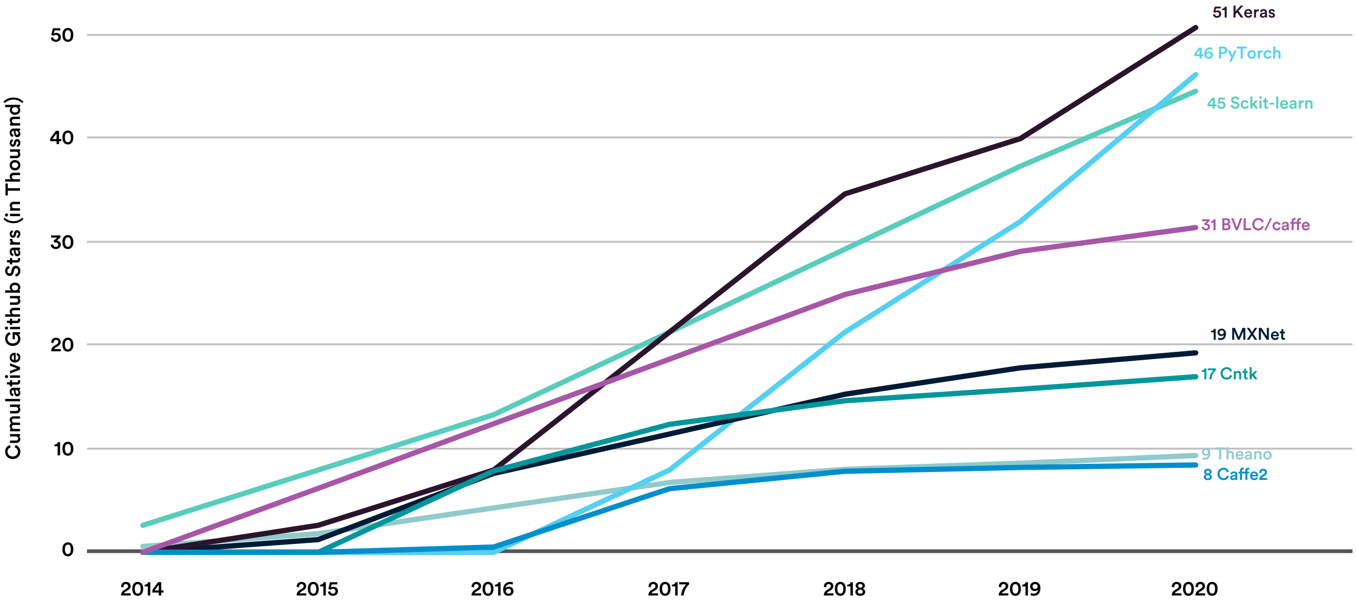

Figure 1.3.2

APPENDIX

ELSEVIER

Prepared by Jörg Hellwig and Thomas A. Collins

Source

Elsevier's Scopus database of scholarly publications has indexed more than 81 million peer-reviewed documents. This data was compiled by Elsevier.

Methodology

Scopus tags its papers with keywords, publication dates, country affiliations, and other bibliographic information.

The Elsevier AI Classifier leveraged the following features extracted from the Scopus records that were returned as a result of querying against the provided approximately 800 AI search terms. Each record fed into the feature creation also maintained a list of each search term that hit for that particular record:

- hasAbs: Boolean value whether or not the record had an abstract text section in the record (e.g., some records are only title and optional keywords)
- coreCnt: number of core-scored search terms present for the record
- mediumCnt: number of medium-scored search terms present for the record
- lowCnt: number of low-scored search terms present for the record
- totalCnt: total number of search terms present for the record
- pcntCore: coreCnt/totalCnt
- pcntMedium: mediumCnt/totalCnt
- pcntLow: lowCnt/totalCnt
- totalWeight = 5*coreCnt + 3*mediumCnt + 1*lowCnt
- normWeight = if (has Abs) { totalWeight / (title.length + abstract.length) } else
- { totalWeight/title.length}
- hasASJC: Boolean value: does the record have an associated ASJC list?
- isAiASJC: does ASJC list contain 1702?
- isCompSciASJC does ASJC list contain a 17XX ASJC code ("1700," "1701," "1702," "1703," "1704," "1705," "1706," "1707," "1708," "1709," "1710," "1711," "1712")
- isCompSubj: does the Scopus record have a ComputerScience subject code associated with it? This should track 1:1 to isCompSciASJC. Scopus has 27 major subject areas of which one is Computer Science. The feature checks, if the publication is within Computer Science or not. This is no exclusion.pcntCompSciASJC: percentage of ASJC codes for record that are from the CompSci ASJC code list

Details on Elsevier's dataset defining AI, country affiliations, and AI subcategories can be found in the 2018 AI Index Report Appendix.

Nuance

- The Scopus system is retroactively updated. As a result, the number of papers for a given query may increase over time.
- Members of the Elsevier team commented that data on papers published after 1995 would be the most reliable. The raw data has 1996 as the starting year for Scopus data.

Nuances specific to AI publications by region

- Papers are counted utilizing whole counting rather than fractional counting. Papers assigned to multiple countries (or regions) due to collaborations are counted toward each country (or region). This explains why top-line numbers in a given year may not match individual country numbers. For example, a paper assigned to Germany, France, and the United States will appear on each country's count, but only once for Europe (plus once for the U.S.) as well as being counted only at the global level.

- "Other" includes all other countries that have published one or more AI papers on Scopus.

Nuances specific to publications by topic

- The 2017 AI Index Report showed only AI papers within the CS category. In the 2018 and 2019 reports, all papers tagged as AI were included, regardless of whether they fell into the larger CS category.
- Scopus has a subject category called AI, which is a subset of CS, but this is relevant only for a subject-category approach to defining AI papers. The methodology used for the report includes all papers, since increasingly not all AI papers fall under CS.

Nuances specific to methodology

- The entire data collection process was done by Elsevier internally. The AI Index was not involved in the keyword selection process or the counting of relevant papers.
- The boundaries of AI are difficult to establish, in part because of the rapidly increasing applications in many fields, such as speech recognition, computer vision, robotics, cybersecurity, bioinformatics, and healthcare. But limits are also difficult to define because of AI's methodological dependency on many areas, such as logic, probability and statistics, optimization, photogrammetry, neuroscience, and game theory—to name just a few. Given the community's interest in AI bibliometrics, it would be valuable if groups producing these studies strived for a level of transparency in their methods, which would support the reproducibility of results, particularly on different underlying bibliographic databases.

AI Training Set

A training set of approximately 1,500 publications defines the AI field. The set is only the EID (the Scopus identifier of the underlying publications). Publications can be searched and downloaded either from Scopus directly or via the API.The training set is a set of publications randomly selected from the initial seven mio publications. After running the algorithm we verify the results of the training set with the gold set (expert hand-checked publications which are definitely AI).

MICROSOFT ACADEMIC GRAPH: METHODOLOGY

Prepared by Zhihong Shen, Boya Xie, Chiyuan Huang, Chieh-Han Wu, and Kuansan Wang

Source

The Microsoft Academic Graph[1] is a heterogeneous graph containing scientific publication records and citation relationships between those publications, as well as authors, institutions, journals, conferences, and fields of study. This graph is used to power experiences in Bing, Cortana, Word, and Microsoft Academic. The graph is currently being updated on a weekly basis. Learn more about MAG here.

Methodology

MAG Data Attribution: Each paper is counted exactly once. When a paper has multiple authors or regions, the credit is equally distributed to the unique regions. For example, if a paper has two authors from the United States, one from China, and one from the United Kingdom, then the United States, China, and the United Kingdom each get one-third credit.

Metrics: Total number of published papers (journal papers, conference papers, patents, repository[2]); total number of citations of published papers.

Definition: The citation and reference count represents the number of respective metrics for AI papers collected from all papers. For example, in "OutAiPaperCitationCountryPairByYearConf.csv," a row stating "China, United States, 2016, 14955" means that China's conference AI papers published in 2016 received 14,955 citations from (all) U.S. papers indexed by MAG.

Curating the MAG Dataset and References: Generally speaking, the robots sit on top of a Bing crawler to read everything from the web and have access to the entire web index. As a result, MAG is able to program the robots to conduct more web searches than a typical human can complete. This helps disambiguate entities with the same names. For example, for authors, MAG gets to additionally use all the CVs and institutional homepages on the web as signals to recognize and verify claims[3]. MAG has found this approach to be superior to the results of the best of the KDD Cup 2013 competition, which uses only data from within all publication records and Open Researcher and Contributor Identifiers (ORCIDs).

Notes About the MAG Data

Conference Papers: After the contents and data sources were scrutinized, it was determined that some of the 2020 conference papers were not properly tagged with their conference venue. Many conference papers in the MAG system are under arXiv papers, but due to issues arising from some data sources (including delays in DBLP and web form changes on the ACM website), they were possibly omitted as 2020 conference papers (ICML-PKDD, IROS, etc.). However, the top AI conferences (selected not in terms of publication count, but rather considering both publication and citation count as well as community prestige) are complete. In 2020, the top 20 conferences presented 103,000 papers, which is 13.7% of all AI conference papers, and they received 7.15 million citations collectively, contributing 47% of all citations received for all AI conference papers. The number of 2020 conference publications is slightly lower than in 2019. Data is known to be missing for ICCV and NAACL. About 100 Autonomous Agents and Multiagent Systems (AAMAS) conference papers are erroneously attributed to an eponymous journal.

Unknown Countries for Journals and Conferences: For the past 20 to 30 years, 30% of journal and conference affiliation data lacks affiliation by country or region, due to errors in paper format, data source, and PDF parsing, among others.

[1] See "A Review of Microsoft Academic Services for Science of Science Studies" and "Microsoft Academic Graph: When Experts Are Not Enough" for more details.

[2] Repository as a publication type in MAG refers to both preprints and postprints. In the AI domain, it predominantly comes from arXiv. See "Is Preprint the Future of Science? A Thirty Year Journey of Online Preprint Services" for details.

[3] See "Machine Verification for Paper and Author Claims" and "How Microsoft Academic Uses Knowledge to Address the Problem of Conflation/Disambiguation" for details.

MICROSOFT ACADEMIC GRAPH: PATENT DATA CHALLENGE

As mentioned in the report, the patent data—especially the affiliation information—is incomplete in the MAG database. The reason for the low coverage is twofold. First, applications published by the patent offices often identify the inventors by their residencies not affiliations. While patent applications often have the information about the "assignees" of a patent, they do not necessarily mean the underlying inventions originate from the assignee institutions. Therefore, detected affiliations may be inaccurate. In case a patent discloses the scholarly publications underlying the invention, MAG can infer inventors' affiliations through the scholarly publications.

Second, to maximize intellectual property protection around the globe, institutions typically file multiple patent applications on the same invention under various jurisdictions. These multiple filings, while appear very different because the titles and inventor names are often translated into local languages, are in fact the result of a single invention. Raw patent counts therefore inflate the inventions in their respective domains. To remediate this issue, MAG uses the patent family ID feature to combine all filings with the original filing, which allows the database to count filings all around the world of the same origin only once.[4] Conflating the multiple patent applications of the same invention is not perfect, and over-conflations of patents are more noticeable in MAG than scholarly articles.

These challenges raise questions about the reliability of data on the share of AI patent publications by both region and geographic area. Those charts are included below.

By Region

AI PATENT PUBLICATIONS (% of WORLD TOTAL) by REGION, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

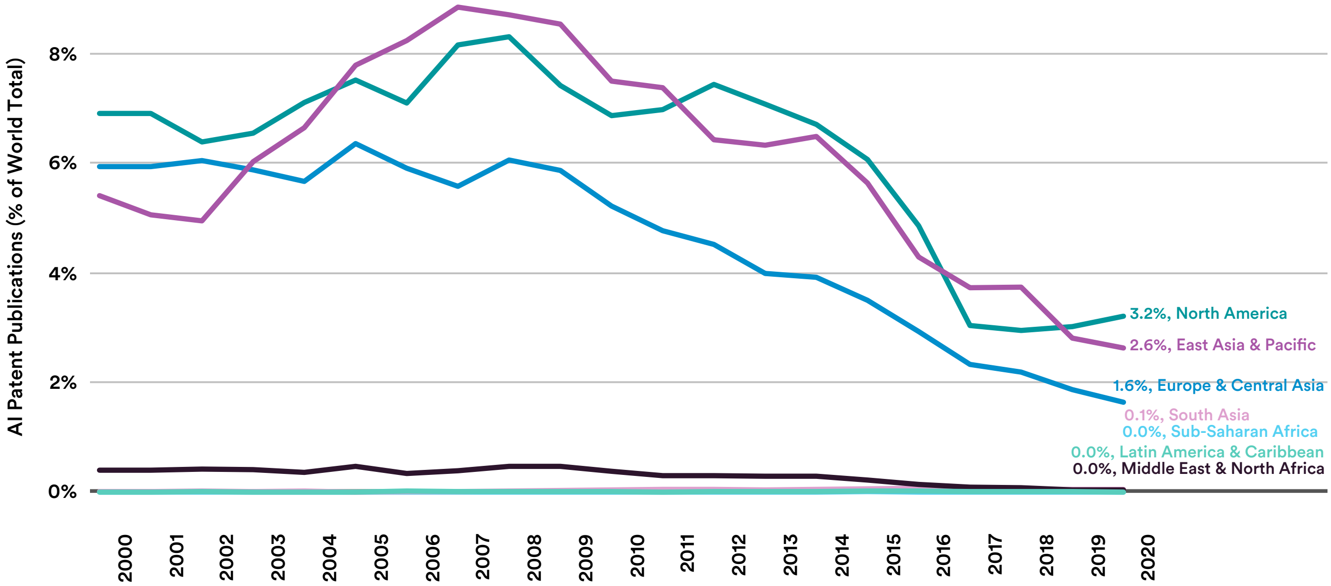

Figure 1.4.1

[4] Read "Sharpening Insights into the Innovation Landscape with a New Approach to Patents" for more details.

By Geographic Area

AI PATENT PUBLICATIONS (% of WORLD TOTAL) by GEOGRAPHIC AREA, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

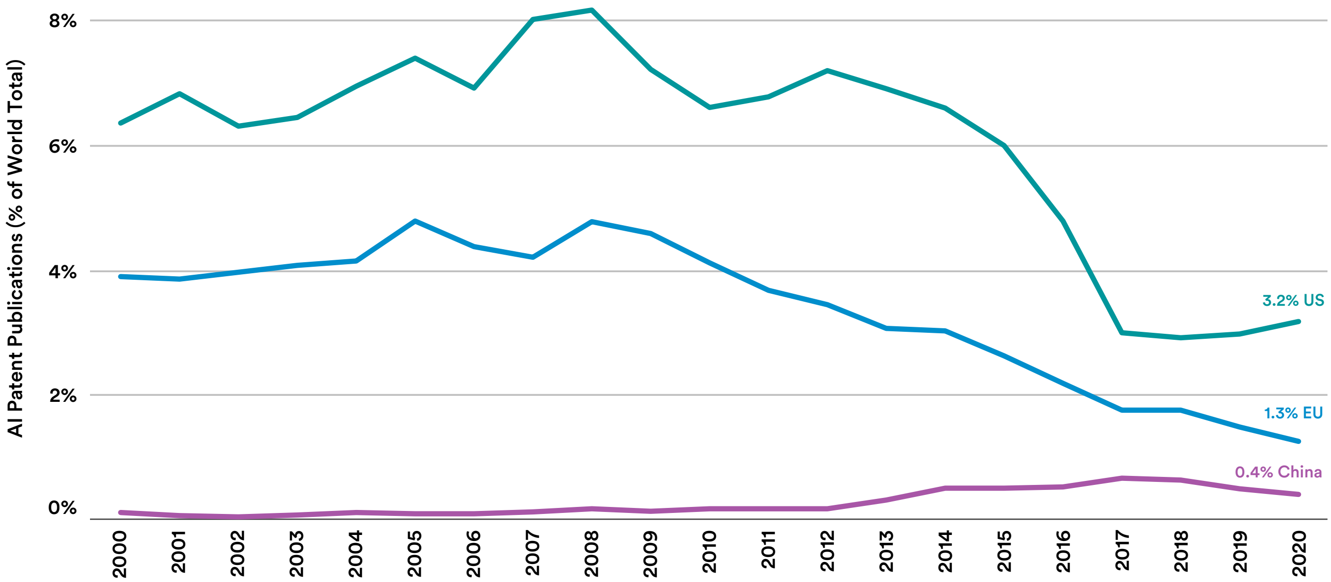

Figure 1.4.2

Citation

AI PATENT CITATIONS (% of WORLD TOTAL) by GEOGRAPHIC AREA, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

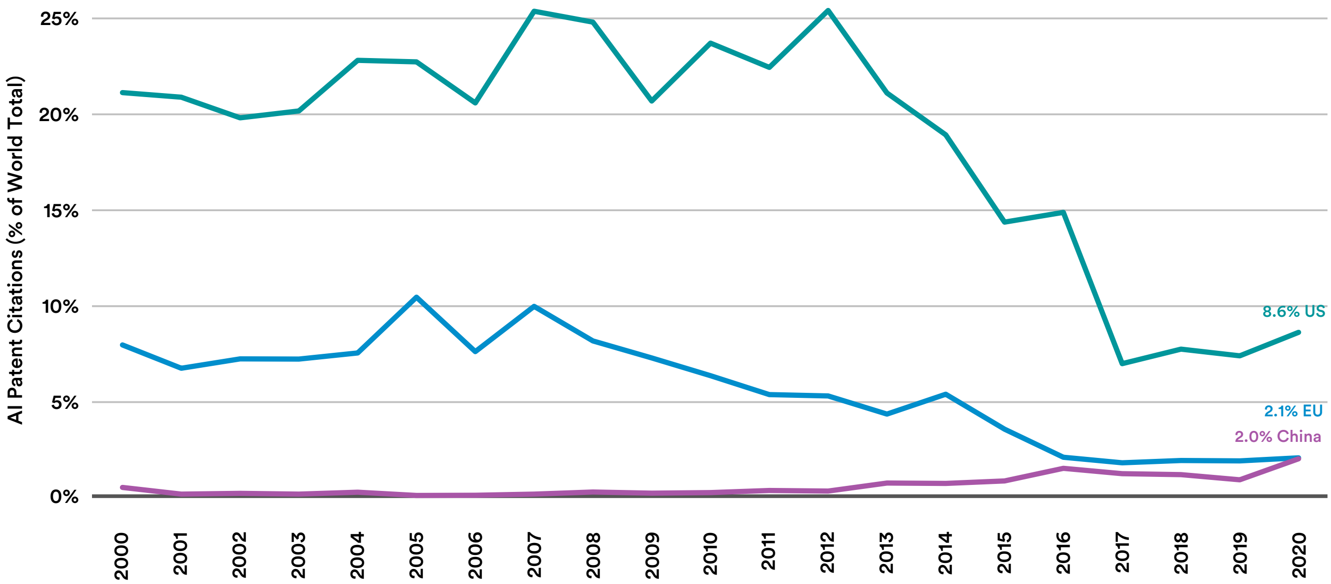

Figure 1.4.3

MICROSOFT ACADEMIC GRAPH: MEASUREMENT CHALLENGES AND ALTERNATIVE DEFINITION OF AI

As the AI Index team discussed in the paper "Measurement in AI Policy: Opportunities and Challenges," choosing how to define AI and correctly capture relevant bibliometric data remain challenging. Data in the main report is based on a restricted definition of AI, adopted by MAG, that aligns with what has been used in previous AI Index reports. One consequence is that such a definition excludes many AI publications from venues considered to be core AI venues. For example, only 25% of conference publications in the 2020 AAAI conference are included in the original conference dataset.

To spur discussion on this important topic, this section presents the MAG data with an alternative definition of AI used by the Organisation for Economic Co-operation and Development (OECD). OECD defines AI publications as papers in the MAG database tagged with a field of study that is categorized in either the "artificial intelligence" or the "machine learning" field of study as well as their subtopics in the MAG taxonomy.[5] This is a more liberal definition than the one used by MAG, which considers only those publications tagged with "artificial intelligence" as AI publications. For example, an application paper in biology that uses ML techniques will be counted as an AI publication under the OECD definition, but not under the MAG definition unless the paper is specifically tagged in the AI category.

Charts corresponding to those in the main text but using the OECD definition are presented below. The overall trends are very similar.

AI Journal Publications (OECD Definition)

OECD DEFINITION: NUMBER of AI JOURNAL PUBLICATIONS, 2000-20
Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

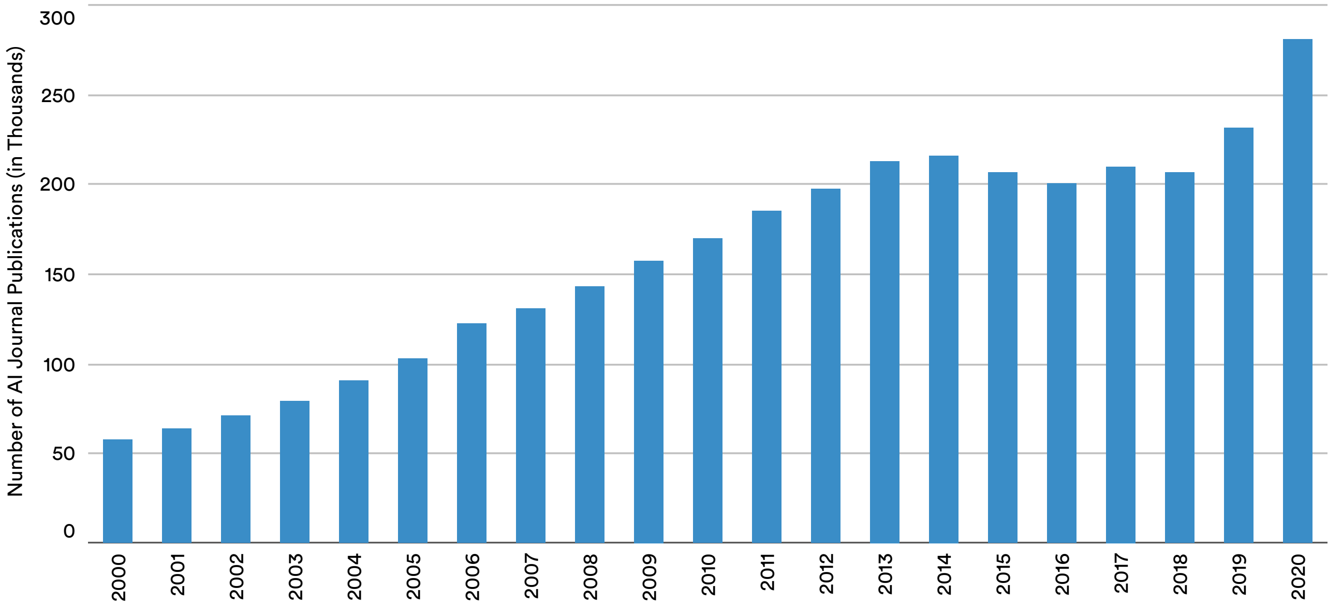

Figure 1.5.1a

[5] Read the OECD.AI Policy Observatory MAG methodological note for more details on the MAG-OECD definition of AI and "A Web-scale System for Scientific Knowledge Exploration" on the MAG Taxonomy.

OECD DEFINITION: AI JOURNAL PUBLICATIONS (% of ALL JOURNAL PUBLICATIONS), 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

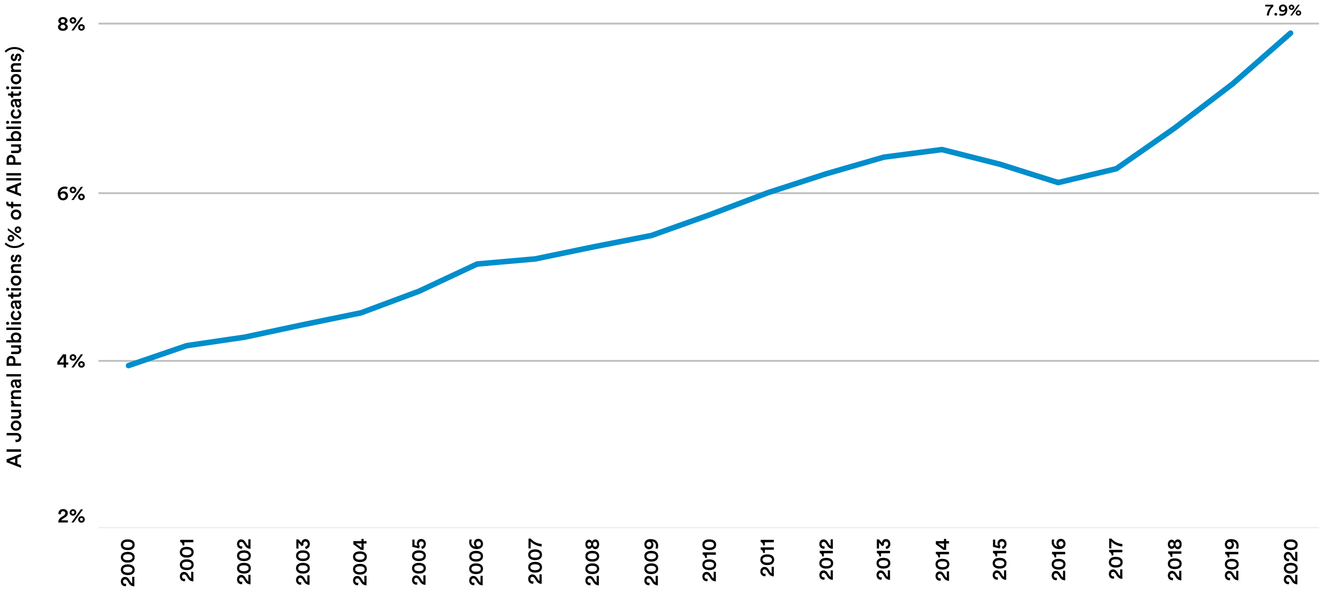

Figure 1.5.1b

OECD DEFINITION: AI JOURNAL PUBLICATION (% of WORLD TOTAL) by REGION, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

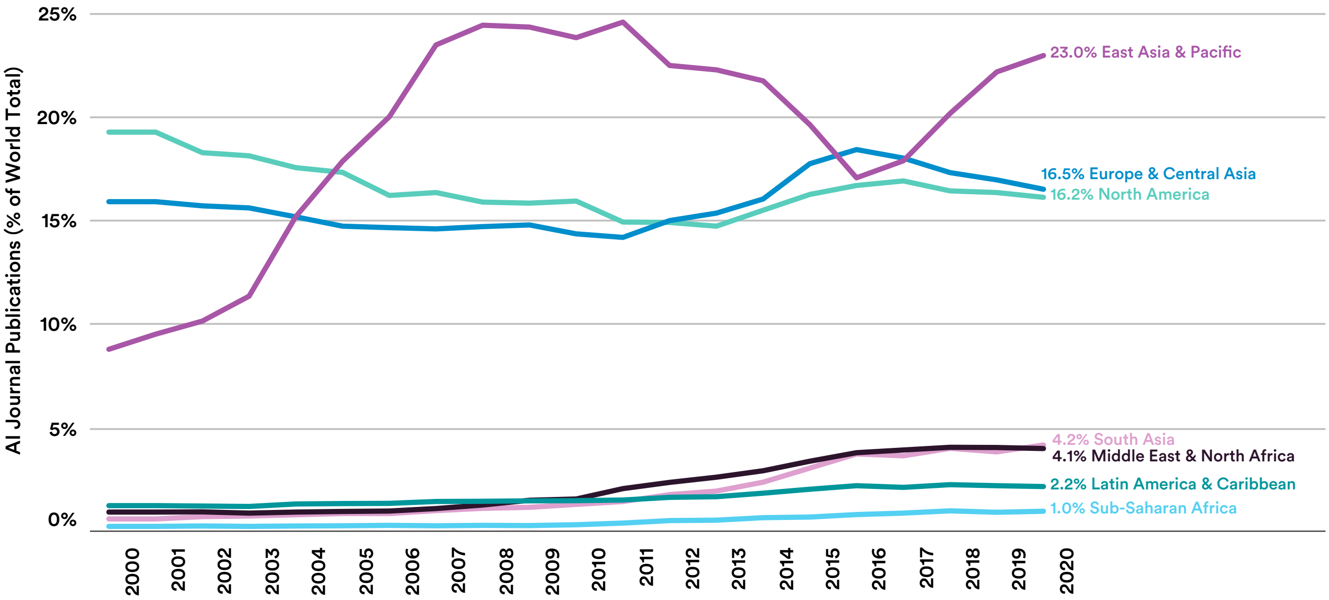

Figure 1.5.2

OECD DEFINITION: AI JOURNAL PUBLICATION (% of WORLD TOTAL) by GEOGRAPHIC AREA, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

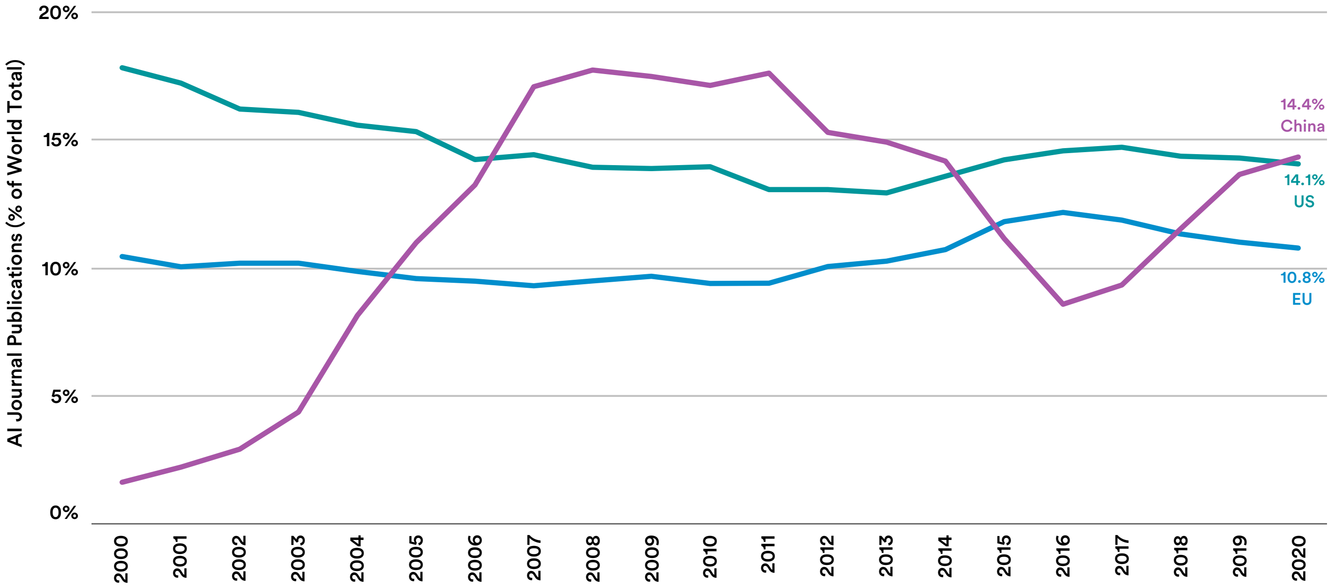

Figure 1.5.3

OECD DEFINITION: AI JOURNAL CITATIONS (% of WORLD TOTAL) by GEOGRAPHIC AREA, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

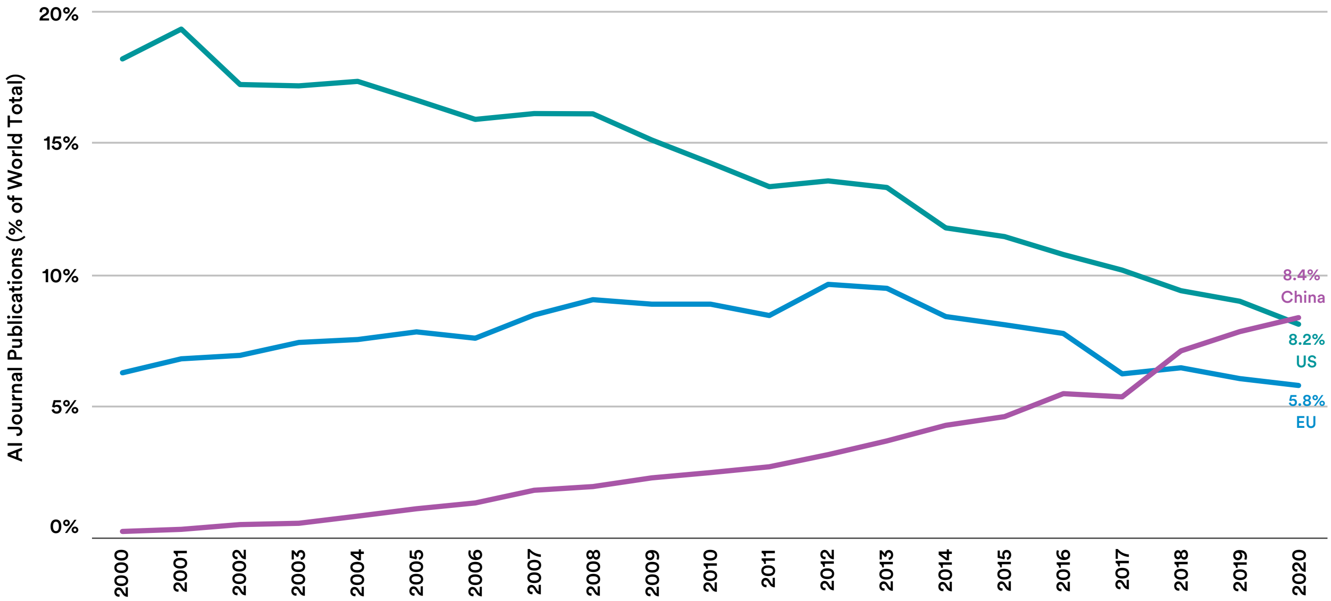

Figure 1.5.4

AI Conference Publications (OECD Definition)

OECD DEFINITION: NUMBER of AI CONFERENCE PUBLICATIONS, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

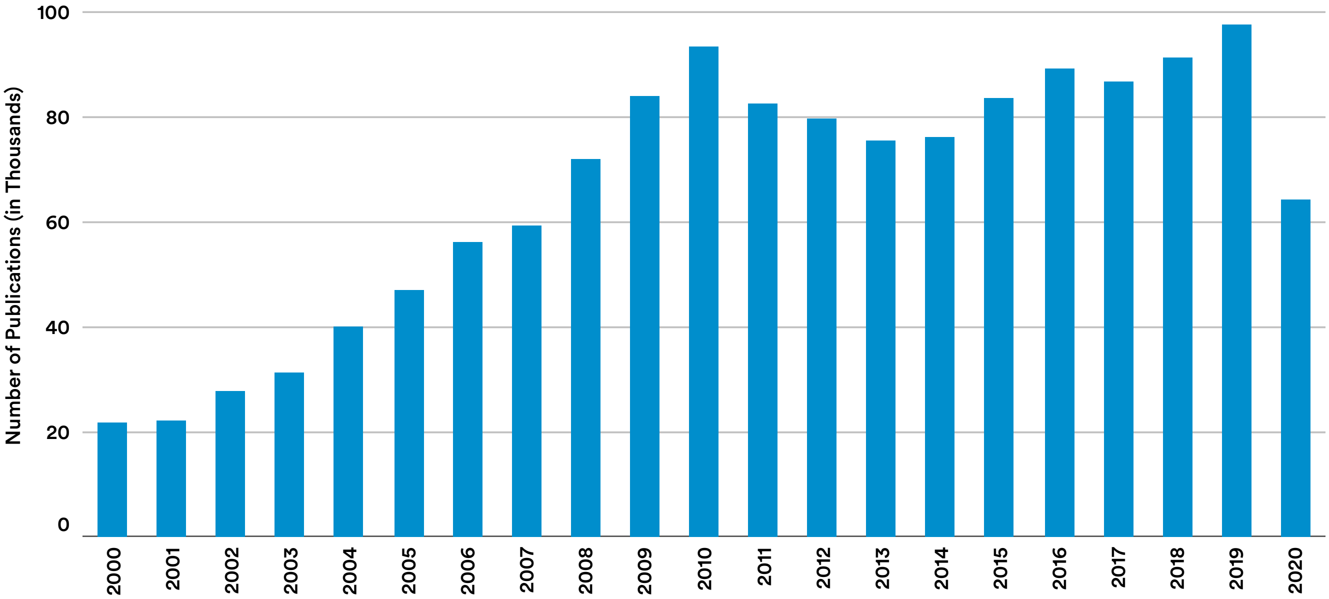

Figure 1.5.5a

OECD DEFINITION: AI CONFERENCE PUBLICATIONS (% of ALL CONFERENCE PUBLICATIONS), 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

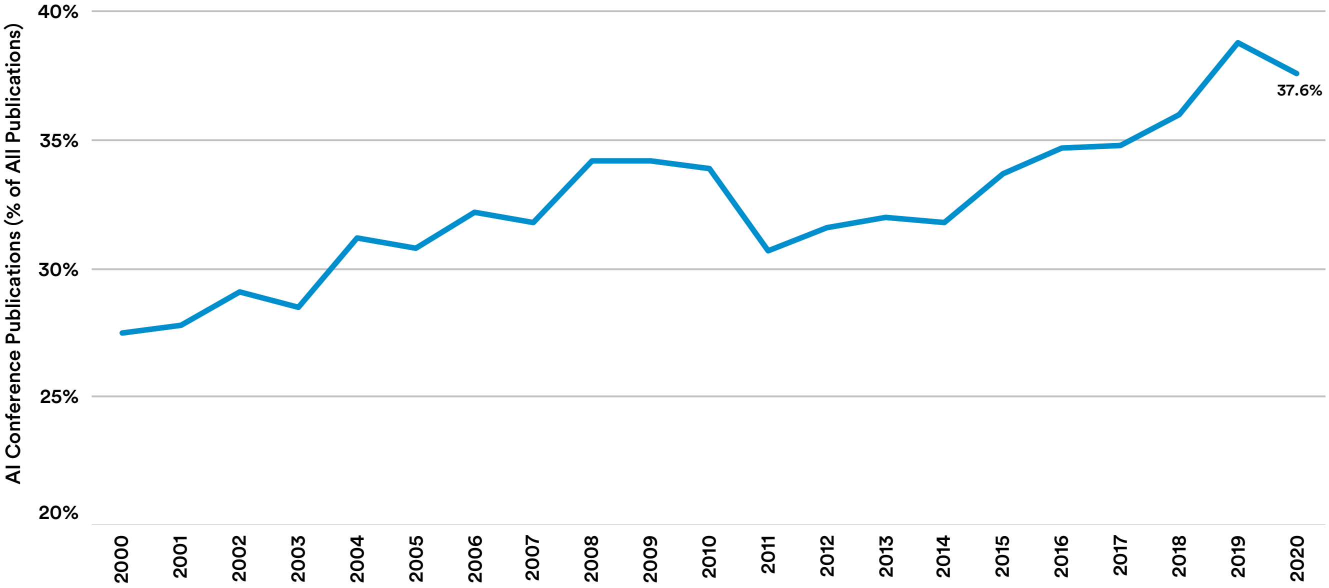

Figure 1.5.5b

OECD DEFINITION: AI CONFERENCE PUBLICATION (% of WORLD TOTAL) by REGION, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

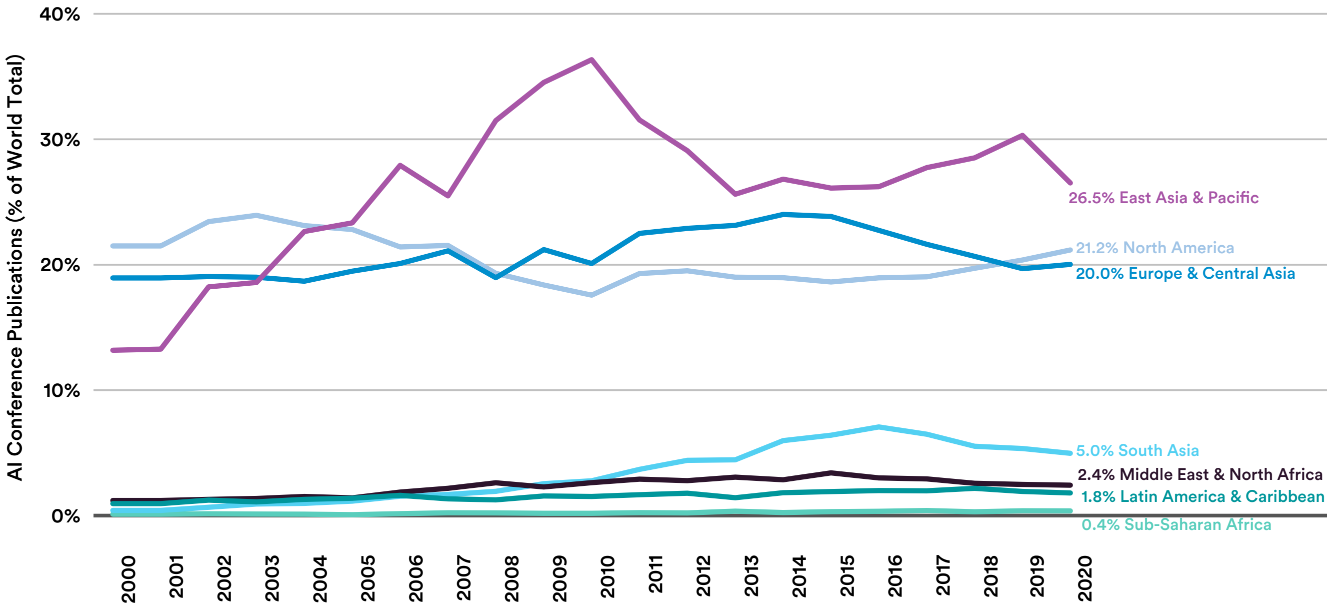

Figure 1.5.6

OECD DEFINITION: AI CONFERENCE PUBLICATION (% of WORLD TOTAL) by GEOGRAPHIC AREA, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

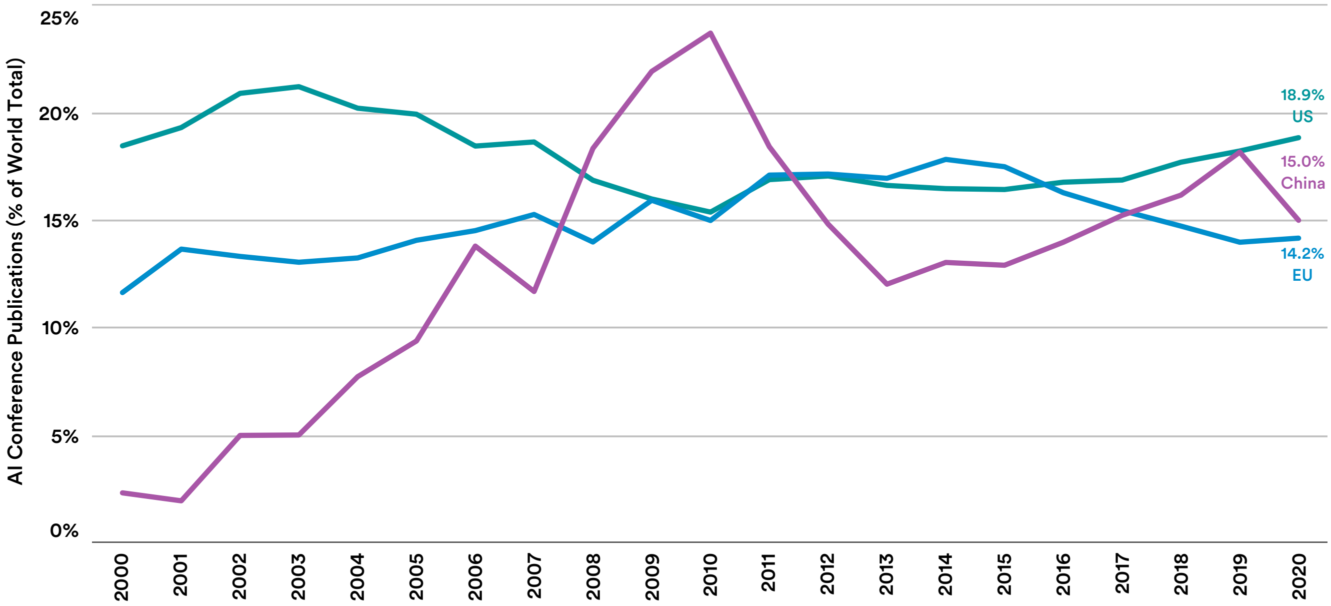

Figure 1.5.7

OECD DEFINITION: AI CONFERENCE CITATION (% of WORLD TOTAL) by GEOGRAPHIC AREA, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

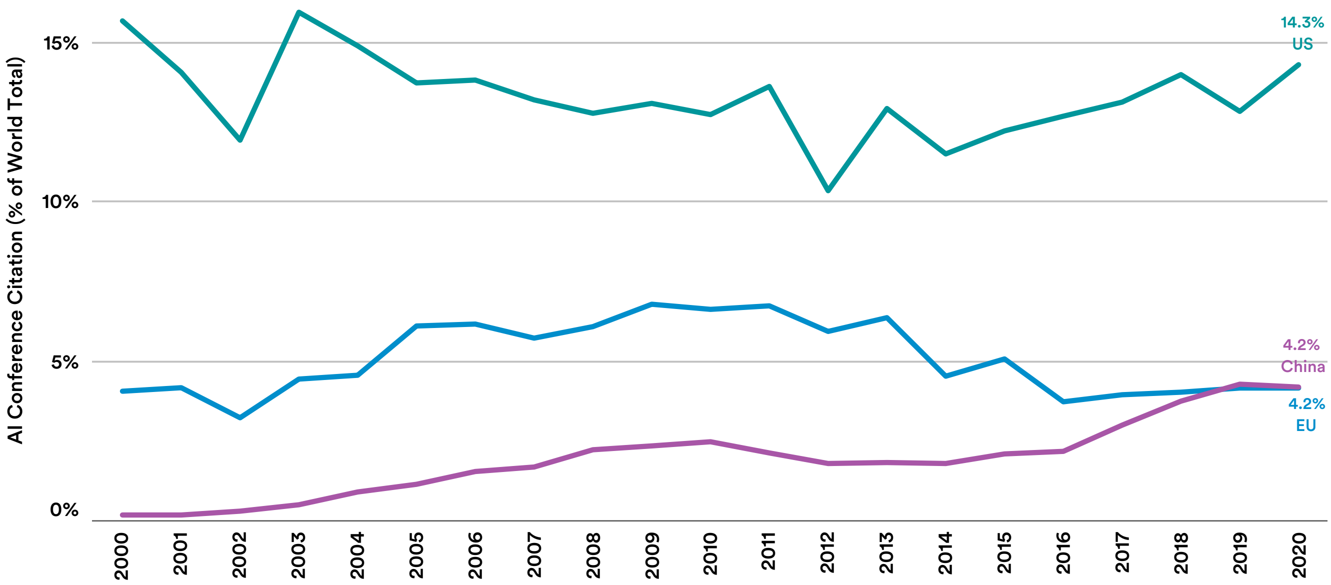

Figure 1.5.8

AI Patent Publications (OECD Definition)

OECD DEFINITION: NUMBER of AI PATENT PUBLICATIONS, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

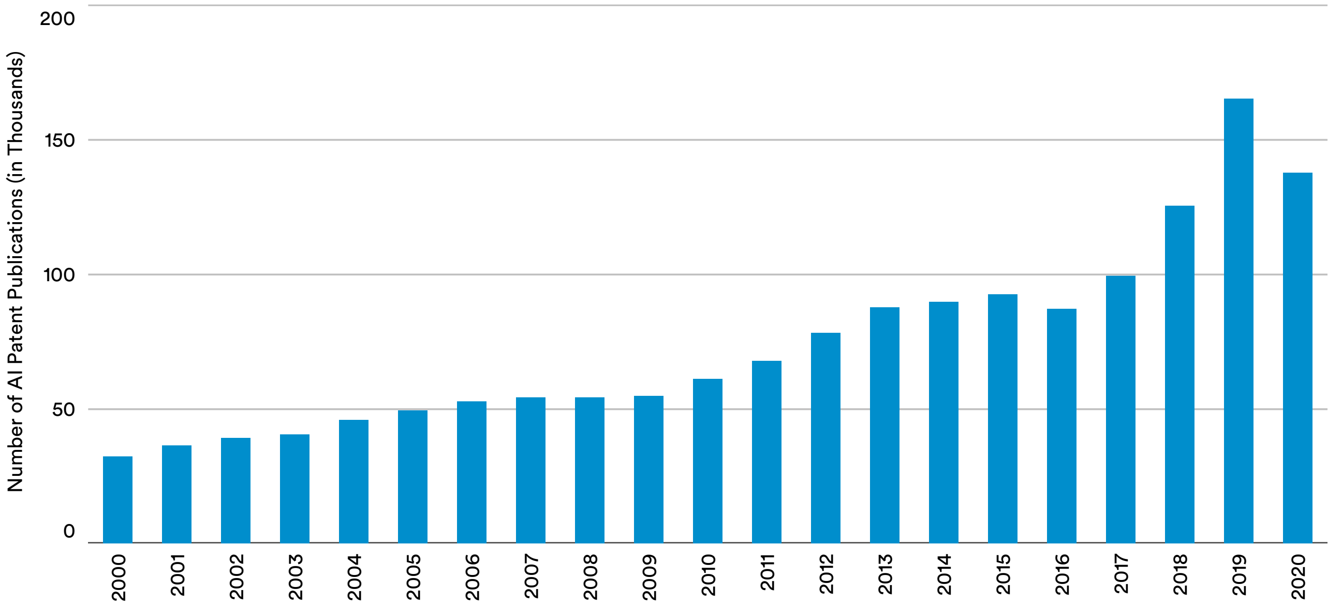

Figure 1.5.9a

OECD DEFINITION: AI PATENT PUBLICATIONS (% of ALL PATENT PUBLICATIONS), 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

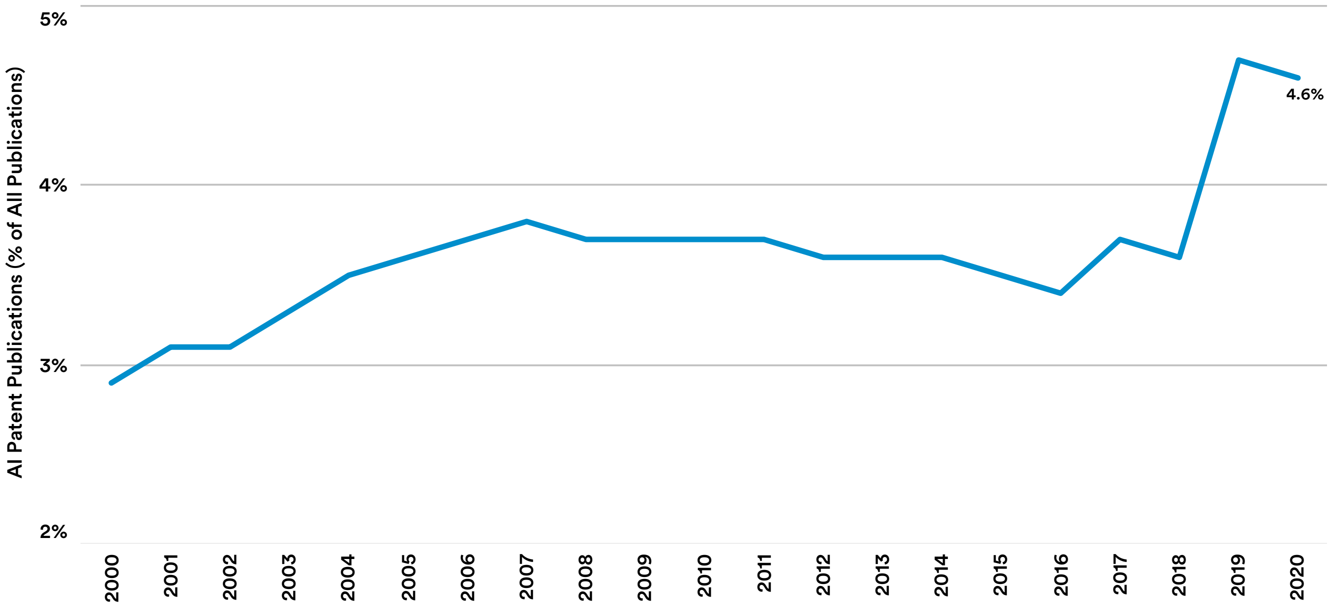

Figure 1.5.9b

OECD DEFINITION: AI PATENT PUBLICATION (% of WORLD TOTAL) by REGION, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

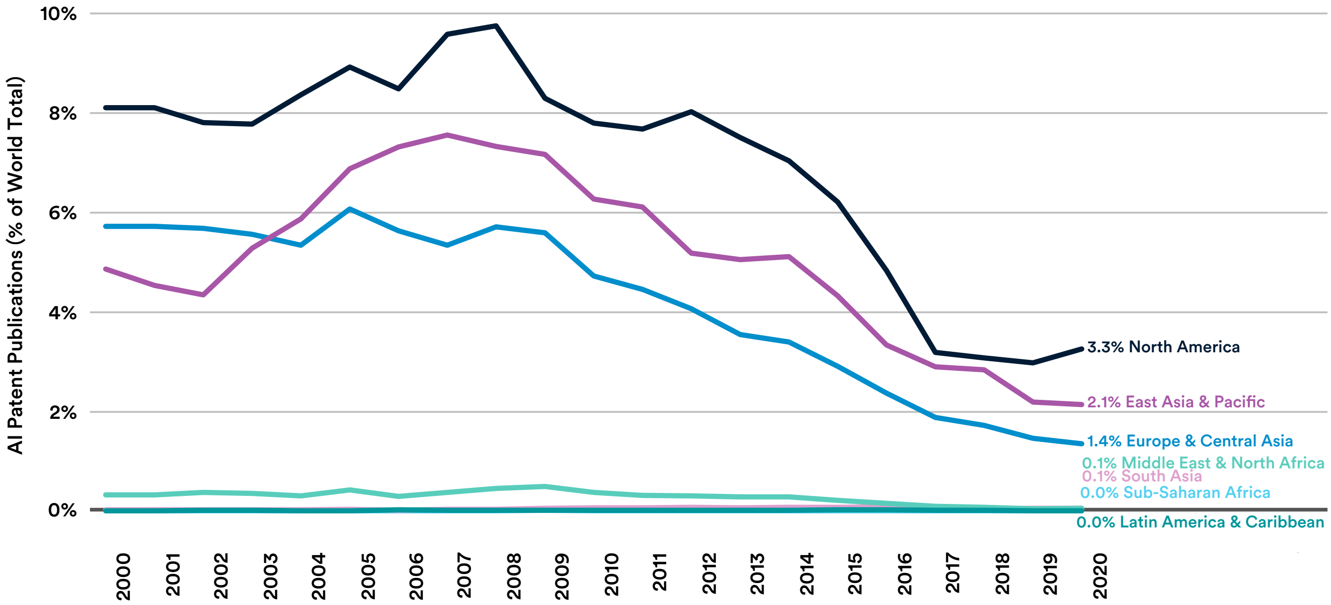

Figure 1.5.10

OECD DEFINITION: AI PATENT PUBLICATIONS (% of WORLD TOTAL) by GEOGRAPHIC AREA, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

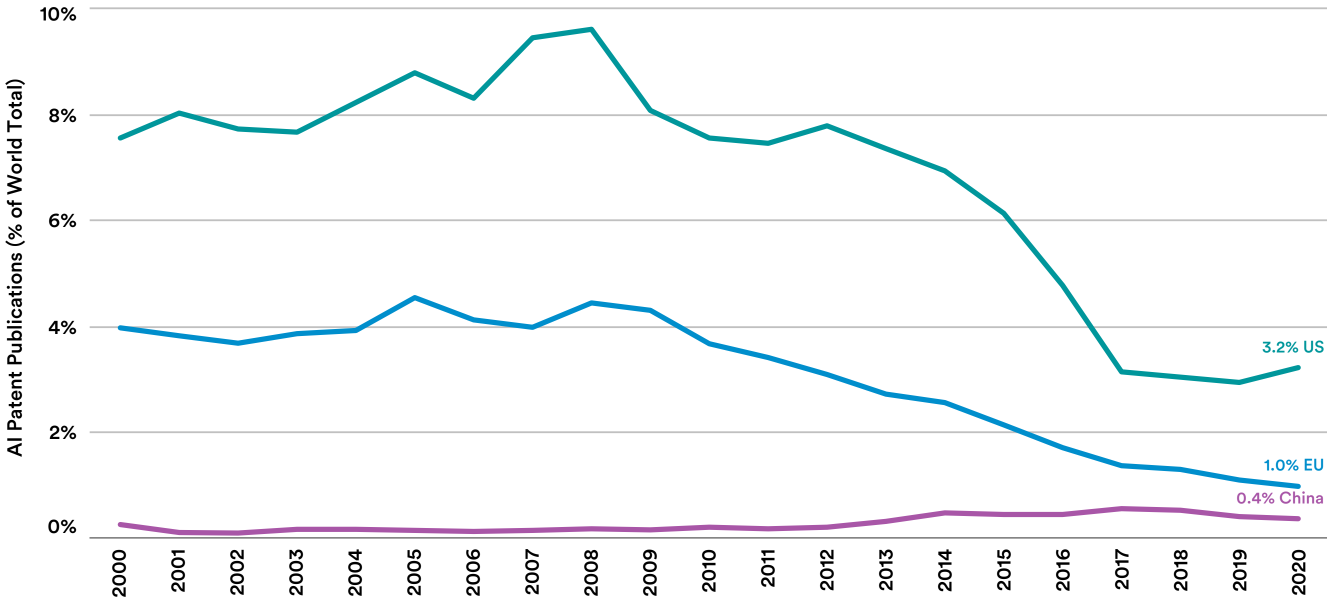

Figure 1.5.11

OECD DEFINITION: AI PATENT CITATION (% of WORLD TOTAL) by GEOGRAPHIC AREA, 2000-20

Source: Microsoft Academic Graph, 2020 | Chart: 2021 AI Index Report

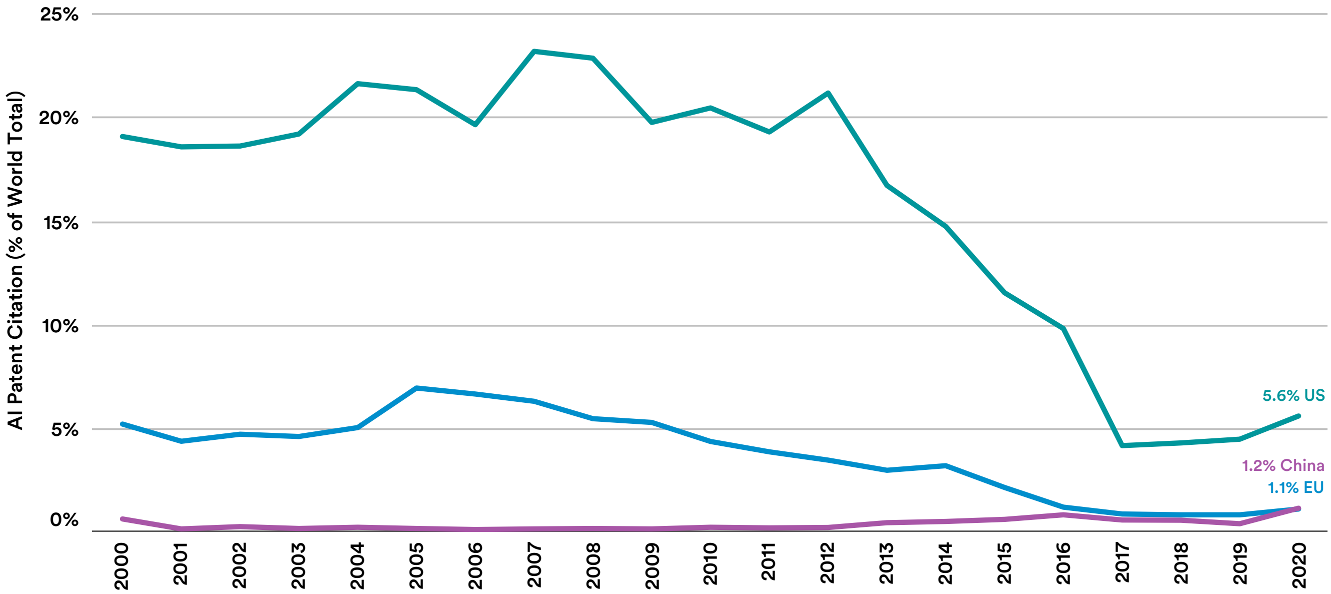

Figure 1.5.12

PAPERS ON ARXIV

Prepared by Jim Entwood and Eleonora Presani

Source

arXiv.org is an online archive of research articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. arXiv is owned and operated by Cornell University. See more information on arXiv.org.

Methodology

Raw data for our analysis was provided by representatives at arXiv.org. The keywords we selected, and their respective categories, are below:

Artificial intelligence (cs.AI)
Computation and language (cs.CL)
Computer vision and pattern recognition (cs.CV)
Machine learning (cs.LG)
Neural and evolutionary computing (cs.NE)
Robotics (cs.RO)
Machine learning in stats (stats.ML)

For most categories, arXiv provided data for 2015–2020. To review other categories' submission rates on arXiv, see arXiv.org's submission statistics.

The arXiv team has been expanding the publicly available submission statistics. This is a tableau-based application with tabs at the top for various displays of submission stats and filters on the side bar to drill down by topic. (Hover over the charts to view individual categories.) The data is meant to be displayed on a monthly basis with download options.

arXiv is actively looking at ways to improve how it can better support AI/ML researchers as the field grows and discovering content becomes more challenging. For example, there may be ways to create finer grained categories in arXiv for machine learning to help researchers in subfields share and find work more easily. The other rapidly expanding area is computer vision, where there is considerable overlap for ML applications of computer vision.

Nuance

- Categories are self-identified by authors—those shown are selected as the "primary" category. Thus there is not a single automated categorization process. Additionally, the artificial intelligence or machine learning categories may be categorized by other subfields or keywords.
- arXiv team members suggest that participation on arXiv can breed greater participation, meaning that an increase in a subcategory on arXiv could drive over-indexed participation by certain communities.

NESTA

Prepared by Joel Kliger and Juan Mateos-Garcia

Source

Details can be found in the following publication: Deep Learning, Deep Change? Mapping the Development of the Artificial Intelligence General Purpose Technology

Methodology

Deep learning papers were identified through a topic modeling analysis of the abstracts of arXiv papers in the CS (computer science) and stats.ML (statistics: machine learning category) arXiv categories. The data was enriched with institutional affiliation and geographic information from the Microsoft Academic Graph and the Global Research Identifier. Nesta's arXlive tool is available here.

Access the Code

The code for data collection and processing can be found here; or, without the infrastructure overhead here.

GITHUB STARS

Source

GitHub: star-history (available at star history website) was used to retrieve the data.

Methodology

The visual in the report shows the number of stars for various GitHub repositories over time. The repositories include the following:
apache/incubator-mxnet, BVLC/cafe, cafe2/cafe2, dmlc/mxnet, fchollet/keras, Microsoft/CNTK, pytorch/pytorch, scikit-learn/scikit-learn, tensorflow/tensorflow, Theano/Theano, Torch/Torch7.

Nuance

The GitHub Archive currently does not provide a way to count when users remove a star from a repository. Therefore, the reported data slightly overestimates the number of stars. A comparison with the actual number of stars for the repositories on GitHub reveals that the numbers are fairly close and that the trends remain unchanged.