January 2022

Comparing U.S. and Chinese Contributions to High-Impact AI Research

CSET Data Brief

AUTHORS
Ashwin Acharya
Brian Dunn

Executive Summary

While the large and growing number of Chinese artificial intelligence publications is well known, the quality of this research is debated. Some observers claim that China is capable of producing a high quantity of AI publications, but lags in original ideas and impactful research.[1] Even Chinese researchers occasionally criticize their country's academic system for its lack of innovation in AI.[2] In recent years, however, quantitative analyses have found that Chinese AI publications are increasingly influential.[3]

AI is an economically and strategically important emerging technology, and the Chinese government has promoted domestic AI progress for years. Chinese and U.S. strengths in AI development will have ramifications for the two countries' relative capabilities in areas ranging from science and medicine to battlefield applications. Further, Chinese researchers' ability to produce impactful AI advances reflects on the more general question of whether Beijing can foster impactful innovation—a capability sometimes called into question by U.S. and European observers.[4]

This brief provides a data-driven comparison of U.S. and Chinese AI research, examining both publications that are highly cited and those published in top AI conferences.[5]

We find that:

- **Chinese researchers' output of highly cited AI publications is increasingly competitive with the work of their U.S. counterparts.** Over the past decade, Chinese researchers have published a growing share of the world's top-5-percent AI publications, rising from half of U.S. output in 2010 to parity in 2019.

- **Top Chinese publications are often cited outside of China, although China still lags behind the United States in international citations.** Highly cited Chinese publications receive 35 percent of their citations from non-Chinese sources, and their citation count from international sources

has steadily increased over time. However, U.S. publications maintain a lead over Chinese ones in international citations, reflecting the United States' closer ties to other leading AI producers.

- **China contributes an increasing share of publications at 13 top AI conferences, while the U.S. share of publications at these conferences is stagnant.** Between 2010 and 2019, China's share of these publications grew from 13 percent to 31 percent, while the U.S. share fell from 55 percent to 51 percent.

- **A notable share of both U.S. and Chinese researchers' high-impact AI publications were U.S.-Chinese collaborations.** For example, such collaborations accounted for 24 percent of both countries' highly cited AI publications in 2019.

- **Some research clusters in CSET's Map of Science contain far more top AI publications than others.** These clusters' topics reflect some areas of interest for Chinese and U.S. researchers.

    - Clusters with a disproportionate share of China's highly cited and top-venue publications include publications on general-purpose computer vision research, as well as applications of AI to surveillance and industry.

    - Clusters with a disproportionate share of the United States' highly cited and top-venue publications cover algorithmic innovations in deep learning, such as transformers and deep reinforcement learning, as well as AI ethics and safety research.

- **The United States and China combined publish about 65 percent of highly cited AI research.** U.S. allies, particularly the European Union and the Five Eyes countries, also make significant contributions to AI research.[6]

Methodology

Identifying AI Publications in the CSET Merged Corpus of Scholarly Literature

In this brief, we investigate the scale of high-quality Chinese and U.S. academic research using CSET's Map of Science, an automated grouping of publications derived from CSET's research clusters and merged corpus of scholarly literature.[7] This corpus incorporates more than 200 million publications from six academic datasets. Three of these are major scientific literature databases commonly used in U.S. bibliometric analyses: Clarivate's Web of Science (WOS), Digital Science's Dimensions (DS), Microsoft Academic Graph (MAG). We also include the China National Knowledge Infrastructure (CNKI), a large database of Chinese-language literature. Finally, we incorporate publications from two open-source datasets: preprints from ArXiv and machine learning papers from Papers With Code.

The Map of Science groups documents in the CSET merged corpus into approximately 120,000 research clusters of several hundred to several thousand publications. A research cluster consists of a group of publications that cite each other more than they cite publications outside the cluster. Since clusters are defined by citation links rather than by topic, they are not guaranteed to contain publications with a common topic. In our experience, publications in a cluster often share a fairly well-defined research area, such as facial recognition. However, a single research area is often covered by multiple clusters.

Our analysis focuses on publications that appear in research clusters that contain a significant share of AI-relevant work; we refer to publications in these AI research clusters as "AI publications." Compared to selecting publications based solely on predicted AI-relevance at paper level, this method allows us to capture more publications that apply artificial intelligence methods to other areas, such as drone piloting. However, our method excludes publications on AI that appear in clusters with a lower AI share. Our results may also include false positives, such as

publications that directly focus on drone piloting without reference to artificial intelligence. In general, publications in these clusters may represent progress in an AI field, the application of AI to a separate domain such as drone piloting, or progress in these AI application domains. We use the same filters as in previously-published work by CSET authors,[8] and identify 1,897 research clusters that meet the following criteria:

- For at least 50 percent of publications, we can predict their AI relevance, either from a SciBERT classifier trained on ArXiv data or from a Chinese-language keyword search.[9]
- Of those publications with an AI prediction, we predict at least 50 percent are AI publications.
- The cluster's publications as a whole have an average publication age of no more than 20 years. This filters out clusters with a large number of older documents, which are unlikely to be relevant for modern AI progress.

Identifying High-Impact Publications

While the 1,897 research clusters we identified as AI-relevant include more than 1.6 million AI publications, we are particularly interested in the high-impact publications from these clusters. Many definitions of "elite" or high-impact research have been used in the literature. This brief primarily focus on highly cited research, following the common bibliometric practice of using a publication's citation rank within a reference group as a proxy measure of impact. We group publications by research field (e.g., computer science, mathematics) to adjust for differing citation practices across fields, and further group them by the year of publication to account for the fact that older publications have had more time to accumulate citations.[10] We then rank papers by citation count for each publication year, taking their rank within these field-year groups as a measure of paper impact or quality.[11] For the purposes of this brief, we focus on research at the 95th citation percentile and above as "highly cited research." These publications received more citations than 95 percent of papers in the same field of study that were published in the same year. We identified 170,000

highly cited publications in our AI research clusters published between 2005 and 2019. Almost half (79,000 in total) were published between 2015 and 2019, reflecting the rapid growth of AI as a research field.[12]

Associating Papers with Countries

We use the metadata from our source datasets to associate publications with their authors' affiliated organizations at time of publication, and to associate those organizations to the country in which they are located. We restrict our analysis to publications with at least one country associated with them in the CSET merged corpus.[13] This brief counts all publications with at least one U.S. author organization as U.S. publications, and all publications with at least one Chinese author organization as Chinese publications. These designations are, therefore, nonexclusive.

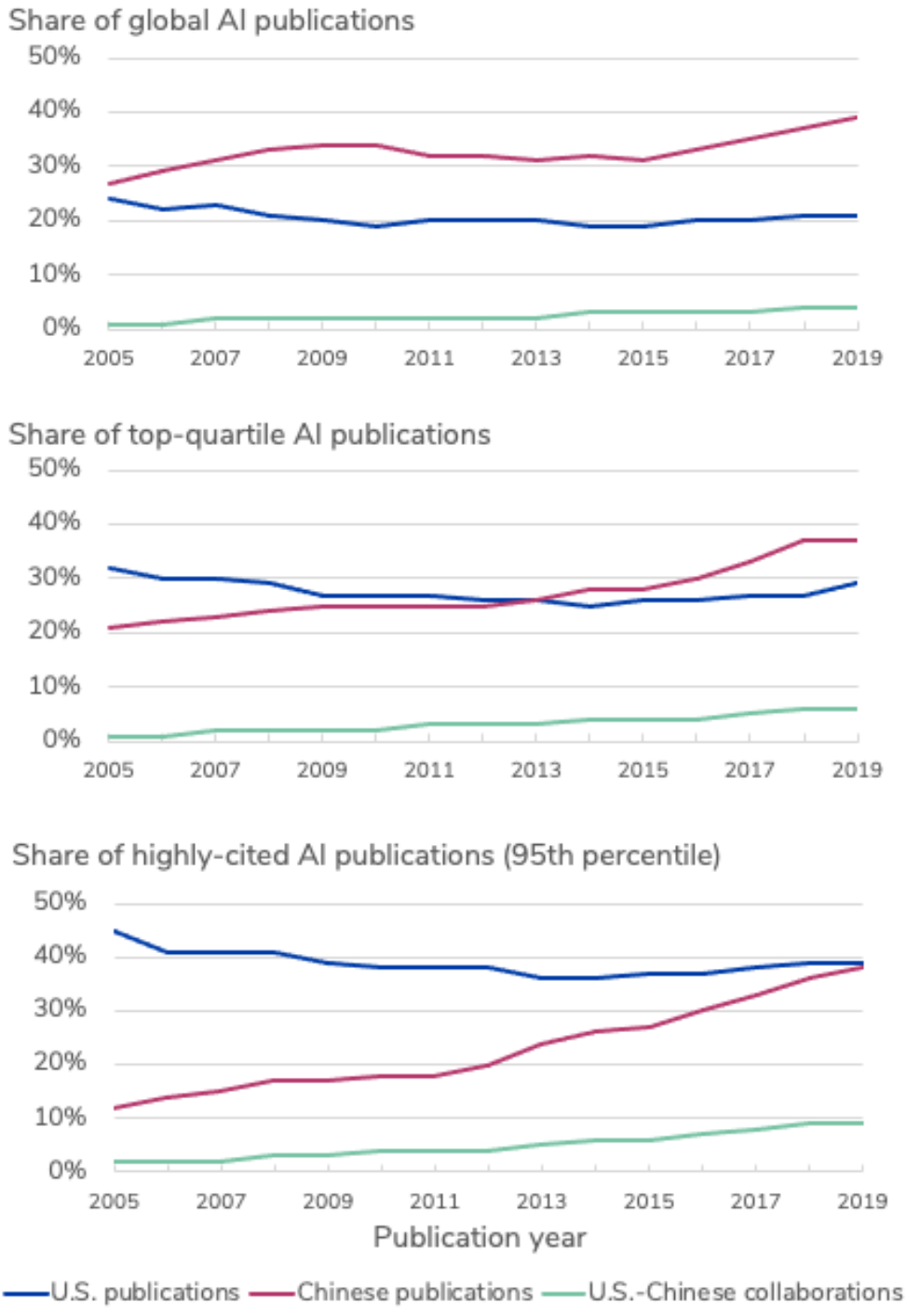
Findings

Comparing Highly Cited Research in AI Over Time

China's Publication of Highly Cited AI Research Is Growing to Rival The United States

The results presented in Figure 1 challenge the common perception that China is capable of publishing research in high quantity but not high quality. Not only has China's *quantity* of AI research grown with respect to the United States since 2010, its publication of high-*quality* research has steadily increased, nearly rivaling U.S. numbers as of 2019.

Figure 1. China has long exceeded the United States in total AI publications, and its share of more highly cited AI publications has steadily grown.[14]



Source: CSET merged corpus. Results generated November 2, 2021.

In 2005, China trailed the United States in highly cited AI research, but had already surpassed it in overall AI output. By 2013, China had reached parity with the United States in top-quartile research, but the United States still exceeded China's 95th-percentile AI output, with 3,600 highly cited publications to China's 2,300. Over the following years, however, China steadily closed the gap in highly cited research. By 2019, the two countries were at parity: 8,000 highly cited AI publications that year had U.S.-affiliated authors, while 7,900 had Chinese-affiliated authors. These totals correspond to 39 percent and 38 percent of the 20,800 highly cited AI publications published that year, respectively.

Chinese AI Papers Are Cited Internationally

The data indicates that Chinese publications receive a consistent share of their citations from international publications, and that the number of citations they receive from non-Chinese sources is growing rapidly over time.

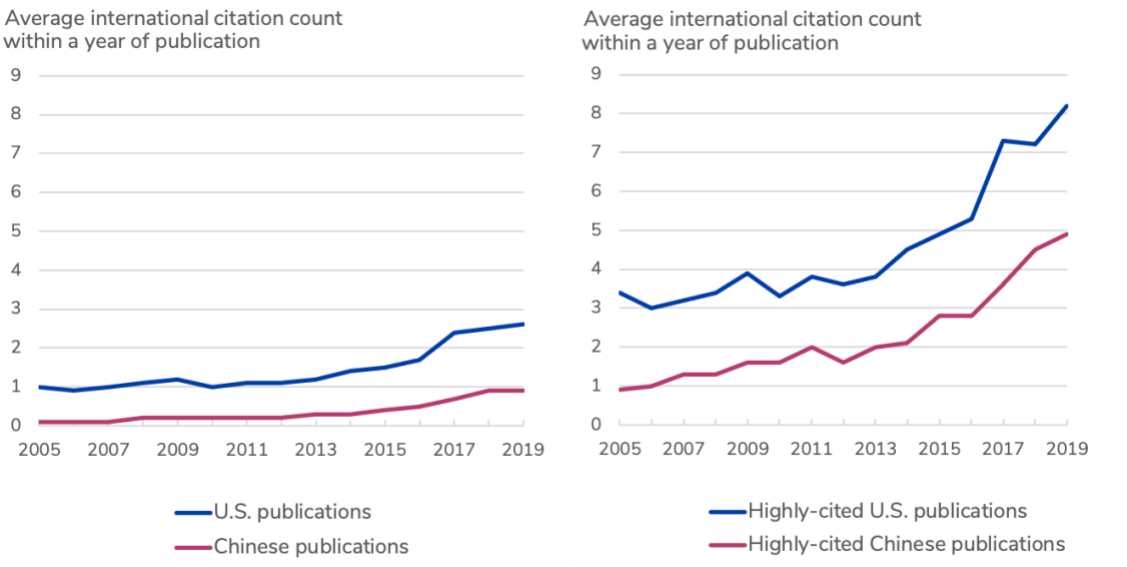
From 2015 to 2019, 35 percent of citations for Chinese AI publications, and 38 percent of citations for highly cited Chinese AI publications, came from non-Chinese publications. Both figures were stable over that time period.

If the rise in highly cited Chinese AI publications were due to a scaling-up of citations from Chinese researchers, unaccompanied by a rise in research quality, we would expect that the share of citations they accrued from international sources would drop over time, but this is not what we observe. The share of international citations received by Chinese publications is, however, notably lower than for U.S. publications, which received approximately 58 percent of their citations from non-U.S. publications in 2019.

To more directly compare the citations received by publications from the United States and China, we consider citations from a fixed pool of publications: international publications with neither U.S. nor Chinese involvement. We find that, in their first year after being published, new Chinese publications receive fewer citations than new U.S. publications from international publications. However, the number of international citations received by new AI

publications in both countries is rapidly growing, especially for the most highly cited papers.

Figure 2. International citations of Chinese and U.S. AI publications are growing rapidly over time.



Source: CSET merged corpus. Results generated September 30, 2021.

There are multiple factors that might cause Chinese publications to receive a smaller share of citations from international sources than U.S. publications, besides a difference in quality. China's AI research output is the largest in the world, and a combination of researcher and funder interests may lead to a large share of this research being focused on areas of AI that are of less interest to foreign researchers. U.S. publications are also more likely to be international collaborations, and we might expect U.S. authors to generally be more well-connected with their colleagues in other productive AI nations, which tend to be U.S. allies. Language barriers and differences in terminology may also prevent foreign researchers from finding or appreciating relevant Chinese publications.[15] As one might expect, collaborations between Chinese and non-Chinese authors are more likely to interest researchers outside of China. Such collaborations receive 43 percent of their citations from non-Chinese publications, while papers with only Chinese authors receive just 30 percent of their

citations from non-Chinese publications.[16] This discrepancy could be due to any number of the factors we have suggested; these papers may be higher quality, may cover topics of greater interest to non-Chinese researchers, or may simply rise to their attention more easily through academic networks.

Overall, these citation patterns indicate that Chinese AI publications receive significant and growing interest from international sources. However, Chinese publications, even highly cited ones, receive a notably smaller rate of international citations than U.S. publications. Our quantitative analysis does not allow us to distinguish between the multiple plausible causes of this gap, but it is likely due in part to the United States' greater degree of integration into the international research community.

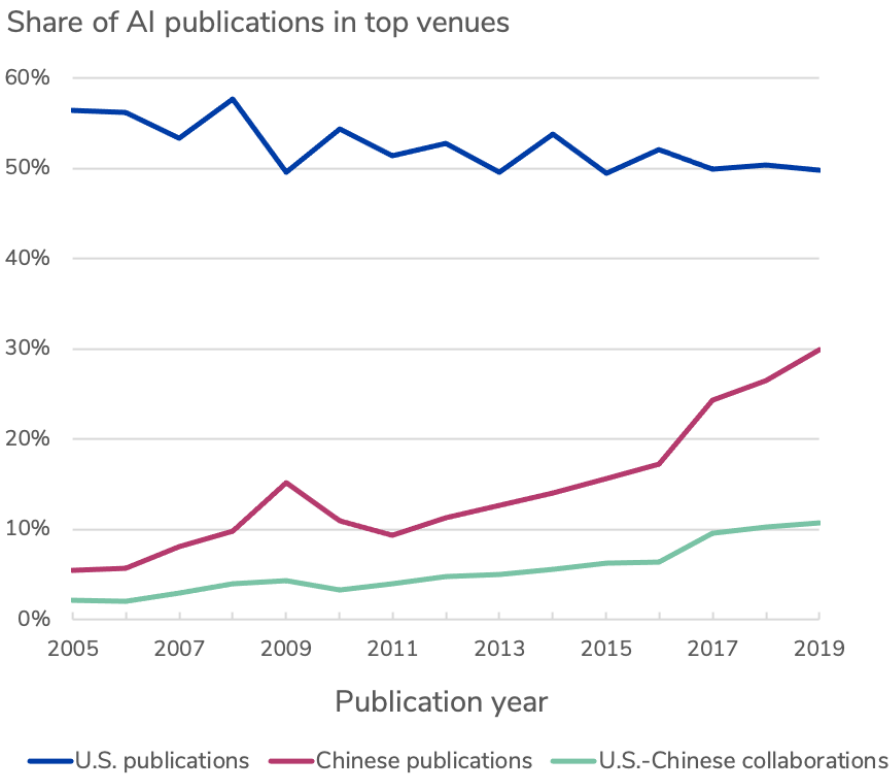
Publications in Top Venues

We find that Chinese research also accounts for a growing share of publications in highly regarded AI venues. CSrankings.org, a practitioner-created ranking of academic institutions by their publications in top computer science venues, lists 13 top conferences for artificial intelligence publications.[17] We identified 46,000 publications in our AI research clusters that were published in these conferences between 2015 and 2019. These top-venue publications are far more likely to be highly cited than the average AI publication: almost half (47 percent) were highly cited, compared to 13 percent of all publications in AI clusters. Highly cited publications in top venues make up a small but noteworthy share (20 percent) of all highly cited publications. (See Appendix B for more details.)

We find that both Chinese and U.S. publication counts in these venues have increased rapidly in recent years. While the United States still publishes significantly more articles in these venues than China in absolute terms, the ratio between the two countries has shrunk: China published 60 percent as many publications in these venues as the United States in 2019, compared to 32 percent in 2015 and 10 percent in 2005. China's share of world output in these venues is also growing, while the United States' share is stagnant; in 2019, 31 percent of publications in these

venues had at least one Chinese author, compared to 13 percent in 2010 and just 6 percent in 2005.

Figure 3. An increasing share of publications in the top 13 AI conferences have Chinese authors, while the U.S. share of these publications is stagnant.



Source: CSET merged corpus. Results generated September 30, 2021.

Similarly, CSRankings' lists of the world's top ten AI institutions has a significant and growing Chinese presence. The list of top AI institutions from 2011 to 2015 includes three institutions from China and six from the United States;[18] the list of top AI institutions from 2016 to 2020 includes four institutions from China and five from the United States.[19] These results indicate that a growing tranche of Chinese research attracts significant interest from the international research community.

Collaborations Between U.S. and Chinese Researchers

Collaborations between U.S. and Chinese researchers contribute a small but notable portion of both countries' AI publications. As we consider higher-impact subsets of research, these collaborations account for an increasing share of Chinese AI contributions.

Table 1. Collaborations between Chinese and U.S. researchers make up a notable share of both countries' high-impact AI publications.

Publication type	Share of U.S. publications from 2019 that were U.S.-Chinese collaborations	Share of Chinese publications from 2019 that were U.S.-Chinese collaborations
AI publications	19%	10%
Top-quartile AI publications	21%	16%
95th-percentile AI publications	24%	24%
99th-percentile AI publications	24%	30%
Top-venue AI publications	22%	36%

In 2019, only 10 percent of Chinese AI publications were collaborations with U.S. researchers, but these collaborations were more highly cited than other Chinese research, accounting for 24 percent of highly cited Chinese publications. This trend continues for smaller subsets of high-impact publications: U.S.-Chinese collaborations made up 30 percent of 99th-percentile Chinese publications and 36 percent of Chinese publications in top AI research venues. As noted above, these publications are not necessarily higher-quality than Chinese-only publications; they may also receive greater international recognition because they are more accessible to non-Chinese researchers.

U.S.-Chinese collaborations also accounted for a moderate portion of the United States' AI publications. Nineteen percent of U.S. AI publications from 2019 were U.S.-Chinese collaborations; the fraction rises to 22 percent for top-venue U.S. publications and 24 percent for highly cited U.S. publications.

Comparing Top AI Research Clusters

We explore research clusters in CSET's Map of Science that contain a high number of these highly cited and top-venue AI publications. Since these clusters are defined by citation links, they are not guaranteed to have a single topic. However, documents in a cluster often do share a fairly well-defined research area or problem, such as a particular machine learning algorithm and its common application areas.

Below, we present a small number of clusters with an unusually high share of highly cited or top-venue AI papers, offering an illustration of some of the research we chart in the previous section. Half of the 1,897 AI clusters in the Map of Science contain fewer than 10 highly cited publications, and three-quarters of these clusters contain fewer than 10 top-venue papers. By contrast, clusters with a large number of such papers are relatively rare: only 36 (2 percent) of AI clusters contained more than 300 highly cited papers, and only 90 (5 percent) contained at least 100 top-venue papers.[20] The 16 unique clusters we describe in this section all meet one or both of these criteria.

Top Research Clusters for Highly Cited and Top-Venue Chinese Publications

Table 2A presents the AI research clusters with the highest amount of highly cited Chinese AI publications from 2015-2019. Table 2B presents the clusters with the largest number of Chinese publications in top AI venues, filtered for clusters where China published more top-venue publications than the United States over this period.[21]

Table 2A. AI research clusters with the highest amount of highly cited Chinese AI research, 2015-2019.

Cluster ID	Description of cluster topic	Number of Chinese highly cited papers, 2015-2019	Number of U.S. highly cited AI papers, 2015-2019	All highly cited papers as a share of the cluster
5806	AI, international regulation, and legal liability	578	14	23%
3568	Using deep convolutional networks for fault diagnosis in industrial machinery	410	53	34%
187	Fuzzy logic methods for decision-making under uncertainty	355	29	27%
214	Object detection and image classification	323	160	13%
148	Recommender systems for e-commerce and streaming sites	319	240	21%

Source: CSET merged corpus. Results generated September 30, 2021.

Table 2B. AI research clusters with the highest amount of top-venue Chinese AI research, 2015-2019.

Cluster ID	Description of cluster topic	Number of Chinese top-venue papers, 2015-2019	Number of U.S. top-venue papers, 2015-2019	All top-venue publications as a share of the cluster
148	Recommender systems for e-commerce and streaming sites	229	190	14%
1436	Person re-identification: identifying multiple instances of the same individual in video feeds	191	98	16%
1989	Using deep learning to analyze graphs, a research area with applications in many areas including recommender systems and knowledge representation	172	168	26%
2891	Deep hashing, a method for organizing databases to support fast retrieval of files (especially image files)	153	109	16%
183	"Visual tracking": tracking objects in video feeds	118	69	6%

Source: CSET merged corpus. Results generated September 30, 2021.

This list presents a small snapshot of overall AI research, but it represents a view into some topics that are commonly addressed by highly cited and top-venue Chinese AI publications.[22]

We find that clusters with unusually high amounts of top Chinese publications cover many areas of computer vision. These areas include the surveillance-relevant task of person re-identification, or re-identifying the same person across multiple video feeds.[23] However, China also publishes work on many general-purpose computer vision tasks, where progress can improve the capabilities of both civilian and security-relevant applications. These tasks—including object detection and visual tracking—make up a large share of computer vision research.[24]

Other top Chinese clusters focus on the practical application of AI in industry, including questions of legal liability, applying AI to fault diagnosis in industrial machinery, and recommending content on e-commerce sites.

Top Research Clusters for Highly Cited and Top-Venue U.S. AI Publications

Table 3A presents the AI research clusters with the highest amount of highly cited U.S. AI publications from 2015 to 2019. Table 3B presents the clusters with the largest number of U.S. publications in top AI venues, filtered for clusters where the United States published more top-venue publications than China over this period.[25]

Table 3A. AI research clusters with the highest amount of highly cited U.S. AI research, 2015-2019.

Cluster ID	Description of cluster topic	Number of U.S. highly cited AI papers, 2015-2019	Number of Chinese highly cited AI papers, 2015-2019	All highly cited papers as a share of the cluster
1193	Natural language processing, especially with transformer architectures	402	103	45%
1609	Deep reinforcement learning	363	36	27%
4358	Societal impacts and fairness in AI deployment, with focuses on social networks and the criminal justice system	326	11	35%
2381	AI robustness against adversarial attacks	292	94	37%
1338	Generative adversarial networks for image and video synthesis	273	172	27%

Source: CSET merged corpus. Results generated September 30, 2021.

Table 3B. AI research clusters with the highest amount of top-venue U.S. AI research, 2015-2019.

Cluster ID	Description of cluster topic	Number of U.S. top-venue AI papers, 2015-2019	Number of Chinese top-venue papers	All top-venue publications as a share of the cluster
1193	Natural language processing, especially with transformer architectures	410	130	50%
1338	Generative adversarial networks for image and video synthesis	277	136	25%
3527	Machine translation with neural nets, particularly using transformer and recurrent neural net architectures	255	195	39%
1609	Deep reinforcement learning	244	60	21%
3446	Using point cloud representations to recognize and model 3D objects from 2D imagery	234	127	33%

Source: CSET merged corpus. Results generated September 30, 2021.

As with Table 2B, this list presents a small snapshot of overall AI research. However, it represents a view into some topics that are commonly addressed by highly cited U.S. AI publications.[26]

One such topic is the development of novel deep-learning architectures and training methods—including several areas that have seen significant progress in recent years. Deep reinforcement learning is used to train algorithms that make decisions in complex domains, such as playing Go or driving cars. Generative adversarial networks (GANs) are used to synthesize images and text, while transformers are a promising new text synthesis architecture, notably used in OpenAI's GPT-2 and GPT-3. The innovations involved in top U.S. clusters are generally used in the high-level AI areas of reinforcement learning and natural language processing, versus the focus on computer vision we see in China's top clusters.

Other top U.S. clusters discuss AI ethics and safety, including fairness in AI decision-making and improving the robustness of AI systems to adversarial attacks.

Widening the Aperture: The International Balance of Highly Cited AI research

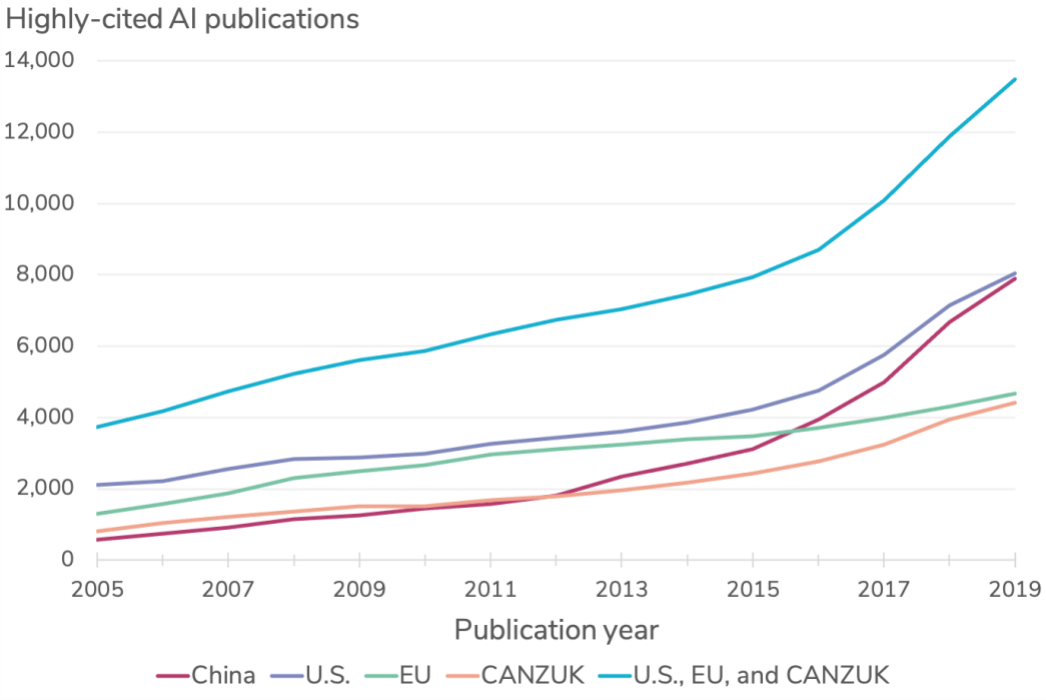
The United States and China Together Contribute Around 65 Percent of Highly Cited Research in AI Clusters[27]

AI competition is an international arena: the United States and China are the most prominent publishers of highly cited AI research, but are far from the only source of innovation. A more representative accounting of the balance of AI innovation would include other key sources of AI innovation, which tend to be U.S. allies.

American Allies, Particularly the EU 27 and CANZUK Countries, Also Publish a Significant Share of Highly Cited AI Research

Previous CSET research has found that Chinese and U.S. R&D spending is comparable in purchasing power parity terms, but the United States and its allies combined spend more than twice as much on R&D as China does.[28] Similarly, Figure 4 illustrates that China is rapidly approaching the highly cited AI research output of the United States, but the United States and its allies combined generate far more highly cited AI research than China alone. In 2019, for example, China published 7,900 highly cited publications in our AI research clusters, compared to 13,500 highly cited AI publications with at least one affiliation from the United States, the European Union, or Canada, Australia, New Zealand and the United Kingdom (CANZUK). Coordination and collaboration with allies will therefore be important to U.S. efforts to maintain an edge in AI research, and in technological competition more generally.

Figure 4. The United States, the EU, and CANZUK all publish significant amounts of highly cited AI research; taken together, they outpublish China by a factor of two. However, China's output is rapidly growing.[29]



Source: CSET merged corpus. Results generated September 30, 2021.

Conclusion

China has long been a major contributor of AI research by quantity. We find that it is increasingly competitive with the United States in high-quality output, as measured by both highly cited publications and publications in top venues. Each metric of research output and quality has limitations, but some general trends appear to hold both across our analyses and in related work.[30] For many years, Chinese researchers have published more publications in AI research clusters than their counterparts in the United States, and their output continues to grow in both quantity and quality. China is approaching the United States' share of the world's high-quality AI publications. The United States maintains a lead in top conference publications, and U.S. research receives significantly more international citations than Chinese work. A notable share of both countries' high-impact research comes from U.S.-Chinese collaborations, which account for 21-24 percent of high-impact U.S. publications and 24-36 percent of high-impact Chinese publications.

Bibliometric analyses give us valuable insight into AI research trends, but it is limited insight. Some research is not reflected in our results. Many of the most impactful advances in AI are being made in industry efforts, which are not fully reflected in the bibliometric data.[31] Our data includes those industry projects that are written up for publication in academic venues or as ArXiv preprints; analyses of patent data could complement our approach.[32] Bibliometric data also necessarily excludes advances that are not published in any form, such as classified government projects or industry trade secrets. In addition, AI publications are not interchangeable; further work could provide a deeper topic-level analysis of high-impact research. Our survey of research clusters in Tables 2 and 3 covers a relatively small portion of both countries' highly cited research, but suggests some points of interest. At a high level, these clusters' topics align with previous research showing that China is particularly prolific in computer vision, while the United States publishes a greater share of highly cited research in other sub-areas of AI such as natural language processing.[33]

Research alone does not guarantee competitiveness: to develop useful technologies, a country needs a healthy innovation ecosystem. Highly cited publications show that a country houses talented researchers, but this does not guarantee that their talents or insights will be applied to practical ends. Conversely, states with fewer top publications can still make use of the insights available in published research.[34] International collaborations, competitive tech companies, and inflows of talented immigrants can enable a state to develop valuable AI applications regardless of its publication output.

Indeed, while China is growing to rival or exceed the United States' publication counts in many areas of AI research, the United States has a valuable resource in its policy and research ties with other technologically sophisticated countries. Prior analyses have shown that the United States and its allies tend to publish significantly more international collaborations than China, and that China's rate of international collaborations is roughly flat while other countries' collaboration rate is growing.[35] We find that, in combination with its allies in the CANZUK and EU-27 nations, the United States far exceeds China's output of highly cited AI research.

Looking ahead, the United States and its allies will need to find a course that preserves the benefits of international collaborations and high-skilled immigration while guarding against adversarial exploitation of these practices. Done poorly, research security restrictions could severely hamper the United States and its allies. A thoughtful, field-specific approach is necessary to make effective use of the U.S. position in the international research ecosystem.[36]

Authors

Ashwin Acharya is a research analyst at CSET. Brian Dunn was a semester research analyst at CSET who contributed to our qualitative descriptions of AI research clusters.

Acknowledgements

For their feedback, we thank Catherine Aiken, Igor Mikolic-Torreira, Dewey Murdick, Will Hunt, Ilya Rahkovsky, and Helen Toner. We particularly thank Kuansan Wang and Field Cady for their comments as external reviewers, and Autumn Toney for her advice and assistance with data analysis. For editorial assistance, we thank Corey Cooper and Melissa Deng.

Document Identifier: doi: 10.51593/20210028

Appendix A: How Highly Cited and Top-Venue AI Papers Are Distributed Across Research Clusters

CSET's Map of Science contains 1,897 AI clusters. The 79,000 highly cited AI papers and 40,000 top-venue AI papers published between 2015 and 2019 are not evenly distributed across these clusters. Rather, most clusters contain few such papers, while a few clusters contain hundreds of them.[37]

Table A-1. Half of all AI research clusters contain fewer than 10 highly cited papers; only 2 percent contain 300 or more such papers.

Publication type	Number (%) of AI clusters with...				
	0 such papers	1-9 such papers	10-99 such papers	100-299 such papers	300 or more such papers
Highly cited AI publications	289 (15%)	621 (33%)	784 (41%)	167 (9%)	36 (2%)
Chinese highly cited AI publications	714 (38%)	711 (37%)	420 (22%)	46 (2%)	6 (0%)
U.S. highly cited AI publications	621 (33%)	693 (37%)	532 (28%)	48 (3%)	3 (0%)

Source: CSET merged corpus. Results generated September 30, 2021.

Table A-2. Almost three-quarters of all AI clusters contain fewer than 10 top-venue publications; only 5 percent contain 100 or more such publications.

Publication type	Number (%) of AI clusters with...				
	0 such papers	1-9 such papers	10-99 such papers	100-299 such papers	300 or more such papers
Top-venue AI publications	706 (37%)	681 (36%)	420 (22%)	78 (4%)	12 (1%)
Chinese top-venue AI publications	1194 (63%)	498 (26%)	186 (10%)	19 (1%)	0 (0%)
U.S. top-venue AI publications	993 (52%)	557 (29%)	310 (16%)	36 (2%)	1 (0%)

Source: CSET merged corpus. Results generated September 30, 2021.

Nonetheless, because highly cited and top-venue publications are spread out across many research clusters, the top few clusters contain only a small portion of all high-impact publications. For example, the five clusters in Table 2A contain only 8 percent of highly cited Chinese AI publications. Collectively, the 16 unique clusters in Tables 2 and 3 contain 15 percent of such publications. However, as noted in Tables 2 and 3, highly cited and top-venue papers make up an unusually large share of these clusters.

Table A-3. The individual lists of top clusters we present in Tables 2A, 2B, 3A, and 3B represent less than 10 percent of the relevant publications. Collectively, the 16 unique clusters we present contain 11 to 16 percent of top U.S. and Chinese publications, and 10 percent of top AI publications overall.

Publication type	Share of publications in this set present in the top 5 relevant clusters (e.g. share of highly cited Chinese publications represented in Table 2A)	Share of publications in this set present in the 16 unique clusters listed in Tables 2 and 3
Chinese highly cited publications	8%	15%
Chinese top-venue publications	8%	16%
U.S. highly cited publications	6%	11%
U.S. top-venue publications	6%	11%
All highly cited publications in AI clusters	N/A	10%
All top-venue publications	N/A	10%

Source: CSET merged corpus. Results generated September 30, 2021.

Appendix B: Overlap Between Highly Cited and Top-Venue AI Papers

We identify 79,000 highly cited and 46,000 top-venue AI publications published between 2015 and 2019. Most of our top-venue publications (34,000, or 74 percent) belong to the AI research clusters that we used to identify highly cited AI publications. Within these AI clusters, 16,000 publications are both highly cited and published in top venues; this means that 47 percent of top-venue publications in AI clusters are highly cited, much higher than the baseline rate of 13 percent of all AI publications.[38] Conversely, 20 percent of highly cited AI publications were published in top venues, compared to 8 percent of AI publications overall.

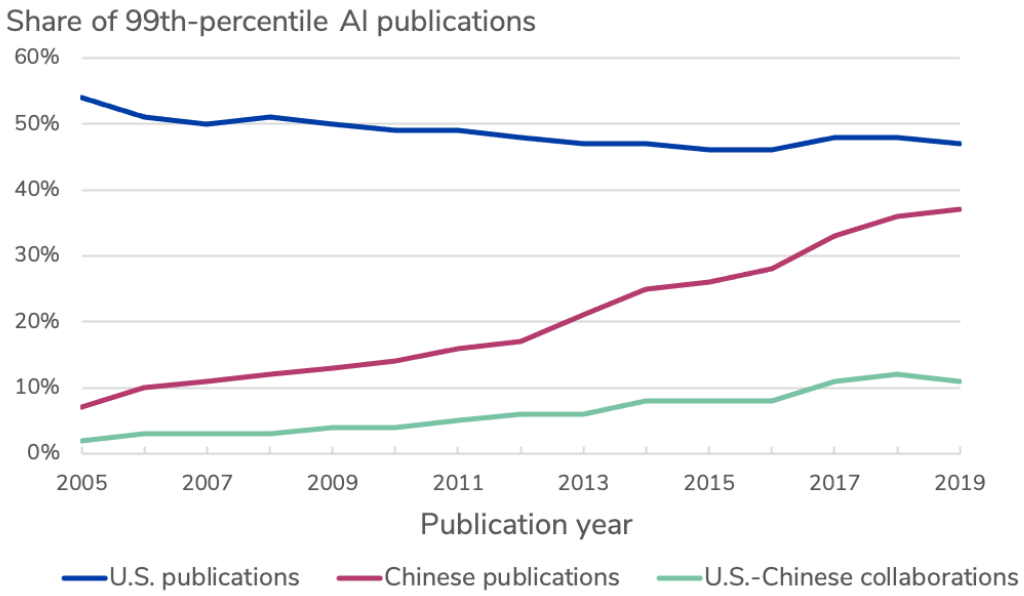
These ratios are similar when broken down by country: 56 percent of Chinese top-venue publications were highly cited, and 19 percent of Chinese highly cited publications were published in top venues. Similarly, 54 percent of U.S. top-venue publications were highly cited, and 31 percent of U.S. highly cited publications were published in top venues. The larger share of U.S. highly cited publications that were published in top venues may reflect the trend that U.S. publications are disproportionately likely to be published in top venues; the U.S. share of top-venue papers is higher than the country's share of highly cited AI publications or of AI publications overall.

Appendix C: Distribution of 99th-Percentile AI Publications

Readers may be curious whether the trends we observe still hold when we consider a more restrictive set of highly cited publications. To explore this question, we re-ran our data analyses for AI publications in the top percentile of their research field.[39] These results are similar to our findings for 95th-percentile publications above.

We noted in Figure 1 that China already exceeded the United States' AI publication count in 2005, reached parity with the United States in top-quartile publications by 2010, and is now reaching parity in 95th-percentile publications. In other words, we see two trends: China is less competitive with the United States in more highly cited subsets of research, but over time it is becoming more competitive in all areas.[40] Here, we see that this somewhat nuanced observation extends to 99th-percentile AI publications: China has not yet reached parity in this area, but its share of these publications is approaching the U.S. share.

Figure C-1. China publishes fewer 99th-percentile AI publications than the United States, but the ratio of U.S. to Chinese publications is decreasing over time. U.S.-Chinese collaborations accounted for almost a third of China's 99th-percentile AI publications in 2019.



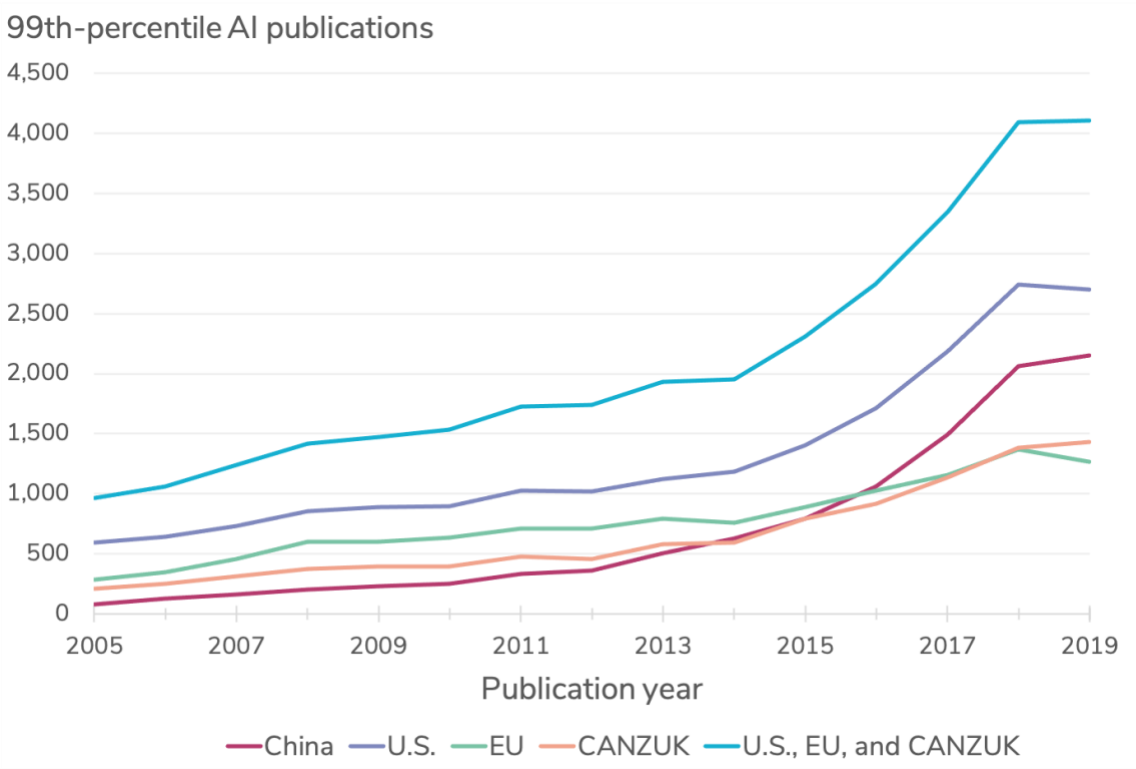
Source: CSET merged corpus. Results generated November 3, 2021.

We also see that U.S.-Chinese collaborations continue to make up an increasing share of Chinese publications as we filter for more highly cited work. In 2019, these publications made up 10 percent of all Chinese AI publications, 24 percent of 95th-percentile Chinese publications, and 30 percent of 99th-percentile Chinese publications.

In Figure C-2, we see that the United States, European Union, and CANZUK combined publish twice as many 99th-percentile publications as China; Figure 4 demonstrates roughly the same ratio for 95th-percentile publications.

Unlike in our other analyses, we observe what may be a trend change in 2019, with much lower growth than in previous years.[41] Similar analyses in the coming years will help us determine whether 2019 was an aberration or the beginning of a slowdown in 99th-percentile AI research.

Figure C-2. The United States, the European Union, and CANZUK all publish significant amounts of 99th-percentile AI research; taken together, they outpublish China by a factor of two.



Source: CSET merged corpus. Results generated November 3, 2021.

Endnotes

[1] For example, a McKinsey report claims that "China lags behind the United States and the United Kingdom in terms of fundamental research that advances the field of AI." Dominic Barton, Jonathan Woetzel, Jeongmin Seong, and Qinzheng Tian, "Artificial Intelligence: Implications for China" (McKinsey & Company, April 2017), https://www.mckinsey.com/~/media/mckinsey/featured percent20insights/China/Artificial percent20intelligence percent20Implications percent20for percent20China/MGI-Artificial-intelligence-implications-for-China.ashx.

[2] For example, several Chinese researchers recently wrote that "although aggregate AI research outputs (e.g., scientific publications, patents) are rising rapidly in China, truly original ideas and breakthrough technologies are lacking." Daitian Li, Tony W. Wong, and Yangao Xiao, "Is China Emerging as the Global Leader in AI?," Harvard Business Review, February 18, 2021, https://hbr.org/2021/02/is-china-emerging-as-the-global-leader-in-ai.

[3] Jiangjiang Yang and Oren Etzioni, "China is closing in on the US in AI research," Allen Institute for AI (Medium), May 11, 2021, https://medium.com/ai2-blog/china-is-closing-in-on-the-us-in-ai-research-ea5213ae80df; Dewey Murdick, James Dunham, and Jennifer Melot, "AI Definitions Affect Policymaking" (Center for Security and Emerging Technology, June 2020), https://cset.georgetown.edu/wp-content/uploads/CSET-AI-Definitions-Affect-Policymaking.pdf.

[4] Robert D. Atkinson and Caleb Foote, "Is China Catching Up to the United States in Innovation?" (Information Technology & Innovation Foundation, April 2019), https://projects.iq.harvard.edu/files/innovation/files/2019-china-catching-up-innovation.pdf.

[5] Our analysis is not limited to publications in academic journals and conferences; it also includes preprints on the ArXiv repository, which private AI labs often use to report their latest innovations. For example, most of the publications linked on OpenAI's publications page are ArXiv preprints. Such preprints can still appear in our subset of highly cited AI publications. For example, OpenAI's ArXiv preprint "Deep Double Descent: Where Bigger Models and More Data Hurt" appears in the CSET merged corpus as one of the most highly cited AI publications of 2019, placing in the highest percentile for computer science publications in that year. Preetum Nakkiran, "Deep Double Descent: Where Bigger Models and More Data Hurt," arXiv preprint arXiv:1912.02292 (2019), https://arxiv.org/abs/1912.02292.

[6] We refer to the Five Eyes countries, excluding the United States, as CANZUK. This group includes Canada, the United Kingdom, Australia, and New Zealand.

In this brief, European Union refers to the 27 member states of the after the departure of the United Kingdom: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, the Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, and Sweden.

[7] CSET merged corpus of scholarly literature includes Digital Science's Dimensions, Clarivate's Web of Science, Microsoft Academic Graph, China National Knowledge Infrastructure, arXiv, and Papers With Code. China National Knowledge Infrastructure is furnished for use in the United States by East View Information Services, Minneapolis, MN, USA.

[8] Ilya Rahkovsky et al., "AI Research Funding Portfolios and Extreme Growth," Frontiers in Research Metrics and Analytics,, April 6, 2021, https://www.frontiersin.org/articles/10.3389/frma.2021.630124/full.

[9] Our SciBERT classifier is described in 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; our Chinese-language keyword search follows the same approach as Rakhovsky et al., "AI Research Funding Portfolios and Extreme Growth."

[10] We classify papers into top-level research fields (e.g. computer science, mathematics) using the Microsoft Academic Graph (MAG) field of study taxonomy. (See Zhihong Shen, Hao Ma, and Kuansan Wang, "A Web-scale system for scientific knowledge exploration," arXiv preprint arXiv:1805.12216 (2018), https://arxiv.org/pdf/1805.12216.pdf). We use a natural language model to estimate field scores for each English-language publication in the CSET merged corpus. This model is highly accurate in replicating MAG's scores for papers in the MAG database, and allows us to extend our tagging to English-language publications not in the MAG dataset. We also impute field scores for publications in other languages (mostly Chinese) by taking an average over the scores of their neighbors in the citation graph.

By far the most common field of study for AI publications, highly cited or otherwise, was computer science.

[11] We limit our data to papers from 2019 and prior years, since papers from 2020 have not yet had time to accumulate many citations and their citation rank is therefore likely to be heavily impacted by small number bias.

[12] Since AI is a particularly fast-growing and impactful research area, a disproportionate share of AI publications are in the top 5 percentiles of their broader research field (e.g. computer science or mathematics). Between 2015 and 2019, 13 percent of AI publications were in the 95th percentile of their

research field, making them "highly cited" as judged by our metric. For the full 2005-2019 period, 11 percent of AI publications were in the 95th percentile of their research field.

[13] Publications with an associated country account for 77 percent of AI publications in recent years (2015-2019), as well as 92 percent of recent highly cited AI publications and 98 percent of recent top-venue AI publications.

[14] "Total AI publications" refers to all AI publications in our dataset with at least one associated country. Throughout this brief, the reported figures for each country include international collaborations featuring researchers from that country.

[15] For example, we find that publications that only appear in the Chinese National Knowledge Infrastructure, which tend to be Chinese-language, receive more than 99 percent of their citations from other Chinese-origin publications.

[16] See also Autumn Toney and Melissa Flagg, "Research Impact, Research Output, and the Role of International Collaboration," (Center for Security and Emerging Technology, November 2021). Here the authors find that excluding international collaborations causes a significant decrease in China's output of highly cited publications.

[17] In the CSRankings taxonomy, the high-level category of AI includes the subcategories of artificial intelligence, computer vision, machine learning and data mining, natural language processing, and "the Web & information retrieval." We include all venues listed under the high-level AI category in our analysis in order to capture top publications from all of these subcategories of artificial intelligence. For future work, we could consult AI domain experts to identify other venues to include. For example, the International Conference on Learning Representations (ICLR) is often mentioned as a top AI conference but is absent from the CSRankings list. (See, e.g., this list of top AI conferences from an AI consulting firm: https://www.am.ai/en/blog/ai-conferences-2021/).

[18] "CSRankings AI Rankings, 2011-2015," CSRankings, accessed September 21, 2021, http://csrankings.org/#/fromyear/2011/toyear/2015/index?ai&vision&mlmining&nlp&ir&world. Permanent link at https://perma.cc/AXS2-26RG; "CSRankings AI Rankings, 2016-2020," CSRankings, accessed September 21, 2021, http://csrankings.org/#/fromyear/2016/toyear/2020/index?ai&vision&mlmining&nlp&ir&world. Permanent link at https://perma.cc/B23T-B3VW.

[19] A recent report similarly finds that both the quantity and quality of doctoral graduates in China is growing over time. See Remco Zwetsloot et al., "China Is Fast Outpacing U.S. STEM Ph.D. Growth" (Center for Security and Emerging

Technology, August 2021), https://cset.georgetown.edu/publication/china-is-fast-outpacing-u-s-stem-phd-growth/.

[20] See Appendix A for more details.

[21] Two clusters led by the United States would otherwise have appeared in Table 2B: clusters 1338 and 981. For the top Chinese clusters listed in Table 2A, China published more highly cited research than the United States over this period.

[22] The clusters listed in Table 2A contain 8 percent of highly cited Chinese AI publications; the clusters in Table 3B contain 8 percent of top-venue Chinese AI publications. Overall, top Chinese publications are spread across many research clusters, but the clusters we present here are among the small fraction of clusters that contain hundreds of these publications. By contrast, most AI research clusters have zero top-venue Chinese publications and fewer than ten highly cited Chinese publications: see Appendix A for details.

[23] A recent report finds that China is by far the largest contributor to several computer-vision surveillance tasks, and publishes a majority of research on person re-identification. See Ashwin Acharya, Max Langenkamp, and James Dunham, "Trends in AI Research for the Visual Surveillance of Populations" (Center for Security and Emerging Technology, December 2021).

[24] See Acharya, Langenkamp, and Dunham, "Trends in AI Research for the Visual Surveillance of Populations."

[25] One China-led cluster would otherwise have appeared in Table 3B: cluster 148. For all top U.S. clusters listed in Table 3A, the United States published more highly cited research than China over this period.

[26] The clusters listed in Table 3A contain 6 percent of highly cited U.S. AI publications; the clusters in Table 3B contain 6 percent of top-venue U.S. AI publications. Overall, top U.S. publications are spread across many research clusters, but the clusters we present here are among the small fraction of clusters that contain hundreds of these publications. By contrast, most AI research clusters have zero top-venue U.S. publications and fewer than ten highly cited U.S. publications: see Appendix A for details.

Note that the clusters presented in Tables 3A and 3B overlap more than the Chinese-led clusters in Tables 2A and 2B, reflecting the fact that U.S. AI publications have a stronger correlation between citation rate and top-venue status than Chinese publications. (See Appendix B.)

[27] In 2019, the two countries published a total of 14,031 highly cited AI publications, including 1,896 U.S.-Chinese collaborations. These accounted for 67 percent of the world total, meaning that 33 percent of world highly cited AI papers had no U.S. or Chinese involvement. Further, a large number of U.S. and Chinese highly cited papers also involved international collaboration with third parties, so this 67 percent of world output was not due to U.S. and Chinese research efforts alone.

[28] Melissa Flagg, "Global R&D and a New Era of Alliances" (Center for Security and Emerging Technology, June 2020), https://cset.georgetown.edu/publication/global-rd-and-a-new-era-of-alliances/.

Purchasing power parity adjusts for the lower cost of a standardized basket of goods in China compared to the United States. As such, it assigns greater value to the Chinese yuan than would be indicated by the yuan-dollar exchange rate.

[29] Note: we count a publication as EU if it is associated with any EU country; the same is true of CANZUK. We do not double-count publications with multiple EU or CANZUK countries: a publication authored by German, French, Canadian, and Australian researchers would only be counted once for the EU and once for CANZUK. Similarly, when we count U.S. and allied publications combined, a U.S., German, and UK collaboration is only counted as a single allied publication.

[30] For related work that reaches similar conclusions, see Murdick, Dunham, and Melot ("AI Definitions Affect Policymaking") and Yang and Etzioni (Allen Institute: "China is closing in on the US in AI research").

[31] For example, impactful AI architectures like the Residual Network and the Transformer were first developed at Microsoft and Google respectively. See Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun on the ResNet ("Deep Residual Learning for Image Recognition," arXiv preprint arXiv:1512.03385 (2015), https://arxiv.org/pdf/1512.03385.pdf) and Ashish Vaswani et al. on the Transformer ("Attention Is All You Need," arXiv preprint arXiv:1706.03762 (2017), https://arxiv.org/pdf/1706.03762.pdf).

[32] See, e.g., Dewey Murdick and Patrick Thomas, "Patents and Artificial Intelligence: A Primer" (Center for Security and Emerging Technology, September 2020), https://cset.georgetown.edu/publication/patents-and-artificial-intelligence/.

[33] See, e.g. Figure 1 in Murdick, Dunham, and Melot, "AI Definitions Affect Policymaking."

[34] For one discussion of the possibility of AI diffusion, see Michael C. Horowitz, "AI and the Diffusion of Global Power," in "Modern Conflict and Artificial

Intelligence" (Centre for International Governance Innovation, 2020), 32, https://www.cigionline.org/sites/default/files/documents/Modern%20Conflict%20and%20AI_web.pdf.

[35] See Autumn Toney and Melissa Flagg, "Comparing the United States' and China's Leading Roles in the Landscape of Science" (Center for Security and Emerging Technology, June 2021), https://cset.georgetown.edu/wp-content/uploads/CSET-Comparing-the-United-States-and-Chinas-Leading-Roles-in-the-Landscape-of-Science-1.pdf.

[36] Melissa Flagg, Autumn Toney, and Paul Harris, "Research Security, Collaboration, and the Changing Map of Global R&D" (Center for Security and Emerging Technology, June 2021), https://cset.georgetown.edu/wp-content/uploads/CSET-Research-Security-Collaboration-and-the-Changing-Map-of-Global-RD.pdf.

[37] In addition, not all top-venue papers appear in AI clusters; between 2015 and 2019, 26 percent of these publications (10,700) appeared in research clusters that do not match our working definition of AI clusters. In Figure 3, we present counts of all top-venue publications, but in our data exploration we found that restricting to top-venue publications in AI research clusters did not significantly alter our findings.

[38] As noted above, AI publications are unusually well-cited; roughly 13 percent of recent AI publications are at or above the 95th percentile for citations within their research field.

[39] Since AI is a particularly fast-growing and impactful research area, a disproportionate share of AI publications are in the top percentile of their broader research field (e.g. computer science or mathematics). Between 2015 and 2019, 4 percent of AI publications were in the 99th percentile of their research field by citation count, and were thus included in our analyses for Appendix C.

[40] These results align with a recent report from the Allen Institute, which finds that China produces an increasing share of the top-1 of percent AI publications. The authors project that China will reach parity with the United States in 2025. See Yang and Etzioni, "China is closing in on the U.S. in AI research."

[41] Another noteworthy difference is that the absolute gap between Chinese and U.S. publications is not clearly significant over time.