Comparison of AI research topic in China and in the West since 2010

In short: I have compiled information about the topic of research as presented in 7 Western AI journals and 7 Chinese ones, between 2010 and 2022. This analysis covered over 50,000 published research papers. My dataset can be found here, along with the code used to collect the data and produce the graphs. Note: some punctual modifications were performed for some issues that did not follow exactly the structure of their journal, and any code aimed at handling all the exceptions would have been unwieldy.

I chose to focus on China's progress in artificial intelligence (AI) research, as news often presents it in menacing terms, and I thought getting actual data on this topic could clarify the matter. Headlines in western media reports frequently reference China's progress in the field of AI.[1] Such reports tend to highlight the power such tools can use to control and repress the population. As a centralised, authoritarian state with 1.4 billion inhabitants and a research budget of 2,7864 trillion yuan (or 409 billion euros)[2], it is clear that China has the data and resources to achieve very advanced goals in AI. Indeed, China's share of research papers in the field of AI went from 4.26% in 1997 to 27.68% of all AI research papers published in 2017, more than any other country. Similarly, the number of AI firms in China reached 1,189 in 2019.[3] For this reason, I wanted to compare the topics of AI research published in the top Western science journals and the leading Chinese research journals.

Although this is only preliminary research, the results are already informative, as we can see that Chinese journals tend to publish on the identified topics with very similar proportions as the West, except for some interesting wide differences which could indicate a different use of AI in Chinese society, companies and institutions.

I will first present the method, then the dataset I obtained through them as well as their limits, then I will present the graphs and some conclusions I gathered from the study.

Methods

To select the journals, I relied on the Scimago Journal and country rank database. This process immediately highlighted a limitation of my method: the first nine top journals are Western: American (IEEE, Microtome Publishing…) or Dutch (Elsevier, Springer…), with an h-index of 190 for the ninth one, whereas none of the nine Chinese AI journals (which are all of them according to Scimago) go higher than 45. The discrepancy between the amount of research on AI in China (and particularly the number of articles published in the country on that topic) and the small number of Chinese scientific journals about AI might be explained by the fact that Chinese researchers publish predominantly in Western

[1] Dennis Nguyen and Erik Hekman, "A 'New Arms Race'? Framing China and the U.S.A. in A.I. News Reporting: A Comparative Analysis of the Washington Post and South China Morning Post," *Global Media and China* 7, no. 1 (March 1, 2022): 58–77, https://doi.org/10.1177/20594364221078626.

[2] National Bureau of Statistics of China, "China's R&D Expenditure Reached 2.79 Trillion Yuan in 2021," National Bureau of Statistics of China, January 27, 2022, http://www.stats.gov.cn/english/PressRelease/202201/t20220127_1827065.html.

[3] Daitian Li, Tony W. Tong, 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.

journals.[4] However, the Chinese journals appear to publish mainly researchers affiliated with a Chinese university. So even though an analysis focusing on the university of affiliation of the authors would be more relevant, a comparison between the topics on which Western and Chinese journals in AI choose to focus can show us how China differs from the global average.

Among the 18 journals selected, 4 were removed (2 Western and 2 Chinese) from the analysis either because their articles did not include keywords, or they were not easily accessible through web-scraping. Finally, another Chinese journal had to be removed from the study as their servers made it difficult to access the papers fast enough.

The datasets

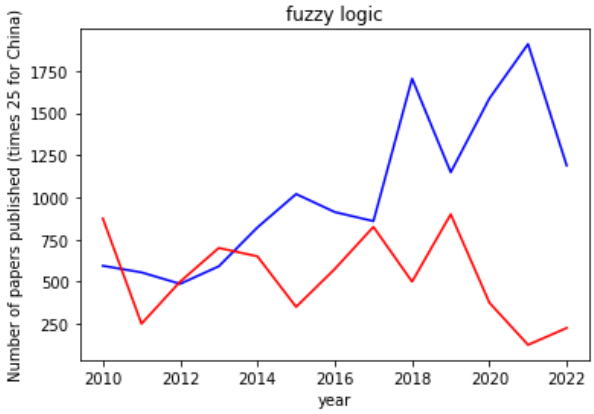
The dataset covers all publications from the selected journals from 2010, to cover the current "AI Spring"[5] (although some journals are younger than that, itself a sign of the flourishing of AI research).

All the journals focus on AI, although some concentrate on subfields, which could increase the proportion of keywords from that topic in the dataset. Nevertheless, I do not consider this bias problematic, since the fact that a journal focused on a topic rose to prominence among more generalist publications is itself a sign of the topic's popularity.

My programme cannot associate keywords related to similar subfields of AI. The use of natural language processing tools could significantly improve the accuracy of this picture, but even without utilising such tools, the most generic concepts tend to be mentioned in large numbers of articles and thus represent reality on the main topics of research over the past twelve years (of course, this also means that over general keywords such as "algorithms" do not illustrate any relevant trend in AI, and were thus ignored). In the following graphs, I have endeavoured to gather many keywords related to the theme, but this is only through looking for a common radical among the keywords. As an illustration, article on fuzzy logic often came with keywords including the word "fuzzy"; 5426 such keywords have been identified in the dataset, making it a good example. Unfortunately, since the program only sums the mentions of individual keywords, it is likely that some papers have been counted several times. But since this bias is present for all years, this should not distort any temporal trend. In all the following graphs the blue line represents the West, and the red one China, multiplied by the ratio of papers in the West dataset by the number of words in the Chinese dataset, which happens to be a nice round 25!

[4] 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/.

[5] James Manyika and Jacques Bughin, "The Coming of AI Spring | McKinsey," McKinsey Global Institute, October 14, 2019, https://www.mckinsey.com/mgi/overview/in-the-news/the-coming-of-ai-spring.



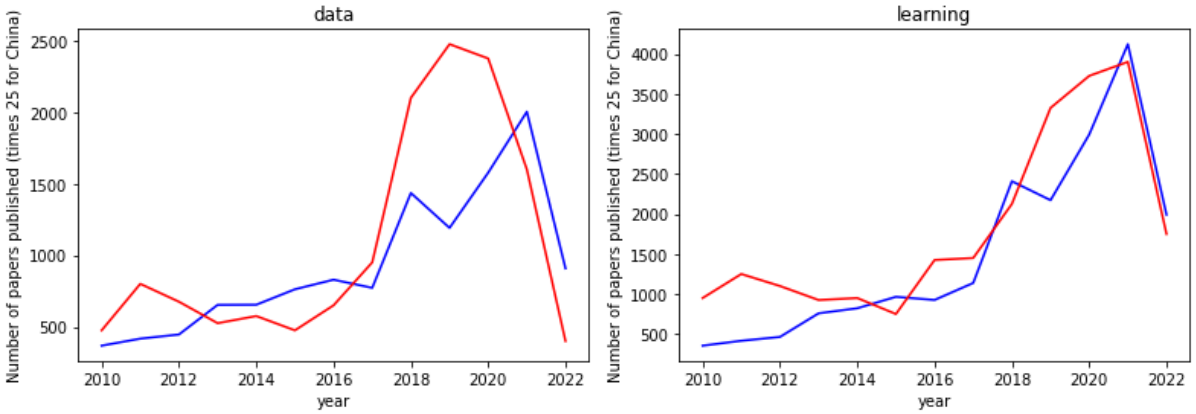
Because AI is such a strategic topic for a state or a company, some organisations might not focus as much on publishing as a university would. As a result, some current works that would be considered notable in a few decades are not included here.

Insights

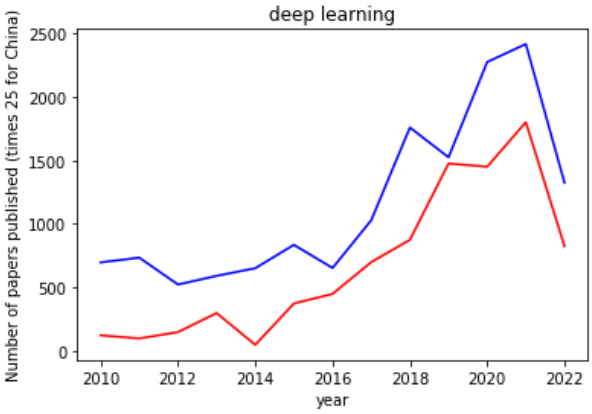
Even though the dataset only includes 14 of the 253 journals mentioned on Scimago, this sample illustrates the tremendous growth in AI research over the past decade.

The relatively low number of Chinese AI journals compared to the number of Chinese AI researchers could be explained by the prominence of English as the de facto language of research. Thus, it prevails over Chinese as a publishing language. Moreover, the number of top journals published in China indicates that using soft power and incentivising researchers to publish in Chinese language journals would come with a reduction in potential international influence.

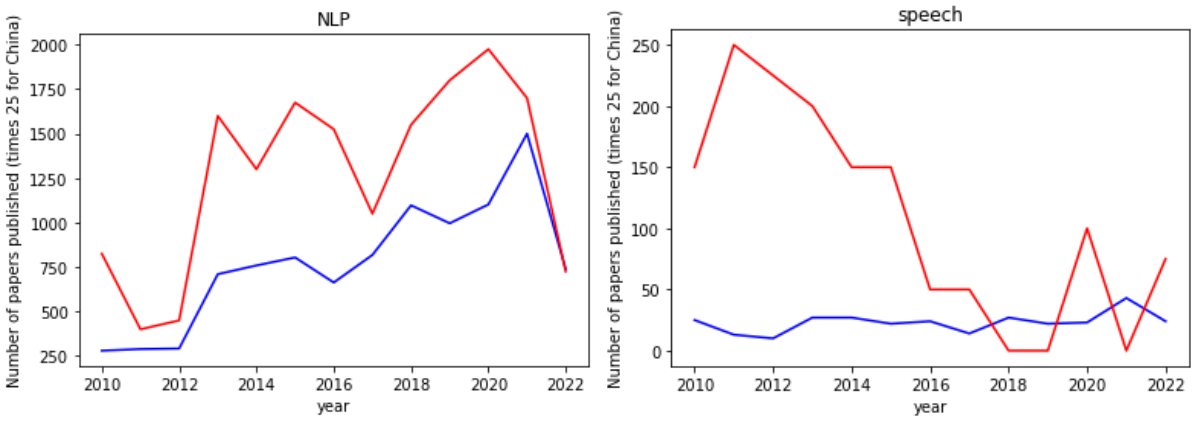
In several topics such as data analysis or learning, it appears that China follows the same rate of publication as the rest of the world.



Mentions of deep learning (and related topics, such as recurrent and convolutional neural networks) have been increasing in China and the West ever since 2010, and the focus on such methods has yet to slow down. The dip at the end is due to the fact few of the papers submitted in 2022 have been published yet; the full data for the current year will not be available until late 2023. Not knowing how long the peer-review process is for each individual journal, it is difficult to estimate which months of year are included for 2022.

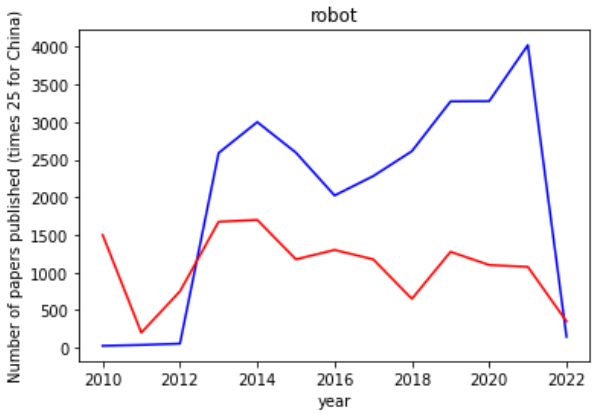


In language-related domains we can see that China publishes far more than we would expect considering the amount of papers published in other subfields of AI. This might be because of the specific needs of the country: the West is not going to show as much interest in the problem of ideogram recognition, or to understanding the meaning of a sentence that relies on Chinese grammar.



Finally, regarding topics with a potentially more political focus, we can see some difference in the way China and the West focused on computer vision over time, China having started considering the topic long before a vision boom occurred in the West. Of course, the possible applications of computer visions are varied, and a finer distinction of the keywords would be required to determine the meaning of those data.

I also considered robotics, wondering whether the different histories of industrialisation of the West and China would have an impact there. Here too, a more accurate dataset would be required to draw definitive conclusion for specific subfields of robotics, but it is clear that this area of AI is still mostly published in Western publications:



China shows a strong focus on decision-making systems however, and here too, a finer study would be required to make hypotheses on the applications of such systems:

This project highlighted the potential of journal databases as a research topic themselves. Through a more detailed analysis, such a framework could for instance be used to decide which topic to research, and identify “overhyped” topics. Unfortunately, the lack of standardisation between journals, even between different articles in the same publication, limits such studies. The vast majority of keywords were used only in one paper each. Moreover, a coherent system of keywords, directly accessible from journal databases, would help researchers find relevant articles and give their publications a wider reach. Thus, journals should help researchers select useful keywords, so their article could be grouped with related research projects. IEEE assigns its own keywords to the papers submitted to their conferences for instance, and that is the reason why the keyword “deep learning (artificial intelligence) ” goes from 0 in 2020 to 197 in 2021: no researcher had phrased it that way before (preferring instead such terms as “deep learning”)

Inversely, some concepts can be ambiguous, and it is difficult outside of the context of the article to understand which field they correspond to. For instance, is pattern classification related to image classification, or were those articles about more general algorithms, or even algorithms meant to classify pattern in a different format altogether?

Words can be grouped in different ways, depending on the focus of the question. We could for instance choose to focus on convolutional neural nets, deep learning, neural networks, or learning, and each category would include the one before.

Mentions of domain specific applications included face recognition, medical image processing and games. The data set included one journal focused on the application to a given field (Artificial Intelligence in Agriculture), contrary to journals focused on specific problems but with no particular requirements on a given application, such as data mining or pattern recognition. Other examples of AI journals mentioned on Scimago that have a focus on a field of applications include design, medicine/healthcare, manufacturing, law, and of course robotics.

Open questions

As mentioned before, such an analysis would be even more useful if it considered the affiliation of the researchers rather the nationality of the publisher. This would better reflect the focus of Chinese research in AI and its impact.

The research does not necessarily tell us much about its applications. For instance, there are many reasons to research topics such as pose estimation, facial detection, image recognition… But such tools can be combined into a technology that identifies political dissidents and create detailed models of their behaviour. Regardless of the country where an algorithm is developed, once it is created, it is hard to predict all the ways it can be used or the ethical implications.