JUNE 2020

AI Definitions Affect Policymaking

CSET Issue Brief

CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

AUTHORS
Dewey Murdick
James Dunham
Jennifer Melot

Definitions of artificial intelligence (AI) are often ambiguous and quickly out of sync with such a rapidly emerging technology. Yet these definitions can directly affect policy. This issue brief demonstrates that a simple model of national competition in AI strongly depends on the specifics of several alternative definitions, such as the method for identifying AI-relevant research and the choice of datasets. This sensitivity illustrates that analysts and policymakers should demand clarity around what is included within the scope of AI and the methods or data used to identify it.

CSET has implemented three principles to identify a functional definition of AI research. This definition leverages judgments from a large group of AI experts using machine learning. It significantly outperforms other methods from 2010 to the present while ensuring future relevance. Applying this method to international AI competition shows that the competitive landscape varies significantly in sub-areas such as computer vision (where China leads), robotics (where China has made significant progress), and natural language processing (where the United States maintains its lead).

Three Principles for Defining AI

Clear and actionable definitions of artificial intelligence and AI systems have proven elusive.[1] This is complicated by the following facts:

- Concepts evolve over time due to rapid technical changes and shifting perspectives;
- People disagree, especially at the boundary of a field; and
- AI is difficult to summarize given its many sub-areas characterized by differing methods, tasks, and application areas.

Furthermore, even when a solid plain-language definition is agreed upon,[2] it is challenging to implement without specific examples of what is and is not related to AI.

In response to these issues, we propose three pragmatic definitional principles to consider when supporting policymakers.[3] In our view, analysis that requires decisions about the "AI relevance" of entities like publications or companies should implement the following principles:

1. Capture the distributed judgments from a large group of skilled practitioners (i.e., the expert crowd) continuously or at least at regular intervals;
2. Compose a set of sub-areas of methods and tasks[4] with active research and practitioner communities formed around them; and
3. Link AI and its sub-areas to specific tangible examples of AI-relevant articles—and eventually, patents and products.

Grounding the definition with these criteria increases the likelihood that it remains a reliable basis for analysis and policy option discovery into the future. To accomplish this goal, it will be important to link such a definition to plain-language versions that are often required in policymaking. We expect that these principles should generalize to a broader range of emerging technologies of interest to policymakers.

CSET developed a functional AI definition using SciBERT—a recent neural network-based technique for natural language processing trained on scientific literature.[5] Compared to other AI definitions, the CSET method performed 88 percent better from 2010 to 2014 and 47 percent better from 2015 to 2019 (averaging 68 percent improvement from 2010 to present relative to the other methods).[6] This performance improvement was achieved by learning how to define AI from experts who use the arXiv pre-print repository.[7] This learning was then applied to predict which of the 35 million articles and preprints were examples of AI-relevant research.[8]

We believe the appropriate definitional choices depend on the analysis required. For instance, which experts we learn from could be a source of disagreement. Rather than argue for a universally applicable definition, we recommend a method that can be tailored to specific policy questions.

Policy Implications for National Competition Models

AI competition is a hot topic in the policy world.[9] Since AI is still an emerging technology, top research production by country or region can serve as a simple, imperfect approximation of national competition.[10] The competition for AI research leadership serves as a rudimentary yet illustrative example of how AI definitions can impact policy recommendations.[11]

Policy Impacts of Different Sub-Areas of AI

AI research is composed of many sub-fields with different methods and tasks. Using CSET's SciBERT definition of AI,[12] Figure 1 (panel a) shows China's rapid growth in AI research output surpassed that of the EU and the United States in 2015 and 2018, respectively. However, the intersection years shifted earlier for computer vision (panel b)—to 2014 (EU) and 2017 (United States)—and later for robotics—to 2017 (EU) and 2019 (United States) (panel c). In natural language processing (panel d), on the other hand, 2019 was an intersection year for China and the EU and there was no intersection for China and the United States. Using only a general AI model to guide AI policy (panel a), one would likely miss the U.S. lead in natural language processing research or fail to notice the significant progress China has made in computer vision.[13]

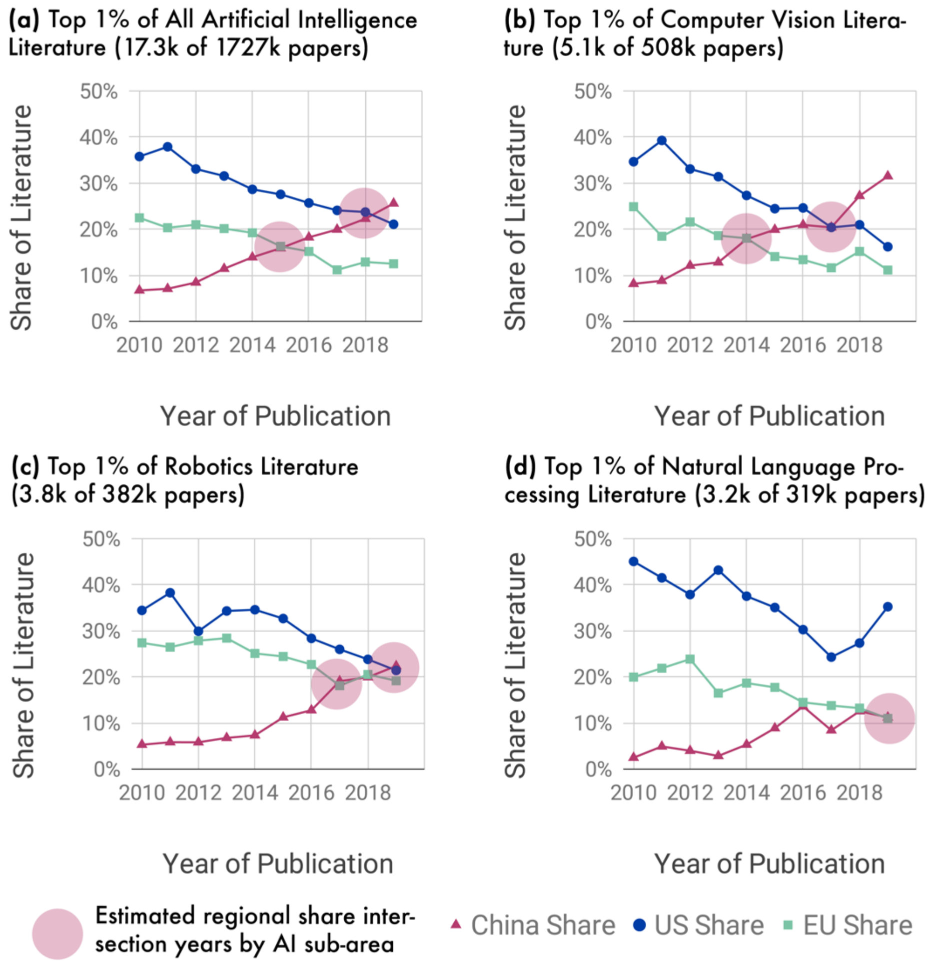

Figure 1. Regional competition in AI is shown using a simple comparison of the share of the top one percent of articles by citation generated by authors with institutional affiliations in China, the United States (US), and the European Union (EU). The years in which these shares intersect are shown in red for (a) all AI and three sub-areas of AI, (b) articles on computer vision, (c) articles on robotics, and (d) articles on natural language processing using CSET's SciBERT-based system

Policy Impacts of Methods for Identifying AI-Relevant Research

The computational methods for identifying AI-relevant articles also carry implications for policy analysis. We compare three methods of identifying AI-relevant articles:

1. Keyword search: the most commonly practiced way to define a field; CSET's version uses more than one hundred terms and patterns iteratively found to be associated with AI articles;[14]
2. Elsevier Hybrid: employs keywords, but added a machine learning component to determine their relative importance when identifying AI-relevant articles;[15] this approach was central to the most recent AI Index report, created to ground the conversation about AI in data; and[16]
3. SciBERT Model: CSET adopted the principles above and leveraged the SciBERT language model to develop a high-performing system that can label AI-relevant articles.[17]

Comparing the approaches to defining AI reveals a limited overlap (19 percent) in the number of articles identified as AI-relevant by all three models (see Figure 2(a)).[18] The overlap increases to 32 percent among articles in the top one percent of citation counts (not shown). It is noteworthy that other implementation choices, such as how country authorship is attributed or how the top one percent is calculated, can also affect the year in which top regional output intersects (as discussed in Figure 1). A recent Allen Institute for AI (AI2) blog projected the U.S. and China intersection year to be 2025,[19] in contrast with our result of 2018,[20] due to choices in implementation details.

The relatively small overlaps of each of these methods reveal the high stakes of method selection. Using different foundations to define AI—with only 20 to 30 percent similarity—will result in a field where experts talk past each other as major economic and political decisions are made. The divergence in results also raises serious concerns about how seemingly simple method choices inform analysis and AI policy in areas such as competition, investment, tech transfer, and application forecasts.

One example of the policy stakes comes from the AI Policy Observatory, launched by the Organisation for Economic Co-operation and Development in early 2020.[21] Due to its widespread reach and easy access, the platform will likely have an impact on policy discussions. The AI Observatory uses the Microsoft Academic Graph (MAG) topic categories to define AI, linking to 2.69 million AI-relevant articles; only 32 percent of these articles overlap with the CSET SciBERT definition.[22] This divergence further demonstrates how different AI definitions will have an unintended impact on policy discussions.

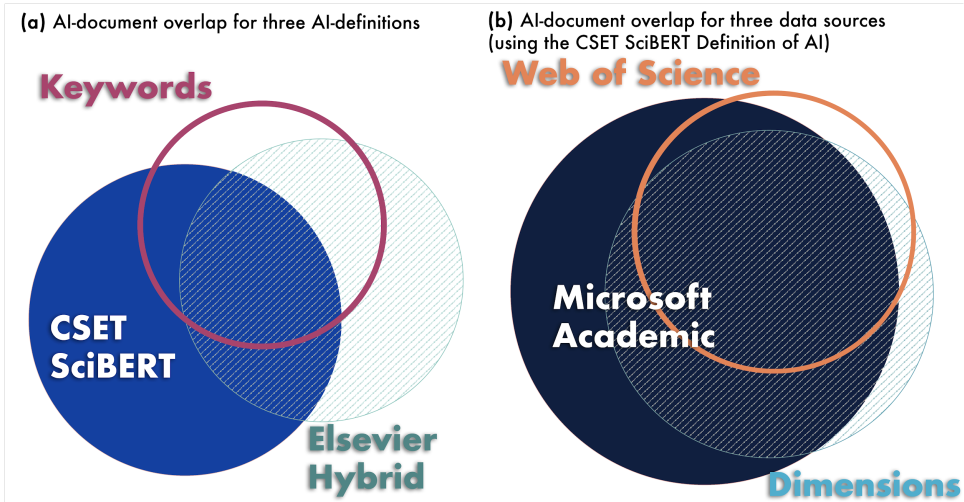

Figure 2. The quantity and overlap of the AI definitions in terms of articles published between 2010 and present are shown to scale in panel (a) using three different approaches: CSET's SciBERT model (1.73 million articles), Keywords (1.08 million articles), and the Elsevier Hybrid (1.42 million articles). The definitions agree for only 18.8 percent of the articles (504 thousand articles). The quantity and overlap of article sources are shown to scale in (b) for CSET's SciBERT AI definition. The vast majority of content is covered by Microsoft Academic Graph, or MAG (1.56 million articles or 90 percent of the 1.73 million total AI-relevant articles), compared with Dimensions (1.12 million articles, 65 percent) or WOS (0.82 million articles, 48 percent). MAG also contains 90 percent of the AI-relevant articles in Dimensions, and 88 percent of those in WOS.

Policy Impacts of Choosing Different Datasets

AI policy-relevant trend analysis is directly affected by what the selected dataset covers.[23] We compared coverage of AI research and its implications for analytic conclusions across three datasets of scholarly literature—MAG, WOS, and Dimensions—relative to a combined dataset. MAG has the largest coverage at 90 percent of the relevant literature. WOS and Dimensions together contain over 171 thousand AI articles not found in MAG (see Figure 2 (b)).

Dataset selection also affects the region-specific share of production of the top one percent of AI-related articles. In fact, the intersection years between China, the EU, and the United States (following the graphs shown in Figure 1) fall within one to three years of each other, depending on which dataset is selected. This discrepancy occurs because datasets collect different articles, so the definition of the top one percent of articles varies between WOS, MAG, and Dimensions. Additionally, even for the same articles, different datasets observe different numbers of citations, affecting the calculation of the year when the leading country changes.[24]

The most reliable method for improving coverage of relevant publications and enhancing their quality is to combine the available data sources. Our results show that the choice of dataset matters, so aggregation will improve the robustness of analysis for AI policy.

Concluding Recommendation

We strongly recommend that any policy advisor or agency carefully explore the policy impact of their choice of AI definition and leverage a data-driven underpinning that can be shared openly with stakeholders. The discussion above illustrates the need for a functional AI definition that operates at a fine-grained, article level over time, but that can also be leveraged by policymakers as they communicate with others. It also highlights the importance of complementing plain-language definitions of AI with positive and negative examples of what is AI-relevant and what is not.

Acknowledgments

For their helpful discussions, comments, and input, the authors would like to thank Tessa Baker, Kevin Boyack, Melissa Flagg, Kyle Lo, Karine Perset, Ilya Rahkovsky, John Rigsby, Carissa Schoenick, Alexandra Vreeman, and Kuansan Wang. The authors are solely responsible for all mistakes.

Appendix: Methods Discussion

Expert Judgments of AI-Relevance

Cornell's arXiv project hosts author-submitted preprints from a variety of scientific disciplines. Authors tag their articles with subjects from a taxonomy.[25] After moderator review,[26] we take these subject tags to reflect expert-crowd judgments of what constitutes AI-relevant research. The labels come from an established taxonomy of research fields. They provide positive and negative examples of AI work in an expanding set of more than 35 million articles and preprints within the combined Microsoft Academic Graph (MAG), Dimensions, and the Web of Science Core Collection (WOS) literature holdings generated by authors around the world. We use arXiv preprints and their subjects to develop a system that can label articles in other datasets as AI-relevant or not with high accuracy.[27] CSET has made the underlying code publicly available to aid in replication.[28]

Data and Measurement

We use data from three sources in this analysis: WOS, Dimensions, and MAG. This analysis includes all unique English-language articles from 2010–2019 for which a title and abstract is available.[29] As discussed above, the three datasets vary in which articles they include and the metadata associated with any given article. The citation counts for articles available in multiple datasets are usually different, and often substantially so. For instance, WOS reports fewer citations than MAG for 75 percent of articles in our analysis that appear in both datasets (Table A1).

Table A1. The same articles observed in MAG generally have higher citation counts then the identical articles in Dimensions (52.0 percent) and WOS (75.3 percent). This affects the calculation of the top one percent of AI articles. All articles are limited to the field of AI as defined by CSET's SciBERT model.

Percent articles with __ citations than MAG	*more*	*equal*	*fewer*
Dimensions	35.7%	15.3%	52.0%
Web of Science (WOS)	19.7%	5.0%	75.3%

Citation counts are important in our analysis because we focus on the top one percent of most-cited articles. For articles that appeared in more than one dataset, we chose to use the largest citation count for each (usually from MAG), on the assumption that the smaller counts (usually from the other, smaller datasets) probably omitted citations.[30] If we instead used the lowest citation count associated with a given article across datasets, the results of the country output analysis would be different. Also, when ranking articles by citation count, we encountered ties among articles for inclusion in the top percentile. Rather than arbitrarily break the ties, we included all tied articles, producing more consistent results. Other, similar analyses do not use this approach.[31]

The three datasets also differ in the information they provide about the institutional affiliations of authors. We use this metadata in our analysis to identify the country or countries with which we should associate an article. The shares of global research output that we report for China, the EU, and the United States include articles where at least one author has an institutional affiliation in one of these countries or regions, in any dataset, and no author has an affiliation in the remaining two, in any dataset. For example, an article with one Chinese author and one U.S. author would be counted in the output share for neither country. This choice had a small impact on the overall count of articles as multiple country affiliations only occurred in four percent of the AI-related articles.

The AI2 and CSET results mentioned above show the intersection year for top-one-percent output from the United States and China to be 2025 versus 2018, respectively. Possible reasons for this large difference include:

- AI2 and CSET use different definitions of AI-relevant articles. The analysis from AI2 used MAG's "artificial intelligence" field to define relevant articles. There is only a 29 percent overlap between the two methods;
- Each analysis used a different method for identifying country affiliations. The AI2 results depend on heuristics applied to the name and website of authors' institutional affiliations and the language of their articles. We used country locations for these institutions from

Web of Science Core Collection and Digital Science GRID,[32] which led to different country affiliations for 10 percent of the articles; and

- AI2 and CSET used different methods for identifying which articles were in the top one percent by citation. CSET avoided arbitrarily breaking ties between articles with the same number of citations and aggregated citation counts across our three data sources (see discussion above). This seemed important due to the very large number of articles with the same number of citations. AI2 used a single source (MAG) and did not explicitly address this tie breaking issue.

Article Deduplication

Most articles in the analysis are available in more than one of our three datasets: Web of Science, Dimensions, and Microsoft Academic Graph. Even relatively unambiguous metadata like publication year can differ across the datasets. Articles also appear more than once within any of these sources, often with incomplete metadata.

To deduplicate, we normalize titles, abstracts, and author last names, and then consider each group of articles within or across datasets that share at least three of the following (non-null) metadata fields to correspond to one article in the merged dataset: normalized title; normalized abstract; publication year; normalized author last names; references (for within-dataset matches); and digital object identifier (DOI).

© 2020 by the Center for Security and Emerging Technology. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

To view a copy of this license, visit https://creativecommons.org/licenses/by-nc/4.0/.

Document Identifier: doi: 10.51593/20200004

Endnotes

[1] P. M. Krafft, Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo, “Defining AI in Policy versus Practice,” (Paper presentation, AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, February 2020), https://doi.org/10.1145/3375627.3375835; Michael Haenlein and Andreas Kaplan, "A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence," *California Management Review* 61, no. 4 (2019): 5–14; and Pamela McCorduck, *Machines Who Think*, *2nd ed*. (Natick, MA: A. K. Peters, Ltd., 2004), 423–4.

[2] Organisation for Economic Co-operation and Development (OECD), “Scoping the OECD AI Principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO),” November 15, 2019, https://www.oecd-ilibrary.org/science-and-technology/scoping-the-oecd-ai-principles_d62f618a-en.

[3] Alternatively, policymakers can view AI primarily from the perspective of the capabilities that it enables or the type of mission impact they are likely to have. AI is often a euphemism for a broader trend that includes computer systems in decision making and other human activity more generally. This discussion of AI definitions is limited to research and emerging technology.

[4] Eventually, we aim to include application areas and capabilities into this framework, but this is still a work in progress.

[5] BERT stands for “Bidirectional Encoder Representations from Transformers” and is a highly effective method of inferring the meaning of words in context. A group at the Allen Institute for AI created SciBERT, a BERT model, for use in scientific literature; I. Beltagy, K. Lo, and A. Cohan, “Scibert: Pretrained language model for scientific text,” *EMNLP* (2019), https://arxiv.org/abs/1903.10676.

[6] We measure performance by the F-1 score that evaluates the AI-relevant/not relevant labeling of each method in a set of articles drawn from arXiv.org. See Figure 1 in James Dunham, Jennifer Melot, and Dewey Murdick, “Identifying the Development and Application of Artificial Intelligence in Scientific Text,” arXiv:2002.07143. The percent relative change is calculated as 100% * ([F-1 measure for CSET SciBERT] - [F-1 for the comparison model]) / [F-1 for the comparison model].

[7] See the methods appendix and a recent article on the topic: Dunham, Melot, and Murdick, “Identifying the Development and Application of Artificial Intelligence in Scientific Text.”

[8] The Microsoft Academic Graph (MAG) data is current as of December 26, 2019; Dimensions is current as of January 28, 2020; and WOS data is from February 4, 2020.

[9] Daniel Castro, Michael McLaughlin, and Eline Chivot, "Who Is Winning the AI Race: China, the EU or the United States?" (Washington, DC: Center for Data Innovation, August 19, 2019), https://www.datainnovation.org/2019/08/who-is-winning-the-ai-race-china-the-eu-or-the-united-states/; Graham Allison, "Is China Beating American to AI Supremacy?" *National Interest*, December 22, 2019, https://nationalinterest.org/feature/china-beating-america-ai-supremacy-106861; and Tim Hwang and Alex Pascal, "Artificial Intelligence Isn't an Arms Race," *Foreign Policy*, December 11, 2019, https://foreignpolicy.com/2019/12/11/artificial-intelligence-ai-not-arms-race-china-united-states/.

[10] The impact of a scholarly article is judged by how much it changes the direction of scientific exploration. This impact can be approximated by how many times it is cited by other AI researchers. Thus, if a country has a considerable share of the most influential AI articles, then it can be viewed as more of a leader than those with smaller shares. We measure quality-adjusted research output by counting the top one percent of articles published each year, determined by the number of citations. We then map publications to countries using the geographic location of researchers' organizational affiliations. A similar approach was taken by the Allen Institute for AI, see Carissa Schoenick. "China May Overtake United States in AI Research." (Blog: Medium, 3/13/2019), https://medium.com/ai2-blog/china-to-overtake-us-in-ai-research-8b6b1fe30595.

[11] The dynamics that drive citations within scholarly publications are complex, see Tahamtan, I. and Bornmann, L., "What Do Citation Counts Measure? An Updated Review of Studies on Citations in Scientific Documents Published between 2006 and 2018," Scientometrics, vol. 121, no. 3, pp. 1635–84, December 1, 2019. DOI: 10.1007/s11192-019-03243-4.

[12] See methods appendix.

[13] The choice of the citation count threshold for top publications is somewhat arbitrary. We use one percent here after experimentation with a few alternatives. As a rule, lower thresholds result in greater shares of output for authors affiliated with institutions in China.

[14] Dunham, Melot, and Murdick, "Identifying the Development and Application of Artificial Intelligence in Scientific Text."

[15] M. Siebert, C. Kohler, A. Scerri, and G. Tsatsaronis, "Technical background and methodology for the Elsevier's artificial intelligence report," Elsevier, 2018, https://www.elsevier.com/?a=829143.

[16] Raymond Perrault, Yoav Shoham, Erik Brynjolfsson, Jack Clark, John Etchemendy, Barbara Grosz, Terah Lyons, James Manyika, Saurabh Mishra, and Juan Carlos Niebles, "The AI Index 2019 Annual Report," (Stanford, CA: AI Index Steering Committee, Human-Centered AI Institute, December 2019), https://hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf.

[17] Beltagy, Lo, and Cohan, "Scibert: Pretrained language model for scientific text."

[18] Dunham, Melot, and Murdick, "Identifying the Development and Application of Artificial Intelligence in Scientific Text"; Venn diagrams were created to scale using eulerAPE; Luana Micallef and Peter Rodgers; "eulerAPE: Drawing Area-Proportional 3-Venn Diagrams Using Ellipses," *PLoS ONE* 9, no. 7 (2014), https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0101717.

[19] The AI2 blog defined AI as articles with the MAG field of study "artificial intelligence" (MAG ID: 154945302), https://github.com/allenai/china_ai.

[20] Schoenick. "China May Overtake United States in AI Research." See Appendix for a brief discussion of potential reasons for the different results.

[21] Visit https://oecd.ai and learn more about the AI Policy Observatory at https://www.oecd.org/going-digital/ai/about-the-oecd-ai-policy-observatory.pdf.

[22] OECD considers an article to be AI-relevant if it is tagged during the MAG's automated concept detection operation with a field of study that is categorized in either the "artificial intelligence" (https://academic.microsoft.com/topics/154945302?fullPath=false) or the "machine learning" (https://academic.microsoft.com/topics/119857082?fullPath=false) fields of study in the MAG taxonomy. OECD only includes results from other fields, such as "natural language processing," "speech recognition," and "computer vision" if they also belong to the "artificial intelligence" or the "machine learning" MAG AI definition, see "OECD.AI Data from Partners: Methodological Note," last updated February 19, 2020 (accessed April 12, 2020), https://www.oecd.ai/assets/files/Methodology_20200219.pdf. Visit https://oecd.ai/oecd-metrics-and-methods to explore the OECD AI Observatory.

[23] The data quality can impact country affiliation fidelity and article citation counts when key information is missing in country or affiliation fields, or when datasets with smaller literature coverage lack references from omitted publications.

[24] MAG generally includes more citations than Dimensions (for 52 percent of the articles) and WOS (for 75 percent of the articles). On average, the top five percent of articles have twice as many citations in MAG as Dimensions and four times as many citations as WOS. While MAG has more citations on average, it is not always the case—35.7 percent of Dimensions AI-relevant articles have more citations than MAG and 19.7 percent of WOS have more citations.

[25] arXiv Category Taxonomy, https://arxiv.org/category_taxonomy.

[26] Moderators confirm or correct the classification, see https://arxiv.org/help/moderation; Michelle McKinney, "arXiv.org," *Reference Reviews* 25, no. 7 (2011): 35–36, doi:10.1108/09504121111168622; Zeeya Merali, "ArXiv rejections lead to spat over screening process," *Nature News*, 2016, https://www.nature.com/news/arxiv-rejections-lead-to-spat-over-screening-process-1.19267.

[27] Dunham, Melot, and Murdick, "Identifying the Development and Application of Artificial Intelligence in Scientific Text."

[28] https://github.com/georgetown-cset/ai-definitions-for-policymaking.

[29] We retrieved the data for this analysis from Web of Science on February 4, 2020; from Dimensions on January 28, 2020; and from Microsoft Academic Graph on December 26, 2019. We describe the details of deduplication and language identification in Dunham, Melot, and Murdick, "Identifying the Development and Application of Artificial Intelligence in Scientific Text."

[30] We did not compare specific citations across datasets in this analysis, or analyze citations by year or other characteristics.

[31] Schoenick. "China May Overtake United States in AI Research." See Appendix for a brief discussion of potential reasons for the difference in results.

[32] "Global Research Identifier Database (GRID)," Digital Science, https://www.digital-science.com/products/grid/.