Data Brief

Counting AI Research

Exploring AI Research Output in English- and Chinese-Language Sources

Author
Daniel Chou

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

July 2022

Introduction

U.S. policymakers care about research output. Tracking research output can, among other things, inform assessments of any given country's innovativeness or assist in evaluating the impact of certain funding initiatives. Specific to artificial intelligence (AI) and other emerging technologies, U.S. policymakers are eager to understand and compare research trends. But measuring research output is not as straightforward as it may seem. There are different ways to define output and different data sources that can be relied on, which can affect the outcomes of assessments.[1]

One way of measuring output is by counting the number of publications in a research area as an indicator of interest or focus in that research area. Capturing research quality, not just quantity, by taking citations into account is another approach to measuring output.[2] This data brief takes the former approach of counting publication quantities as a way to explore trends in AI research.[3]

When counting AI publication output, it is critical to consider where the publications are published and the language in which a work is published. In developing CSET's merged corpus of scholarly literature, we intentionally incorporated China National Knowledge Infrastructure (CNKI; 中国知网), a key Chinese-language data source, in addition to predominantly English-language sources, specifically Web of Science, Digital Science Dimensions, Microsoft Academic Graph, arXiv, and Papers With Code.

The inclusion of CNKI means that counting AI research at CSET may produce different results, compared to analyses without CNKI. While there may be times when it is appropriate to separate CNKI and predominately English-language publications, the ability to analyze them jointly offers a broader view of the research landscape and an in-depth exploration of Chinese-language AI research. This data brief explores the implications of including CNKI by presenting AI research publication trends in CSET's merged corpus including and excluding CNKI.

Table of Contents

Data and Methodology

To count artificial intelligence (AI) research output, we examine publications in CSET's merged corpus of scholarly literature, which includes Clarivate's Web of Science, Digital Science Dimensions, Microsoft Academic Graph, arXiv, Papers With Code, and the China National Knowledge Infrastructure.[4] The inclusion of CNKI in our merged corpus results in 43 million Chinese-language publications (17 percent of publications), whereas without CNKI we observe 4.2 million Chinese-language publications (2 percent of publications).[5]

We use "publication" to refer to all document types in the merged corpus, which includes journal articles, conference papers, book chapters, and more. To identify AI publications in the merged corpus, we use a human-curated list of bilingual AI keywords. Publications with a keyword match in their title, abstract, or, when available, full-text were tagged as AI-relevant.[6] This approach differs from previous CSET research that identifies AI research using a model trained to predict AI-relevance for English-language publications.[7] We opted to use the same keyword-based search to identify AI publications in both Chinese and English to avoid using different methodologies for different languages.

When analyzing publications, we extract reported author organization affiliations, the country of affiliated organizations, and any acknowledged funding organizations. To assign a publication to a country, we use the country of the affiliated organizations. If the same organization is listed multiple times on a publication, then we assign the paper to that organization only once. Likewise, if multiple author-affiliated organizations from the same country are listed on the paper, we assign the paper to that country only once. See Appendix A for more details and Appendix D for list of keywords.

Global AI Research Output

We first compare trends in AI research output by analyzing the number and share of AI publications by contributing country, the location of organizations producing AI research, and the organizations funding AI research.

China and the United States were the leaders in AI research output in 2020 by raw publication counts, as seen in Figure 1A. This holds when examining CSET's merged corpus with and without CNKI, though the number of AI publications with Chinese-affiliated authors skyrockets with the inclusion of CNKI, jumping from 61,999 to 254,098—an additional 192,099 AI publications captured by CNKI as seen in Figure 1B.[8]

Figure 1A: Top Contributing Countries to AI Publications in 2020, without CNKI

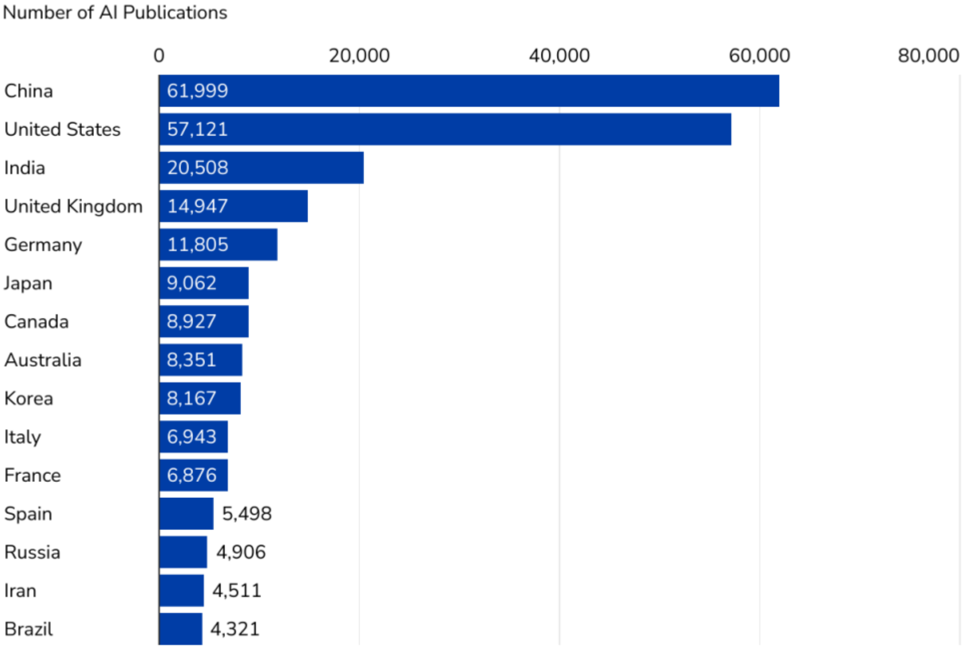
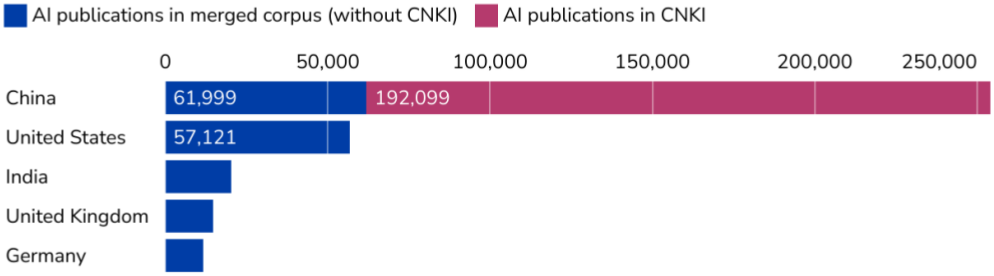


Figure 1B: Top Contributing Countries to AI Publications in 2020, with CNKI



Source: CSET merged corpus.

Next, we present the share of AI publications from the United States, China, and the rest of the world (ROW) from 2010–2020 in CSET's merged corpus, first excluding CNKI (Figure 2A) and then including CNKI (Figure 2B).[9]

Figure 2A: Country Share of AI Publications 2010–2020, without CNKI

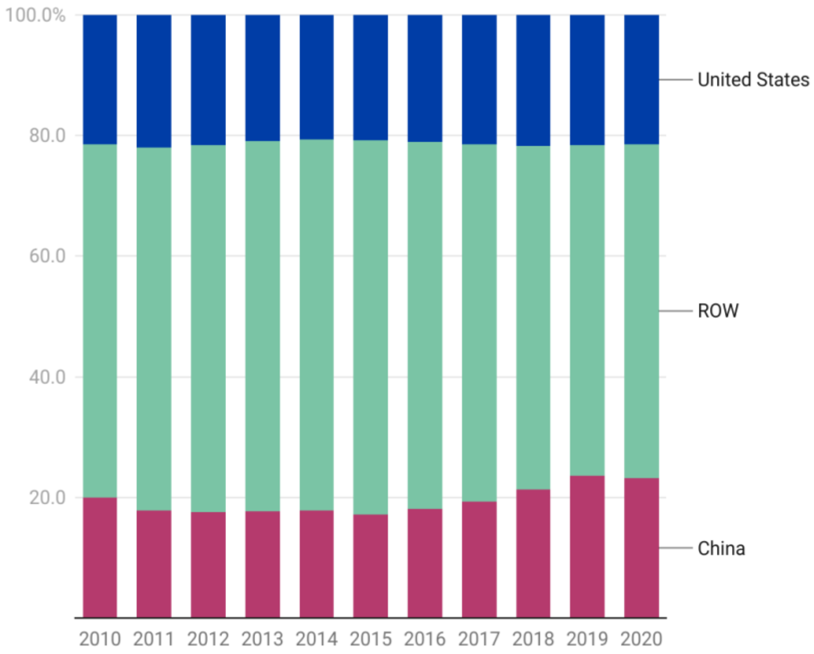
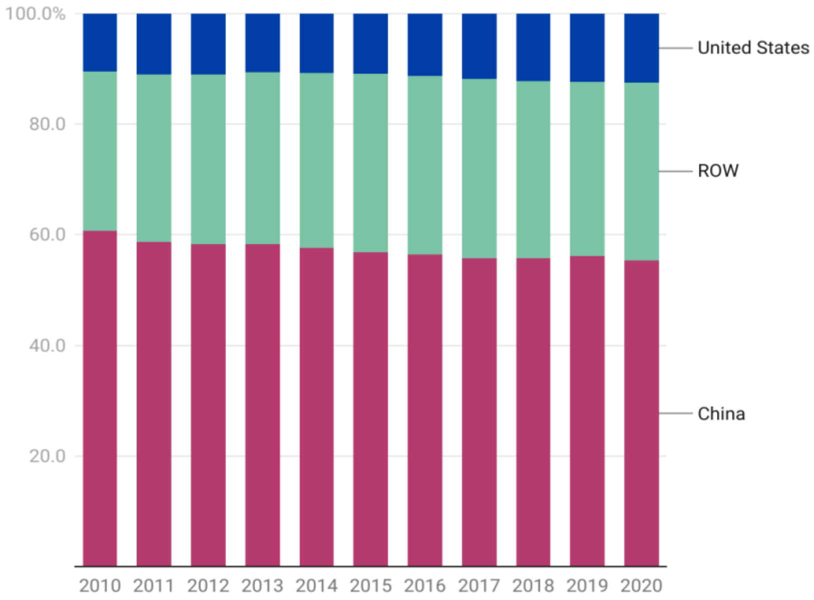


Figure 2B: Country Share of AI Publications 2010–2020, including CNKI

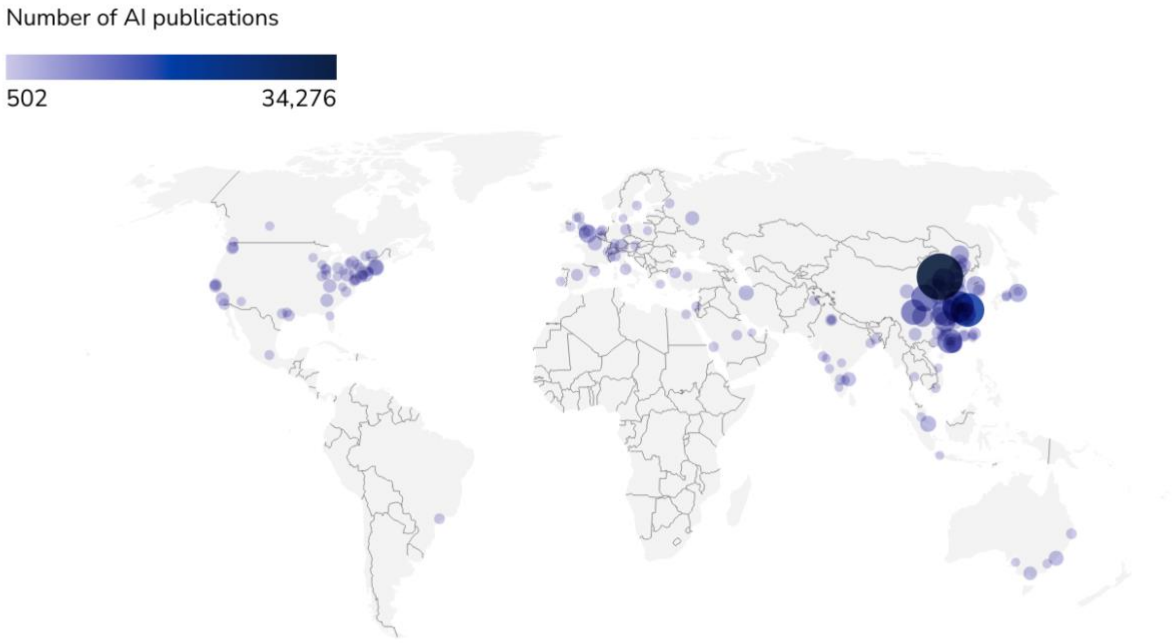


Source: CSET merged corpus.

Looking at the relative shares of AI publications from 2010–2020, we see that including CNKI presents China as a top producer of AI research in a much starker contrast, with nearly half of the world's AI publications coming from the country.

Figure 3 maps where this global AI output is coming from.[10] Each bubble denotes a city which houses organizations that collectively produced more than five hundred AI publications in 2020 in CSET's merged corpus, including CNKI. Note that some publications are missing location information, but we extracted city-level location information from 73 percent of AI publications from 2010–2020.

Figure 3: Location of Global AI Research Output by Publication Count, 2020



Source: CSET merged corpus.

Comparing publication counts for Chinese cities in Figure 3 to a version of the same map that excludes CNKI, we find that nearly two-thirds of AI publications coming from China in Figure 3 are exclusively in CNKI.[11] For larger cities, roughly half of AI publications are CNKI-only, but for smaller cities Chinese-language CNKI publications far exceed AI publications from other sources. For example, AI publication counts increase notably for Chinese cities such as Baoding (108 to 1,207), Wuxi (354 to 1,488), and Guiyang (316 to 900) when including CNKI. This suggests that some information about Chinese AI research output—especially beyond known hubs—is missing when CNKI is excluded.

Lastly, we explore the funders acknowledged in AI publications. Table 1 shows top funders in CSET's merged corpus without CNKI.[12]

Table 1: Top Funding Organizations of AI Publications in CSET's Merged Corpus, without CNKI, 2020

Funding Organization	Number of Funded Publications
National Natural Science Foundation of China	30,156
Ministry of Science and Technology (China)	10,199
National Key Research and Development Program of China	8,318
National Science Foundation (United States)	7,279
European Commission	7,245
Ministry of Education (China)	5,649
National Institutes of Health (United States)	4,873
Fundamental Research Funds for the Central Universities (China)	3,556
National Research Foundation of Korea	2,453
China Postdoctoral Science Foundation	1,896

Source: CSET merged corpus excluding CNKI.

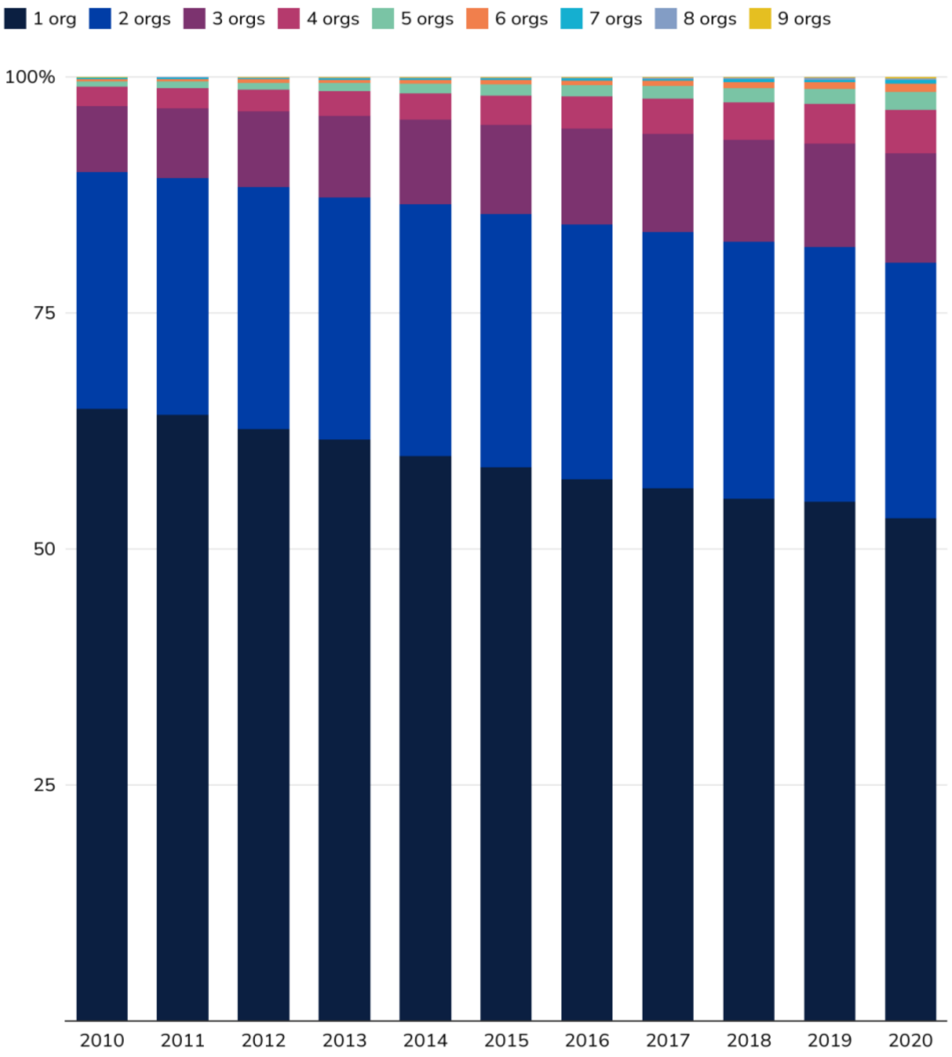
Table 1 excludes CNKI because we have not yet standardized the names of CNKI funding organizations in CSET's merged corpus. Therefore, it likely underestimates the contribution of China's funding organizations. Yet even without CNKI, the National Natural Science Foundation of China (国家自然科学基金委员会; NSFC; NNSF) is the most frequently acknowledged funding source among AI publications in CSET's merged corpus. The inclusion of CNKI would have no impact on this ranking, because more CNKI publications reported NSFC support than any other funder.

It is worth noting that funder acknowledgements are only observed in roughly one-third of publications in the merged corpus (35 percent of CNKI AI publications specifically). Also note that Table 1 shows top funding organizations acknowledged in publications, but not funding amounts, which we do not observe in the data.

AI Research Collaboration

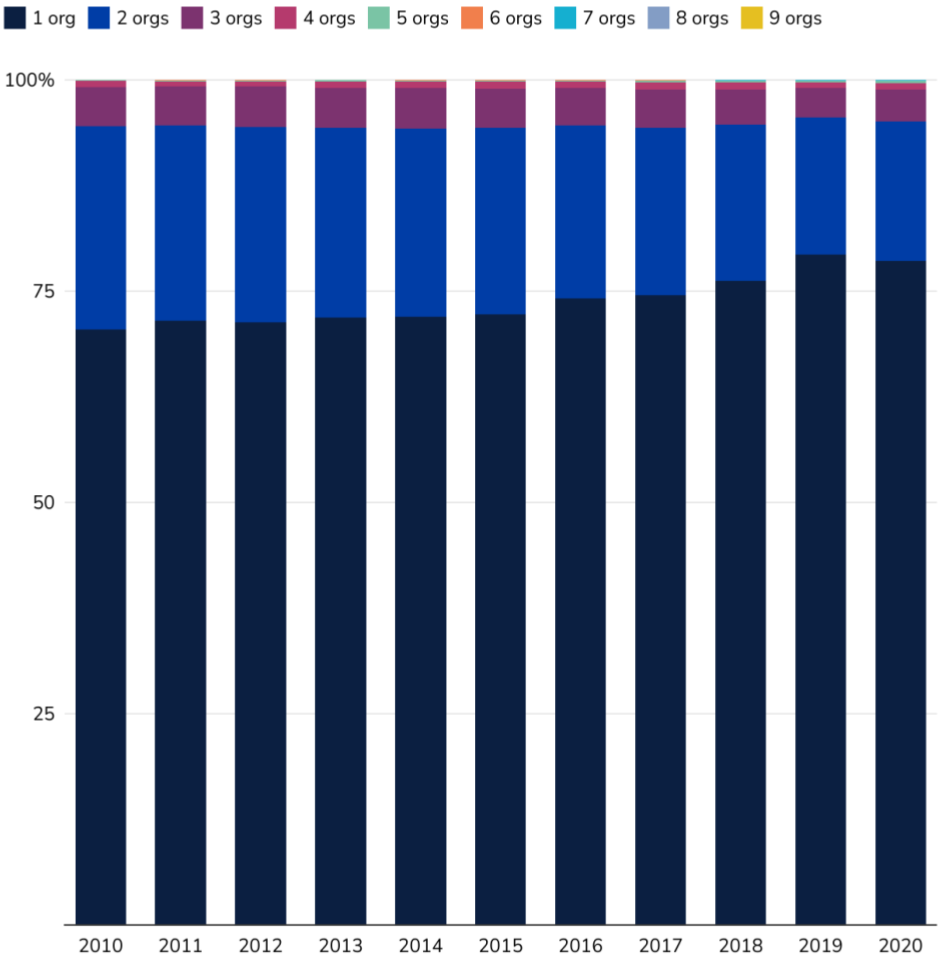
We examine collaborative trends in AI research output across CNKI publications, as well as across publications from other sources used to generate the merged corpus. One way to identify collaborative output is to look at the organization affiliations of publication authors to see if authors from different organizations are publishing together. Figure 4A displays the number of different author-affiliated organizations on AI publications from 2010 to 2020 in the merged corpus without CNKI, while Figure 4B shows the number of different organizations on AI publications in CNKI over the same period.

Figure 4A: Share of AI Publications by Number of Collaborating Organizations, 2010–2020



Source: CSET merged corpus without CNKI.

Figure 4B: Share of CNKI AI Publications by Number of Collaborating Organizations, 2010–2020



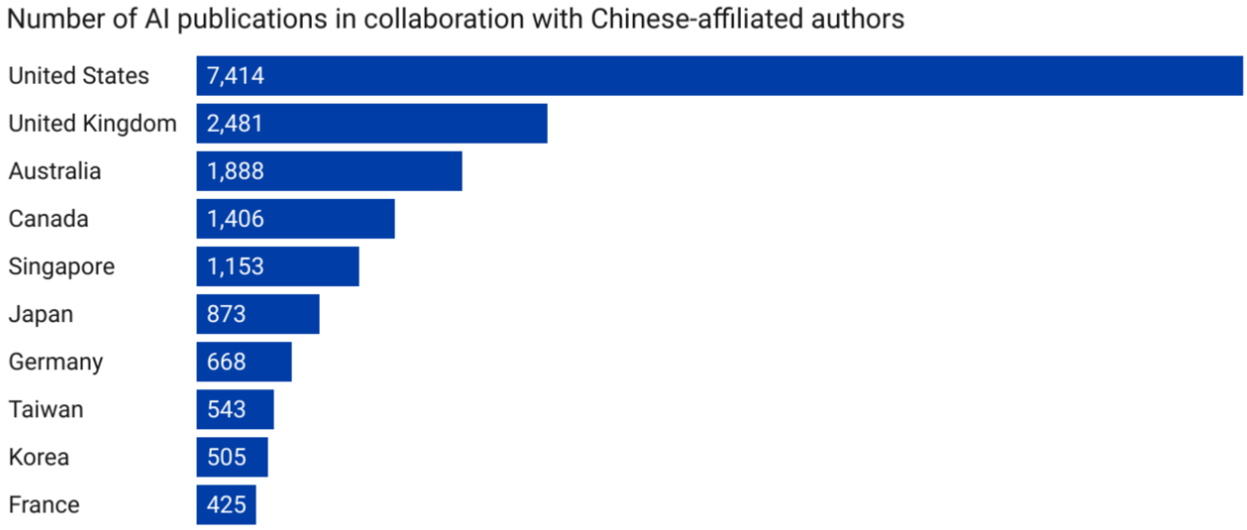
Source: CNKI.

We see a steady increase in the aggregate share of publications with authors from two or more organizations in Figure 4A.[13] However, this trend looks different when we analyze only AI publications in CNKI, as displayed in Figure 4B. Here, a smaller share of papers have authors affiliated with multiple organizations, and that holds over time, but CNKI single-organization publications are actually on the rise, making up about 70 percent of CNKI AI publications in 2010 and nearly 80 percent in 2020.[14]

Collaboration across organizations appears relatively less common among CNKI AI publications, and the same is true for the low share of China's research collaboration with other countries. Previous CSET research found that across all fields between 2010–2019, only 7 percent of Chinese-affiliated publications in CSET's merged corpus were international collaborations, compared to countries like the United Kingdom and United States, with 53 and 37 percent, respectively.[15]

Specific to AI publications, 22 percent of AI publications in 2020 with Chinese-affiliated authors in CSET's merged corpus were international collaborations. As shown in Figure 5, collaborations with authors in the United States, United Kingdom, Australia, and Canada were most frequent among AI publications with Chinese-affiliated authors. These top AI research collaborators have remained nearly unchanged since 2015.

Figure 5: Top International AI Publication Collaborators with China, 2020



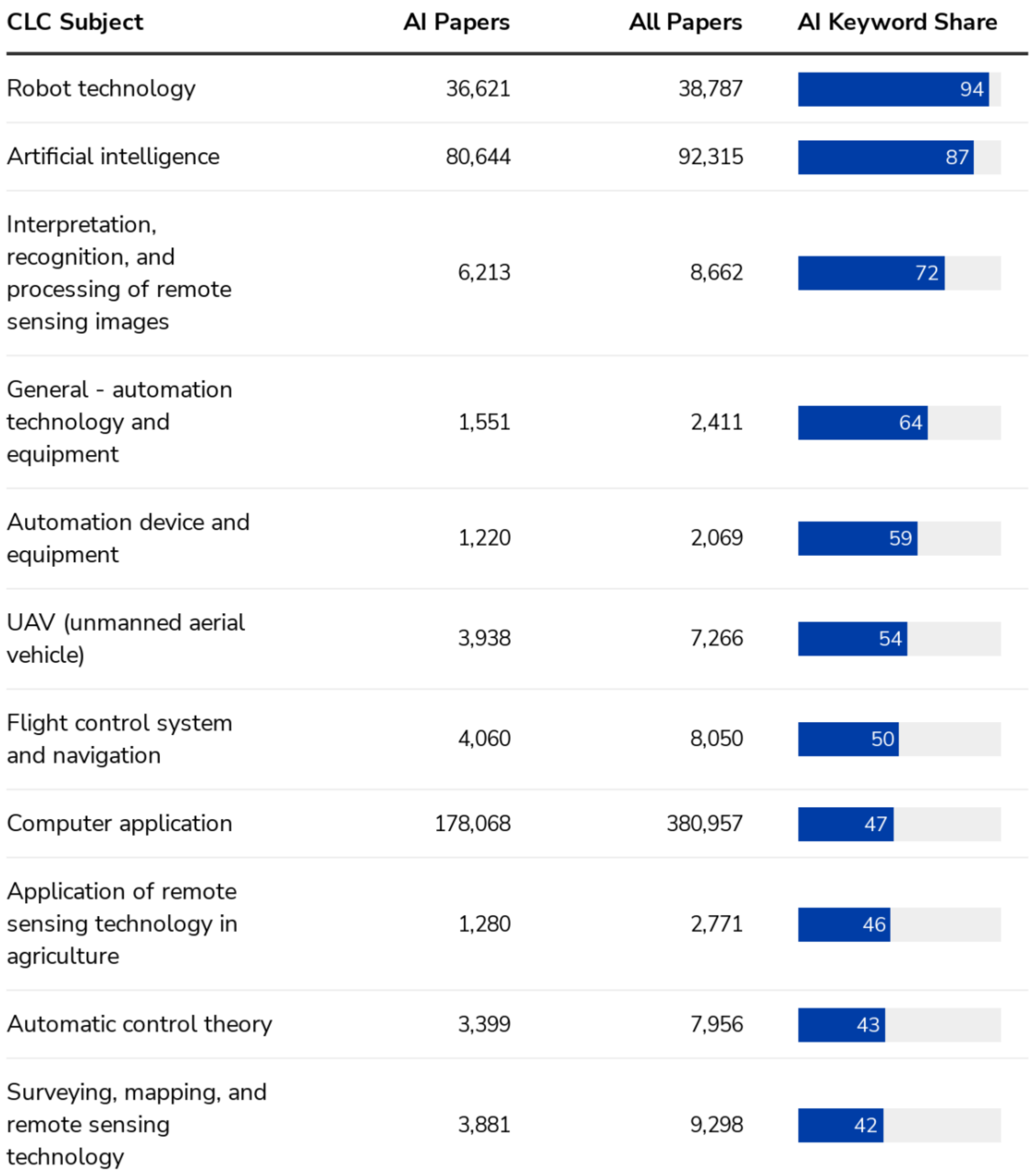
Source: CSET merged corpus.

AI Research in China

In addition to capturing more AI publications coming from China, including a source of Chinese-language publications in an examination of AI research output enables a detailed look into characteristics of China’s AI research output, such as subject areas and affiliated organizations.

For example, we can examine the Chinese Library Classification (中国图书馆分类法; 中图分类法; CLC) subject topics assigned to AI publications. CLC subjects are assigned to publications by Chinese academic journals. Table 2 lists leading CLC subjects assigned to AI publications in CNKI between 2010 and 2020, according to the share of all publications assigned that subject.

Table 2: AI Publications in CNKI by CLC Subject



CLC Subject	AI Papers	All Papers	AI Keyword Share
Robot technology	36,621	38,787	94
Artificial intelligence	80,644	92,315	87
Interpretation, recognition, and processing of remote sensing images	6,213	8,662	72
General - automation technology and equipment	1,551	2,411	64
Automation device and equipment	1,220	2,069	59
UAV (unmanned aerial vehicle)	3,938	7,266	54
Flight control system and navigation	4,060	8,050	50
Computer application	178,068	380,957	47
Application of remote sensing technology in agriculture	1,280	2,771	46
Automatic control theory	3,399	7,956	43
Surveying, mapping, and remote sensing technology	3,881	9,298	42

Source: CNKI.

Subjects including robotics, automation, and computing contain a high share of AI publications. Many AI publications in CNKI pertain to subjects in emerging application fields, such as navigation and remote sensing. That many publications in these fields are AI publications suggests these are potential application areas of AI research. Note that because AI publications were identified using a keyword-based approach, there is a good but incomplete overlap with publications assigned the “artificial Intelligence” CLC subject during this time period.

We can also examine the organizations publishing AI research specifically in CNKI.[16] Table 3A lists the top organizations by number of AI publications in CNKI in 2020.[17]

Table 3A: Organizations with Top AI Research Output in CNKI Journals, 2020

Organization	Number of Publications
University of Chinese Academy of Sciences (中国科学院大学)	1316
Wuhan University School of Information Management (武汉大学信息管理学院)	1249
Tsinghua University Department of Automation (清华大学自动化系)	1132
Shanghai Jiao Tong University (上海交通大学)	928
Sichuan University College of Computer Science (四川大学计算机学院)	889
Peking University School of Electronics Engineering and Computer Science (北京大学信息科学技术学院)	795
Renmin University of China (中国人民大学)	770
Zhejiang University (浙江大学)	770
Tongji University (同济大学)	762
University of Shanghai for Science and Technology (上海理工大学)	732

Source: CNKI.

Note: We tallied organizations that contributed to each AI paper according to preliminary results of CSET’s organization entity resolution efforts. For the set of AI papers, sometimes the authors identify more often with the department than with the university. This is common with single-subject filtering of CNKI publications.

One interesting take away from Table 3A is Wuhan University School of Information Science’s commanding lead in AI publications, which appears steady from 2015 to 2020.[18] Another takeaway is Sichuan University College of Computer Science’s jump in AI publications over recent years, from approximately five hundred published in 2015 to 889 published in 2020.[19] This regional AI research hub tends to attract less attention in analyses of China's AI research, which tend to focus on Beijing and Shanghai.

For comparison, Table 3B lists the top organizations by number of AI publications in CSET's merged corpus, including CNKI, in 2020.

Table 3B: Organizations with Top AI Research Output in CSET Merged Corpus, 2020

Organization	Number of Publications
Tsinghua University	2395
Shanghai Jiao Tong University	2076
University of Chinese Academy of Sciences	1855
Zhejiang University	1852
Harvard University	1768
Stanford University	1757
Chinese Academy of Sciences	1625
Massachusetts Institute of Technology	1548
University of Electronic Science and Technology of China	1484
Peking University	1468

Source: CSET merged corpus.

Note: Organization names in Table 3B have been standardized to facilitate analysis. Table 3B contains CSET's enhancements for CNKI-specific organization entity resolution efforts. Without these data enhancements, different organization name stylings of the same organization would be tallied separately, and this would affect the ranking of organizations by publication count.

Comparing Table 3A and Table 3B offers a perspective of China's lead in AI research output. Many leading AI research producers in Table 3B are U.S. universities, including Harvard University, Stanford University, and the Massachusetts Institute of Technology, even when we include CNKI in the merged corpus. Nevertheless, Chinese organizations from Table 3A such as Tsinghua University, Shanghai Jiao Tong University, and the University of Chinese Academy of Sciences surface as top AI research producers in Table 3B.

Conclusion

We examined AI research output across different, multilingual sources using a keyword-based approach and counting AI publications across various dimensions of interest. We show that basic trends in AI research look different when including and excluding a corpus of Chinese-language research publications. Additionally, we show that the inclusion of such a corpus enables some specific explorations of AI research in China. Our analysis shows that China's lead in AI research output is even more pronounced than many English-language sources suggest when analyzed individually.

Authors

Daniel Chou is a data scientist at CSET.

Acknowledgments

The author is especially grateful to Kevin Boyack, Zhichao Fang, and José Hernández-Orallo for their comprehensive reviews.

For feedback and assistance, the author would like to thank Catherine Aiken for data storyboarding, and Anna Puglisi and Ashwin Acharya for helpful discussions and feedback on analytical use cases of our data analysis. The author would also like to thank James Dunham, Autumn Toney, and Ben Murphy for helpful suggestions, and Shelton Fitch and Lynne Weil for editorial support. Finally, the author would like to thank Bill Hannas and Huey-Meei Chang for their encouragement to pursue this analysis. The author is solely responsible for all remaining errors.

Document Identifier: doi: 10.51593/20220010

Appendix A: Definitions and Baseline

The CSET merged corpus of scholarly literature combines a number of aggregated datasets from scholarly literature, specifically Clarivate Web of Science, Digital Science Dimensions, Microsoft Academic Graph, arXiv, Papers With Code, and CNKI. CNKI includes a wider variety of periodicals, compared to the other datasets in the merged corpus, including popular science and trade magazines. For example, a particular magazine article in CNKI touts deep learning as the next great innovation, but we do not necessarily consider it as an AI research publication. For this analysis, we consider a paper to be a "research publication" if it has a bibliographic record either in:

1. Web of Science, Dimensions, or Microsoft Academic Graph; or
2. CNKI and includes a list of references to previous publications.

Throughout the brief we drop the moniker "AI research publications" and use instead "AI publications" for brevity.

Using the keyword-based search outlined in the Data and Methodology section, we identified the following number of AI research publications from 2010 to 2020:

- CNKI: over 1.1 million
- CSET merged corpus (excluding CNKI): nearly 1.7 million

Unless otherwise noted, we remove CNKI records from the rest of the merged corpus for comparative purposes. Comparing CNKI and the rest of the merged corpus, we note:

1. There are English-language journals in the merged corpus whose publishers are based in China;
2. English-language scholarly literature sources curate some China-based journals in both English and Chinese in their collections; and
3. If the same publication is curated by both CNKI and an English-language source, then it contributes to publication tallies of both CNKI and CSET merged corpus (without CNKI) in our analysis.

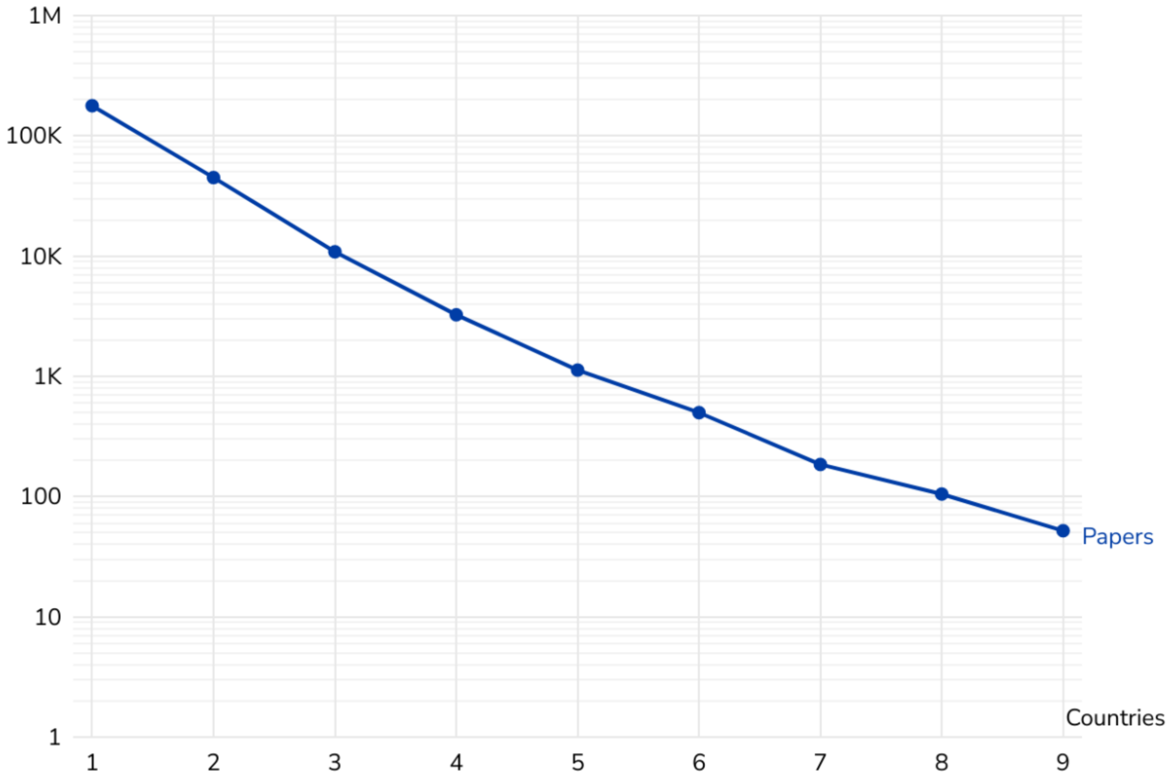
Appendix B: Collaboration Trends and Power Law

International AI Collaboration

We investigated the concept that the share of AI research collaboration (countries and organizations) in any given year follows a power law. In other words, the number of publications with N contributing organizations is a power of N for each year. A plot using the log-normal scale illustrating a functional relationship between two quantities that follow a power law tends to approximate a straight line. The following charts indicate the power-law idea has some credibility, but more in-depth analysis is needed to establish collaboration trends as a power law.[20]

Figure B1 shows collaboration trends in 2020 according to the number of distinct countries that contributed to AI research papers in CSET's merged corpus without CNKI. The vertical axis is shown in log scale, so the nearly linear trend line indicates that the number of countries collaborating in AI research follows a power-law distribution with a long tail.

Figure B1: International AI Research Collaboration in 2020

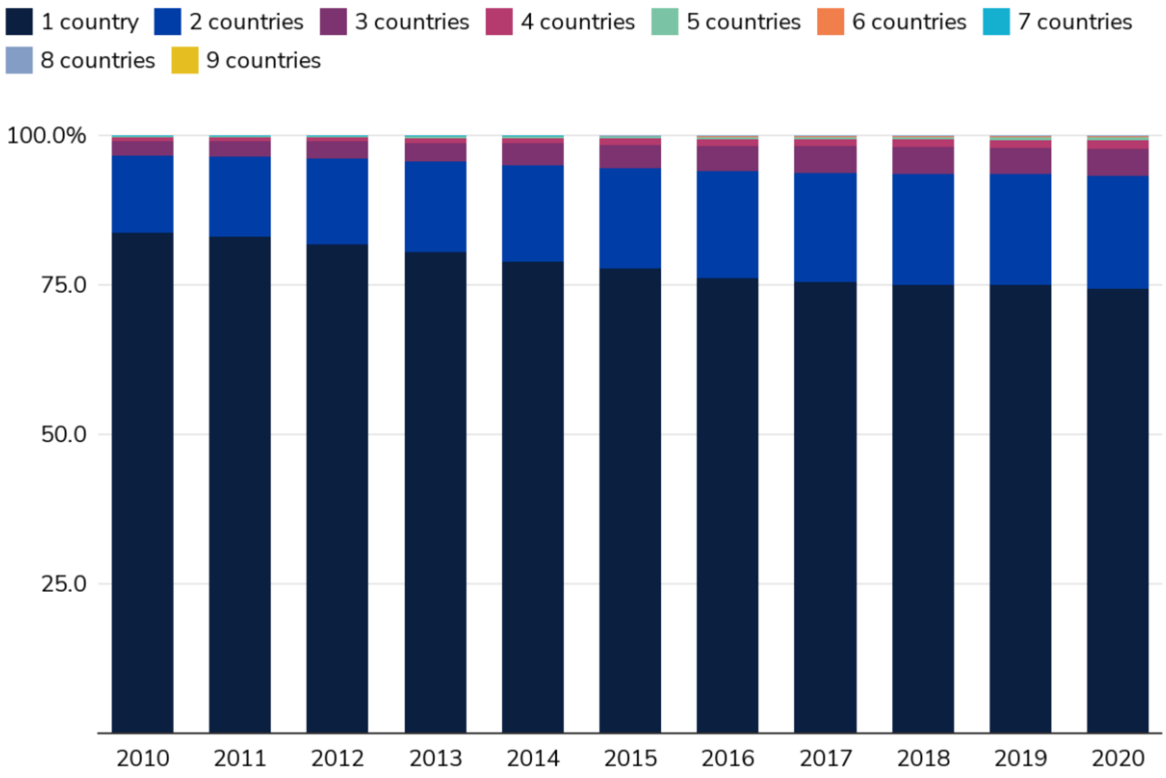


Source: CSET merged corpus without CNKI.

We omit a comparative analysis of Figure B1 with CNKI because while there are authors of CNKI publications with non-Chinese organization affiliations, the number is very low compared to publications with only Chinese organization affiliations.

If we examine international AI research collaborations over a longer period of time, then we see more publications with authors from organizations in different countries. Figure B2 shows the share of international collaborations for publications in CSET's merged corpus without CNKI from 2010 to 2020. The number of publications with two and three contributing countries increased over the period from 2010 to 2020.

Figure B2: Share of International AI Research Collaborations, 2010–2020



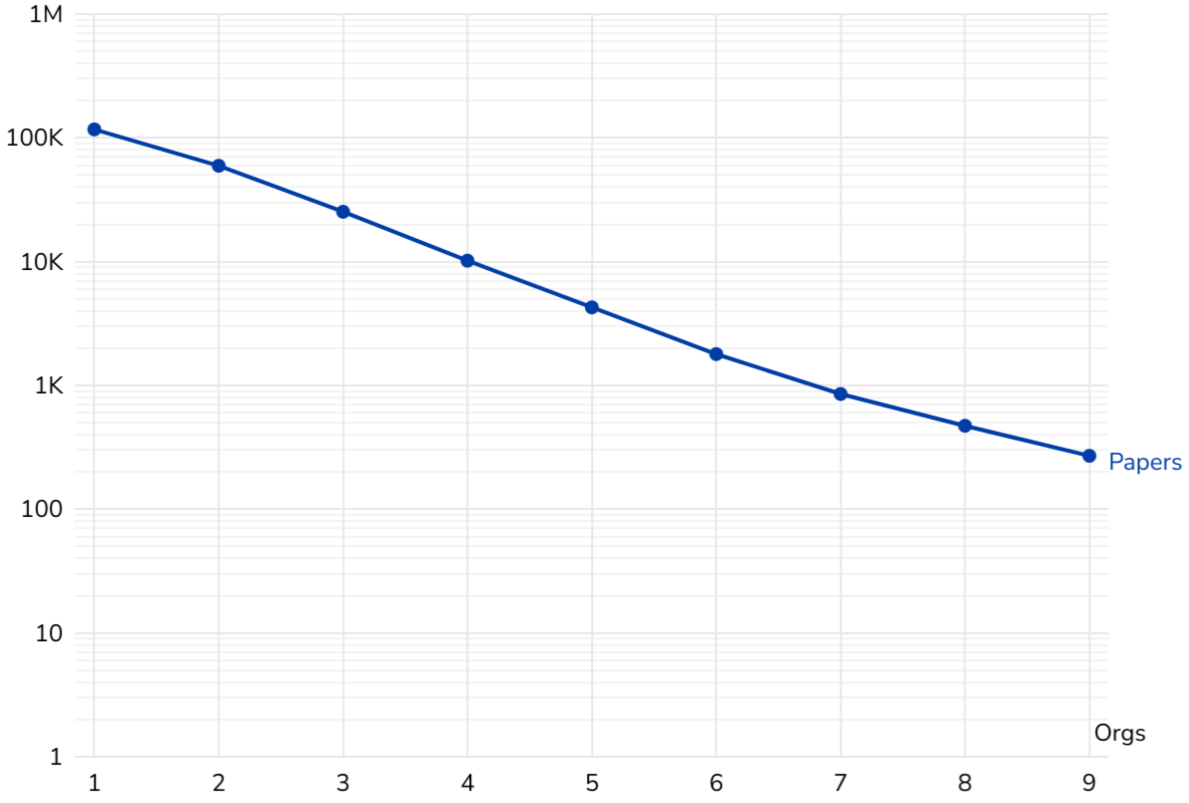
Source: CSET merged corpus without CNKI.

Figure B2 multiyear trends all exhibit similar power-law distribution with a long tail as seen in the single-year trend found in Figure B1. Note that the share of publications varies inversely with the number of contributing countries. Shares of publications with more than five collaborating countries are very small, as seen in Figure B2, and the share of papers with 20 or more contributing countries is on par with the share of papers with 10 contributing countries.

Organization AI Collaboration

Looking at interorganizational collaboration, we observe a similar trend as in international collaboration in AI research in Figure B1. Figure B3 shows AI publications in 2020 according to the number of distinct organizations that contributed to them. The vertical axis is shown in log scale, so the nearly linear trend line indicates that the number of organizations collaborating in AI research follows a power-law distribution with a long tail.

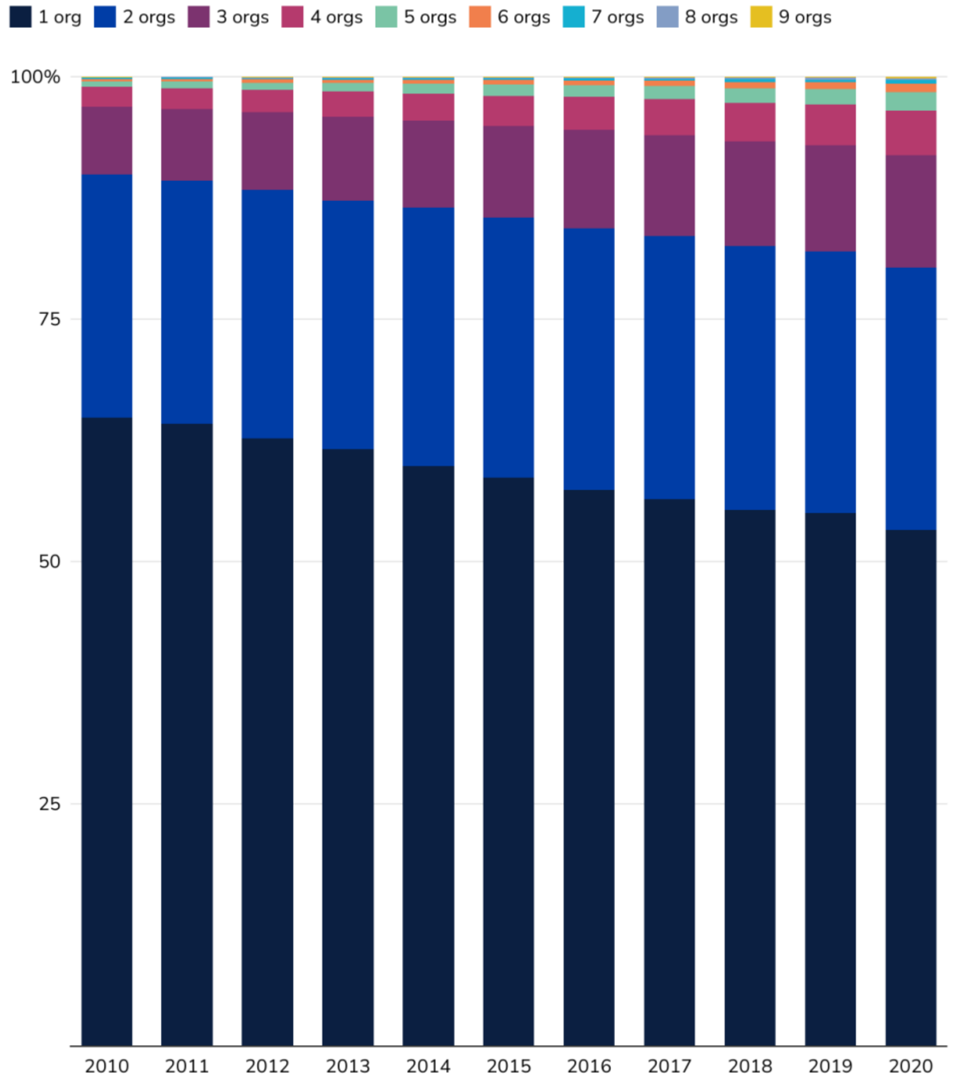
Figure B3: Organization AI Research Collaboration in 2020



Source: CSET merged corpus without CNKI.

As we did for collaborating countries in Figure B2, if we examine organization AI research collaborations over a longer period of time, then we discover more papers have authors affiliated with different organizations. We can look at Figure B4 (the same as Figure 4A above), which shows the share of interorganizational collaborations among papers with different numbers of contributing organizations from 2010 to 2020. Again, we see that the number of papers with two, three, and four contributing organizations is increasing over time, even through 2020.

Figure B4: Share of Organization AI Research Collaborations, 2010–2020



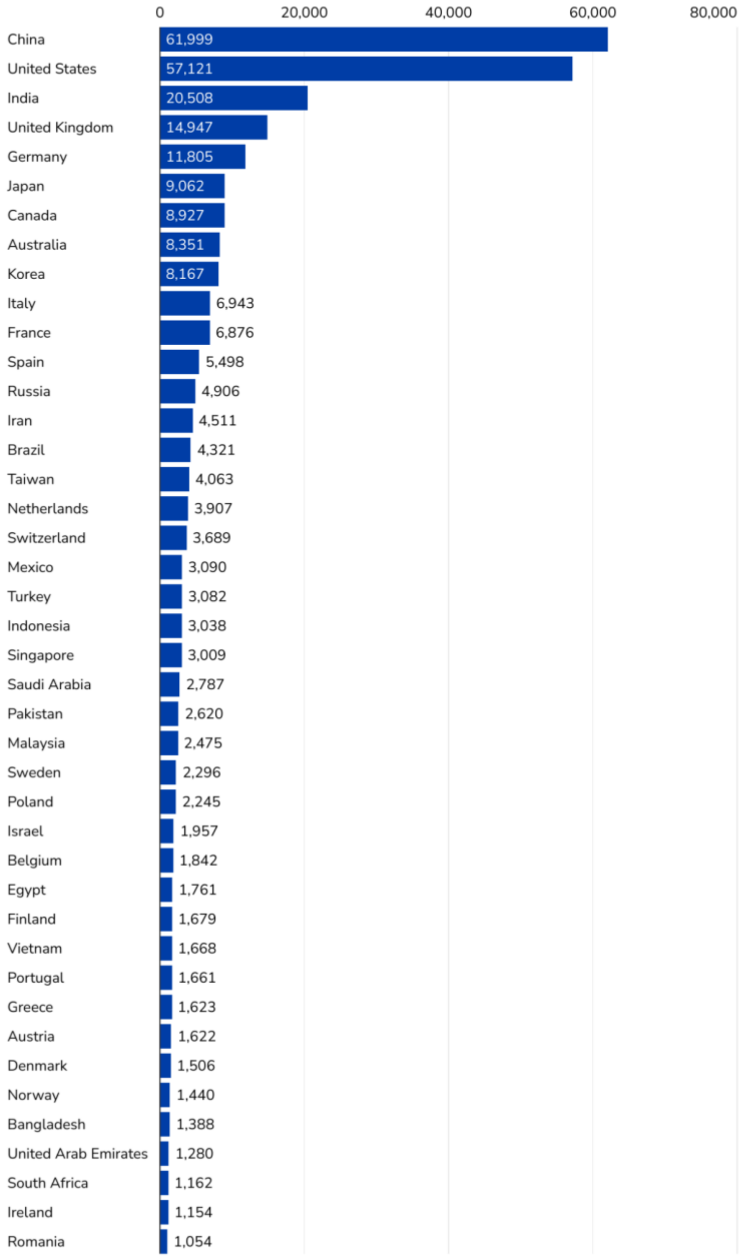
Source: CSET merged corpus without CNKI.

Both international collaboration and interorganizational collaboration for AI research tell similar data stories. From 2010 to 2020, AI researchers participated in more international collaboration and with researchers from different organizations.

Appendix C: Extended Versions of Figures and Tables

Here we provide extended versions of some of the figures presented in earlier sections for interested readers. First is Figure C1—an extended version of Figure 1A above—which presents the number of AI publications for 15 contributing countries in 2020. Figure C1 extends the list to the 42 countries that contributed to more than one thousand AI publications in 2020.

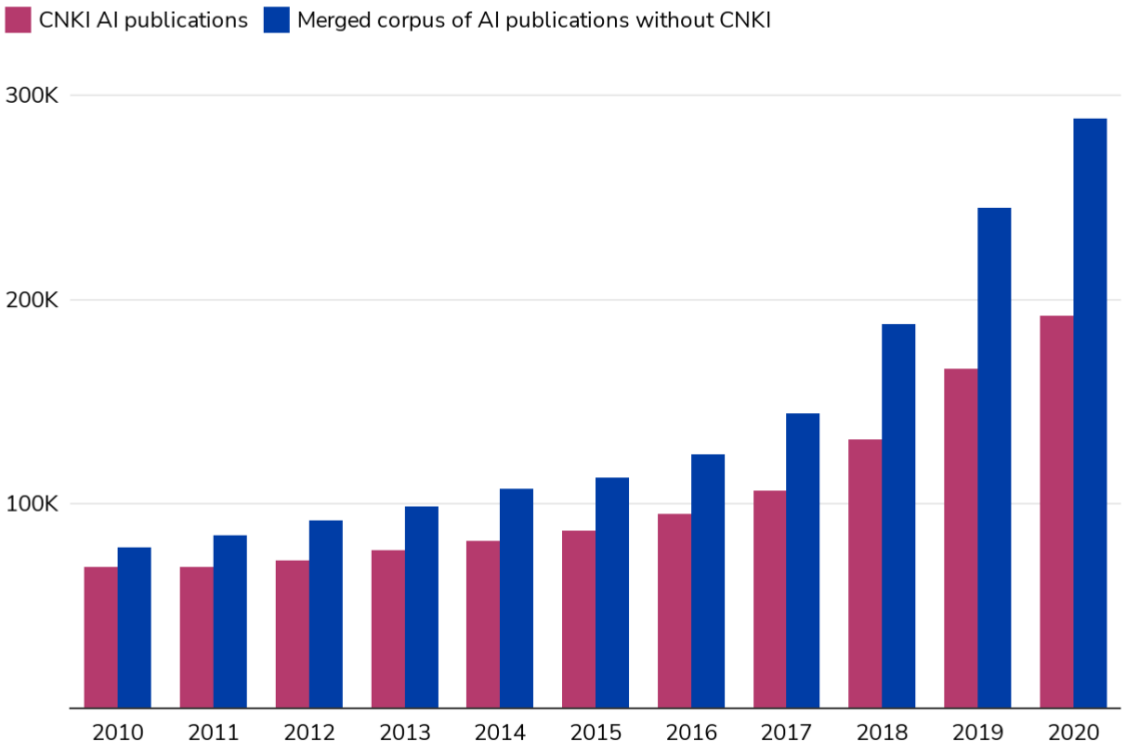
Figure C1: AI Publications by Contributing Country, 2020



Source: CSET merged corpus without CNKI.

Figure C2 offers a different view of the data in Figures 2A and 2B. AI publications increase from 2010–2020 in both Chinese-language and predominantly English-language sources. However, the growth of AI publications in the latter appears to outpace the growth of AI publications in the former.

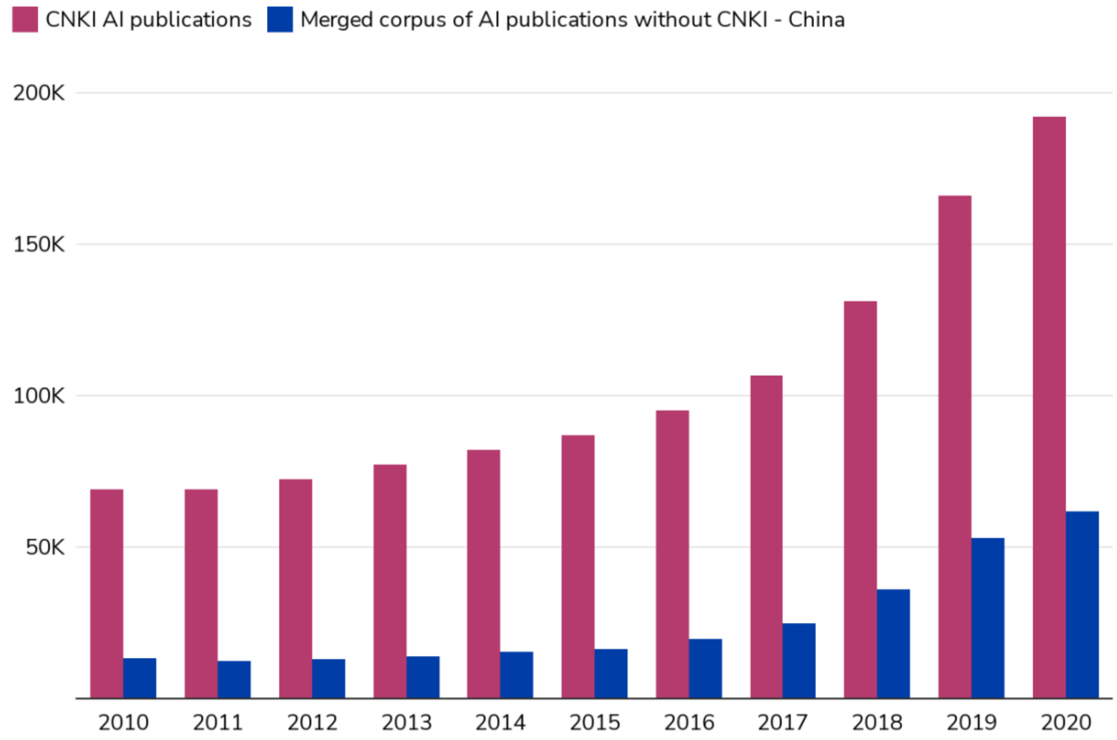
Figure C2: Global AI Research Publications by Year



Source: CSET merged corpus.

Figure C3 displays CNKI AI publications, like in Figure C2, and compares that to the number of publications in the merged corpus without CNKI with Chinese affiliated-authors. Despite significant growth in contributions to AI publications in non-CNKI venues, Chinese-affiliated AI researchers continue to publish more in Chinese-languages venues.

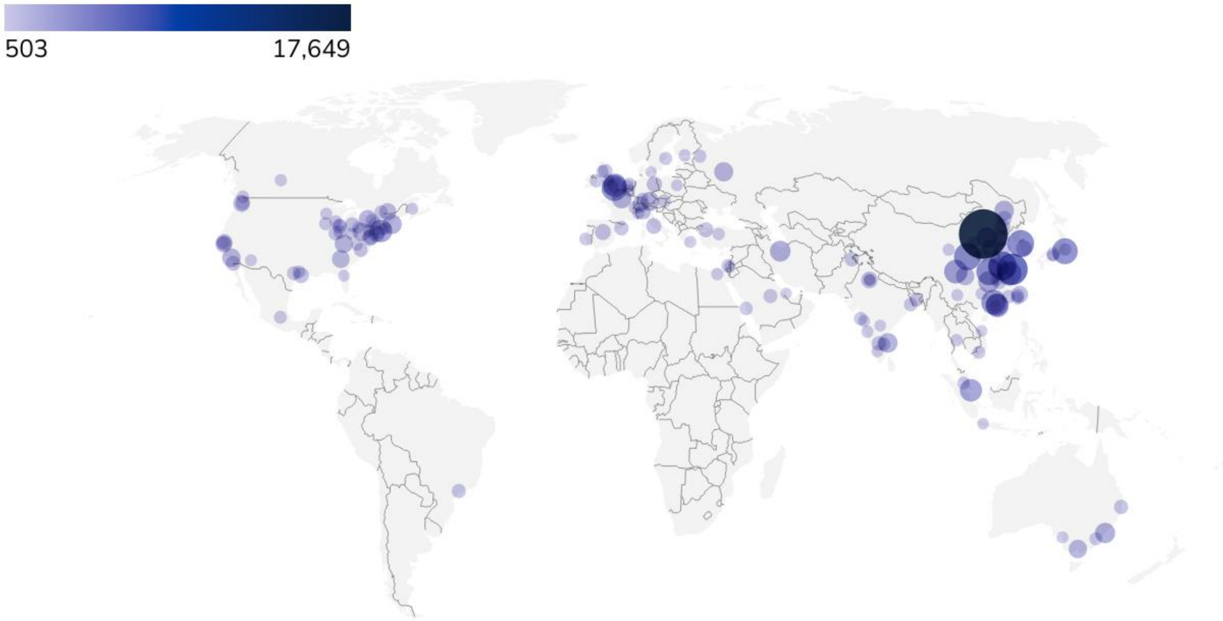
Figure C3: Chinese-Affiliated AI Research Publications by Year



Source: CSET merged corpus.

Figure C4 is a variation of the map presented in Figure 3 that omits CNKI publications. In Figure C4, each bubble denotes a city which houses organizations that produced more than five hundred AI publications in 2020. For larger cities in China, the bubble sizes in Figure C4 are roughly half as big as their Figure 3 counterparts. Furthermore, Figure 3 has far fewer (and much smaller) bubbles for smaller cities. This supports the previously discussed understanding that the breadth of AI research is widespread beyond major cities in China.

Figure C4: Location of Global AI Research Output, 2020



Source: CSET merged corpus without CNKI.

Next is an extended version of Table 1, which presents the top 10 funding organizations for AI publications in CSET's merged corpus. Table C1 extends that table to the 20 organizations.

Table C1: Top Funding Organizations of AI Publications in CSET Merged Corpus, without CNKI

2020 Rank	2020 Papers	2015 Rank	2015 Papers	Funding Organization
1	30,156	1	7,922	National Natural Science Foundation of China
2	10,199	4	2,899	Ministry of Science and Technology (China)
3	8,318	26	315	National Key Research and Development Program of China
4	7,279	5	2,802	National Science Foundation (United States)
5	7,245	2	3,485	European Commission
6	5,649	3	3,029	Ministry of Education (China)
7	4,873	6	1,746	National Institutes of Health (United States)

8	3,556	7	1,297	Subsidy for Basic Scientific Research Expenses at Central Government-Funded Universities (China)*
9	2,453	11	599	National Research Foundation of Korea
10	1,896	13	518	China Postdoctoral Science Foundation
11	1,822	8	1,041	Japan Society for the Promotion of Science
12	1,704	19	415	Ministry of Science ICT and Future Planning
13	1,644	12	565	Natural Sciences and Engineering Research Council (China)
14	1,626	10	618	Engineering and Physical Sciences Research Council (China)
15	1,571	18	429	Chinese Academy of Sciences
16	1,565	14	513	German Research Foundation
17	1,339	15	509	European Research Council
18	1,331	29	303	China Scholarship Council
19	1,185	17	490	National Council for Scientific and Technological Development (China)
20	1,167	24	351	Coordenação de Aperfeicoamento de Pessoal de Nível Superior (Brazil)

Source: CSET merged corpus.

* 中央高校基本科研业务费资助.

Next is an extended version of Table 2, which presents the top 10 CLC subjects assigned to AI publications in CNKI in 2020. Table C2 extends that table to the 30 CLC subjects with the highest AI keyword share assigned to AI publications in CNKI that year.

Table C2: AI Publications in CNKI by CLC Subject (Part 1 of 3)

CLC Subject	AI Papers	All Papers	AI Keyword Share
Robot technology	36,621	38,787	94
Artificial intelligence	80,644	92,315	87
Interpretation, recognition, and processing of remote sensing images	6,213	8,662	72
General - automation technology and equipment	1,551	2,411	64
Automation device and equipment	1,220	2,069	59
UAV (unmanned aerial vehicle)	3,938	7,266	54
Flight control system and navigation	4,060	8,050	50
Computer application	178,068	380,957	47
Application of remote sensing technology in agriculture	1,280	2,771	46
Automatic control theory	3,399	7,956	43

Source: CNKI.

Table C2: AI Publications in CNKI by CLC Subject (Part 2 of 3)

CLC Subject	AI Papers	All Papers	AI Keyword Share
Surveying, mapping, and remote sensing technology	3,881	9,298	42
Fuzzy mathematics	1,763	4,262	41
Guidance and control	2,165	5,324	41
Computer applications in agriculture	1,779	4,550	39
Game Theory	1,325	3,499	38
Computer applications in highway transportation and highway engineering	4,266	11,394	37
Pump	2,609	7,022	37
Information science	1,295	3,608	36
Application of remote sensing technology	1,349	3,865	35
Mechanical manufacturing process	6,075	18,638	33

Source: CNKI.

Table C2: AI Publications in CNKI by CLC Subject (Part 3 of 3)

CLC Subject	AI Papers	All Papers	AI Keyword Share
Nonlinear physics	1,120	3,881	29
Teaching theory, teaching method	1,437	5,022	29
Radar	7,964	27,836	29
Optimal mathematical theory	1,017	3,608	28
Circuit and network	4,614	16,485	28
High voltage insulation technology	1,189	4,266	28
Automated system	32,051	117,160	27
Mental process and mental state	1,305	4,783	27
Missile	2,894	10,848	27
General - computing and computer technology	17,989	68,105	26

Source: CNKI.

Next is an extended version of Table 3A, which presents the top 10 organizations according to CNKI AI publications in 2020. Table C3 extends Table 3A to the 20 organizations as well as 2015 AI publication counts and rankings.

Table C3: Organizations with Top AI Research Output in CNKI Journals[21]

2020 Rank	2020 Papers	2015 Rank	2015 Papers	Organization
1	1316	2	777	University of Chinese Academy of Sciences (中国科学院大学)
2	1249	1	1037	Wuhan University School of Information Management (武汉大学信息管理学院)
3	1132	4	713	Tsinghua University Department of Automation (清华大学自动化系)
4	928	6	616	Shanghai Jiao Tong University (上海交通大学)
5	889	18	502	Sichuan University College of Computer Science (四川大学计算机学院)
6	795	31	386	Peking University School of Electronics Engineering and Computer Science (北京大学信息科学技术学院)
7	770	80	221	Renmin University of China (中国人民大学)
8	770	5	680	Zhejiang University (浙江大学)
9	762	14	528	Tongji University (同济大学)
10	732	15	527	University of Shanghai for Science and Technology (上海理工大学)
11	718	46	311	Beijing Normal University (北京师范大学)
12	715	27	435	Nanjing University (南京大学)
13	701	3	748	Nanjing University of Aeronautics and Astronautics (南京航空航天大学)
14	688	16	522	Jilin University (吉林大学)
15	686	12	531	Southeast University (东南大学)
16	658	10	567	Tianjin University (天津大学)

17	632	42	345	Wuhan University of Technology (武汉理工大学)
18	607	20	492	North China Electric Power University (华北电力大学)
19	605	23	469	Southwest Jiaotong University School of Electrical Engineering (西南交通大学电气工程学院)
20	595	25	459	Huazhong University of Science and Technology (华中科技大学)

Note: We tallied organizations that contributed to each AI paper according to preliminary results of CSET's organization entity resolution efforts. For the set of AI papers, sometimes the authors identify more often with the department than with the university. This is common with single-subject filtering of CNKI publications.

Appendix D: AI Keywords

Below are AI keywords used in our analysis. For clarity, we display them in base lowercase form and omit some variants, including punctuation and styling variations. We searched using case-insensitive regular expressions using these AI keywords. For example, we used “fac\S+ identi” to search for “facial identification” and its variants. We intended to identify publications about AI and machine learning applications in addition to fundamental research. These AI keywords are neither exhaustive nor authoritative.

- active learning
- adaptive learning
- anomaly detection
- artificial intelligence
- associative learning
- autoencoder
- autonomous navigation
- autonomous system
- autonomous vehicle
- average link clustering
- backpropagation
- binary classification
- bionlp
- boltzmann machine
- character recognition
- classification algorithm
- classification label
- clustering method
- complete link clustering
- computer aided diagnosis
- computer vision
- deep learning
- ensemble learning
- evolutionary algorithm
- facial expression recognition
- facial identification
- facial recognition
- feature extraction
- feature learning
- feature matching

- feature selection
- feature vector
- feedforward network
- fuzzy clustering
- generative adversarial network
- generative model
- gradient algorithm
- graph matching
- graphical model
- handwriting recognition
- hierarchical clustering
- hierarchical model
- human robot
- image annotation
- image classifi
- image matching
- image processing
- image registration
- image representation
- image retrieval
- incremental clustering
- information extraction
- information fusion
- information retrieval
- k-nearest neighbor
- knowledge-based system
- knowledge discovery
- knowledge representation
- language identification
- language model
- machine learning
- machine perception
- machine translation
- multi-class classification
- multi-label classification
- multitask learning
- natural language generation
- natural language processing
- natural language understanding

- neural network
- object recognition
- one-shot learning
- pattern matching
- pattern recognition
- random forest
- recommender system
- recurrent network
- reinforcement learning
- robotics
- scene classification
- scene understanding
- self-driving auto
- self-driving car
- semi-supervised learning
- sentiment classification
- single-link clustering
- spatial learning
- speech processing
- speech recognition
- speech synthesi
- statistical learning
- supervised learning
- support vector machine
- text mining
- text processing
- transfer learning
- transformer-based
- translation system
- unsupervised learning
- video classification
- video processing
- zero shot learning
- 人工智能
- 自编码
- 自动编码
- 自主导航
- 自动驾驶
- 无人驾驶
- 反向传播

- 玻尔兹曼机
- 计算机视觉
- 深度学习
- 深层学习
- 人脸识别
- 面部识别
- 面像识别
- 面容识别
- 前馈网络
- 生成对抗网络
- 生成模型
- 图模型
- 信息抽取
- 知识表示
- 语言模型
- 机器学习
- 机器翻译
- 自然语言处理
- 自然语言理解
- 神经网络
- 一次性学习
- 模式识别
- 随机森林
- 机器人学
- 循环网络
- 回馈网络
- 回归网络
- 递归网络
- 反馈网络
- 强化学习
- 语音识别
- 语音合成
- 监督学习
- 支持向量机
- 迁移学习
- 转移学习
- 基于变换器
- 人机交互
- 机器视觉
- 零次学习
- 零样本

Endnotes

[1] Dewey Murdick, James Dunham, and Jennifer Melot, "AI Definitions Affect Policymaking" (Center for Security and Emerging Technology, June 2020), https://cset.georgetown.edu/publication/ai-definitions-affect-policymaking/.

[2] While a definition for "output" that accounts for citations has several benefits, it may be limited when examining Chinese- and English-language research together, as Chinese-language publications tend to be cited less often by international publications; see Ashwin Acharya and Brian Dunn, "Comparing U.S. and Chinese Contributions to High-Impact AI Research" (Center for Security and Emerging Technology, January 2022), https://cset.georgetown.edu/publication/comparing-u-s-and-chinese-contributions-to-high-impact-ai-research/). Thus, lower citation counts of Chinese-language AI publications may not necessarily imply low quality or irrelevant research.

[3] For an overview of alternative approaches to identifying and aggregated AI-related research utilized in CSET research, see "Identifying AI Research" (Center for Security and Emerging Technology, *forthcoming*).

[4] Dimensions is an interlinked research information system provided by Digital Science (http://www.dimensions.ai). All China National Knowledge Infrastructure content is furnished for use in the United States by East View Information Services, Minneapolis, MN, USA.

[5] CSET's merged corpus includes 80,355,458 non-English language publications, of which 54 percent (43,774,509) are Chinese-language publications. Omitting CNKI publications from CSET's merged corpus reduces the Chinese-language share among non-English language publications to 10 percent (as of August 2021).

[6] Full text is available for CNKI publications, but not for many publications in the merged corpus from other sources. See Appendix D for the list of keywords used in our analysis.

[7] For examples of CSET research using this alternative approach, see Autumn Toney, "Creating a Map of Science and Measuring the Role of AI in it" (Center for Security and Emerging Technology, June 2021), https://cset.georgetown.edu/publication/creating-a-map-of-science-and-measuring-the-role-of-ai-in-it/; Margarita Konaev et al., "Headline or Trend Line? Evaluating Chinese-Russian Collaboration in AI" (Center for Security and Emerging Technology, August 2021), https://cset.georgetown.edu/publication/headline-or-trend-line/; and Husanjot Chahal et al., "Quad AI: Assessing AI-related Collaboration between the United States, Australia, India, and Japan" (Center for Security and Emerging Technology, May 2022), https://cset.georgetown.edu/publication/quad-ai/. For information on development of classifier to identify AI-relevant research see James Dunham, Jennifer Melot, and Dewey Murdick, "Identifying the Development and Application of Artificial Intelligence in Scientific Text," arXiv preprint arXiv:2002:07143 (2020), https://arxiv.org/abs/2002.07143. This is also the approach used in the 2022 AI Index Report, "Artificial Intelligence Index Report 2022" (Stanford Institute for Human-Centered Artificial Intelligence, 2022), https://aiindex.stanford.edu/report/), which relied on CSET's merged corpus to analyze AI research. CSET has used alternative methodologies to identify AI-relevant research, including identifying research clusters in CSET's Map of Science (Jennifer

Melot and Ilya Rahkovsky, "CSET Map of Science" (Center for Security and Emerging Technology, October 2021), https://cset.georgetown.edu/publication/cset-map-of-science/) that have some threshold of AI-relevant papers, and considering the publications within those clusters to constitute AI research (see Ashwin Acharya and Brian Dunn, "Comparing U.S. and Chinese Contributions to High-Impact AI Research" (Center for Security and Emerging Technology, January 2022), https://cset.georgetown.edu/publication/comparing-u-s-and-chinese-contributions-to-high-impact-ai-research/). See forthcoming CSET report "Identifying AI Research" for a more complete overview of these various approaches.

[8] See Appendix C for a version of Figure 1A that includes AI research output of more countries. Figure 1B does not include any additional papers for countries other than China because we omitted non-Chinese affiliations for the small number of CNKI publications with non-Chinese organization affiliations, which we estimate to be less than 0.5 percent of CNKI publications (less than one hundred thousand). Our efforts to extract and standardize non-Chinese organization affiliations from CNKI publications are ongoing.

[9] Figure 2A presents shares of AI publications by country that differ from previously published CSET research, such as Acharya and Dunn, "Comparing U.S. and Chinese Contributions to High-Impact AI Research" (see specifically Figure 1), because they rely on different methodologies for identifying AI-relevant publications. For an overview of the different methodologies CSET developed for identifying AI-relevant research, see "Identifying AI Research" (*forthcoming*).

[10] Similar to publication country assignment, to assign a publication to a city, we extracted the city of the affiliated organization(s) if available. If the same city is listed multiple times on a publication, then we assign the paper to that city only once. Likewise, if multiple author-affiliated organizations from the same city are listed on the paper, we assign the paper to that city only once.

[11] See Figure C4 in Appendix; See also Figure 3 in Anna Puglisi and Daniel Chou, "China's Industrial Clusters" (Center for Security and Emerging Technology, June 2022), https://cset.georgetown.edu/publication/chinas-industrial-clusters/, for a map zoomed in on China for only CNKI AI publications.

[12] See Appendix C for an extended table of top acknowledging funding sources.

[13] Appendix B includes a similar analysis comparing contributing countries instead of contributing organizations, suggesting a similar trend of more countries contributing to papers over time.

[14] We explored the idea that the share of organization AI research collaboration in any given year follows a power law; this might suggest AI research collaboration network is scale-free. Scale-free networks exhibit properties that simplify computation of network statistics. See Appendix B for more discussion.

[15] See Autumn Toney and Melissa Flagg, "Research Impact, Research Output, and the Role of International Collaboration" (Center for Security and Emerging Technology, November 2021),

https://cset.georgetown.edu/publication/research-impact-research-output-and-the-role-of-international-collaboration/, especially Appendix Figure A.

[16] Note that the number of publications is one measure of an organization's contributions to AI research but not the only one. It does not directly capture research quality or innovativeness, and excludes research that is not openly published.

[17] See Appendix C for an extended table of top acknowledging funding sources.

[18] See Table C3 for comparison.

[19] See Table C3 for comparison.

[20] There are established methods for demonstrating whether trends follow a power law, but this analysis is beyond the scope of our discussion. See Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman, "Power-Law Distributions in Empirical Data," *SIAM Review* 51, no. 4 (2009): 661–703 and Michael P. H. Stumpf and Mason A. Porter, "Critical Truths about Power Laws," *Science* 335, no. 6069, (2012): 665–666.

[21] See a variation of this table in William C. Hannas et al., "China Advanced AI Research" (Center for Security and Emerging Technology, July 2022).