December 2021

Measuring AI Development

A Prototype Methodology to Inform Policy

CSET Data Brief

AUTHORS
Jack Clark
Kyle Augustus Miller
Rebecca Gelles

Table of Contents

Introduction: Upgrading AI Measurement for the 21st Century

This report demonstrates how contemporary approaches to bibliometric and technical analysis can give policymakers an adaptable tool to better understand ongoing AI developments.[1] By combining bibliometric analysis (via CSET's recently developed Map of Science derived from CSET's research clusters and merged corpus of scholarly literature) with qualitative knowledge of specific AI subfields, we show how detailed pictures of AI progress can be created to help policymakers understand the current state of research for specific AI technologies. This report applies our methodology to three topics: re-identification, speaker recognition, and image synthesis. Additionally, we offer ideas for how policymakers can integrate this approach into existing and future measurement programs.[2]

By combining a versatile and frequently updated bibliometrics tool with a more hands-on analysis of technical developments within a particular AI subfield, we can build out a detailed picture of the state of published research into specific topics. In particular, tracking how AI subfields show improvements in performance benchmarks can send a powerful signal about the state of technical progress. Benchmarks have played a critical role in spurring AI development for topics as wide-ranging as image recognition, self-driving cars, and robotics,[3] and they can be used to supplement bibliometric-based approaches, allowing us to see whether rapid research growth is resulting in genuine improvements in performance. This methodology differs from most contemporary approaches to technology assessment because it is structured to allow deeper system-led approaches, places a significant premium on the analysis of fast-moving AI publications (including bibliometric analysis of preprints), and is generically applicable to a variety of AI capabilities, rather than custom-designed for a specific area of study.

Why Does Measurement Matter?

Measurement is uniquely intertwined with policymaking, especially in the case of AI. For the past few decades, attempts by policymakers and academic researchers to assess and measure AI have been linked. The National Institute for Standards and Technology's (NIST) MNIST dataset of handwritten digits became a valuable source to help gauge the overall pace of AI progress as people used it as a simple benchmark to assess capabilities. The XVIEW dataset, designed by the United States DOD's Defense Innovation Unit Experimental (DIUx), played a role in stimulating the development of computer vision capabilities applied to satellite imagery and served as a resource that generated data about AI progress. Similarly, the Defense Advanced Research Projects Agency's (DARPA) self-driving car and robotics competitions have themselves been exercises in measurement and assessment that catalyzed work in a technical area and gave researchers a sense of emerging capabilities in a policy-relevant field. Perhaps the most prominent example of how measurement and assessment impacted policymakers was the development of ImageNet, a Stanford University project to help researchers test the performance of computer vision systems against a large, deliberately challenging (at the time) dataset.

This report aims to go one step further. The U.S. government today periodically conducts tests or assessments of AI capabilities, or develops new tests and datasets at the behest of experts or agencies to understand AI technologies. We instead propose a system for continuously monitoring and assessing AI-enabled capabilities for publication patterns that might highlight consequential trends to the government, and for continuously analyzing technical benchmarks that can help the government detect significant advancements. While many different organizations regularly try to assess and measure AI capabilities for a specific policy purpose, we propose a system for measuring capabilities and patterns within AI as a whole.[4]

After outlining our methodology, this report discusses how qualitative expert knowledge combined with bibliometric tools,

like CSET's Map of Science, can generate insights regarding developments in the specific research areas of re-identification, speaker recognition, and image synthesis. We also discuss some limitations of this approach and end with recommendations for how policymakers can support the development of similar types of analytical tools.

Methodology

Bibliometric Clustering: The technical basis of our methodology is a bibliometric analysis tool, in this case CSET's Map of Science—a merged corpus of research representing most of the world's scientific literature that groups papers into clusters based on the citation links between them.[5] For each research cluster, the Map of Science provides detailed metadata, such as keywords and core subjects, the country or countries in which the authors of the cluster papers are based, and the types of organizations commonly listed as funders for research. Additionally, it provides several CSET-calculated indicators, such as the percentage of papers within a cluster related to AI, computer vision, natural language processing, or robotics.[6]

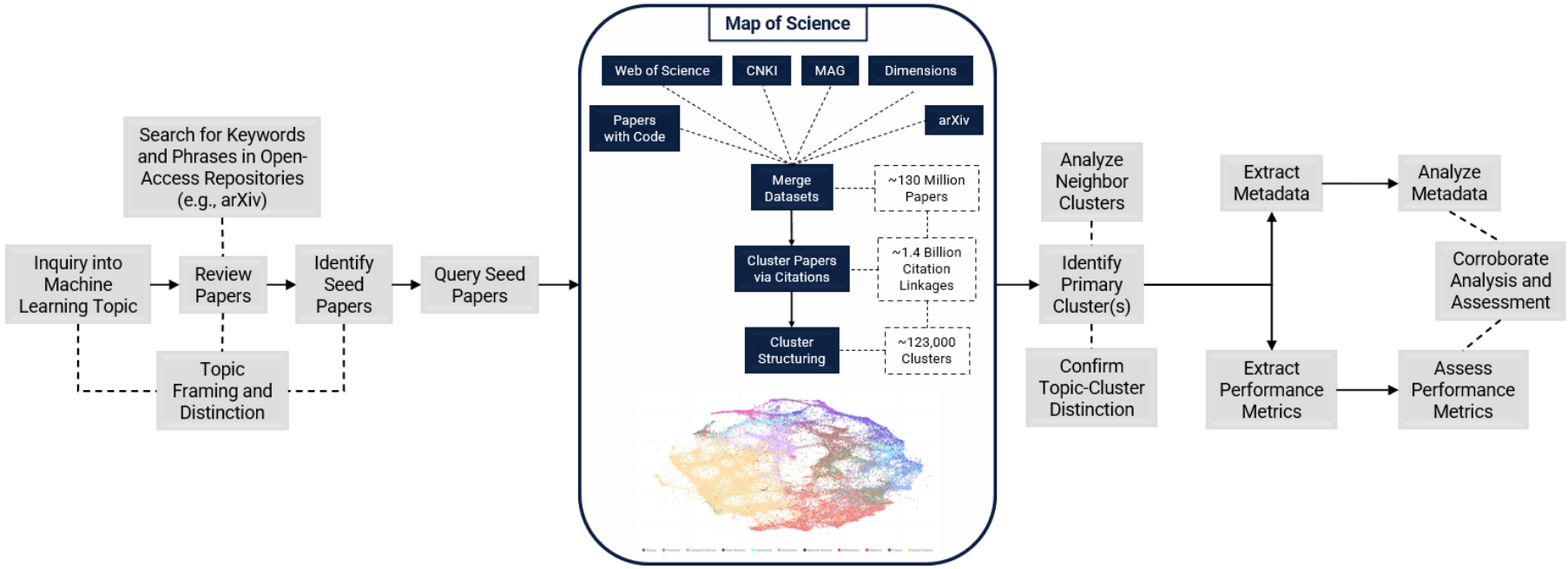
Bibliometric Analysis: To begin our analysis, we first selected topics related to machine learning and reviewed some of the relevant literature to identify 10–20 seed papers that are central to the given fields.[7] The seeds included highly cited works that were published across multiple years and encompassed key elements of the topic at hand (the seed papers used in the case studies can be found in Appendix B). We then searched for these papers in the Map of Science to identify the clusters that encompassed the highest number of them. Identifying quality seed papers requires a degree of domain expertise, but the approach to identifying clusters can be altered to be more systems-led and operable for non-experts. For example, one could use keyword searches in the Map of Science to identify relevant clusters and highly cited papers, and then manually inspect the papers.

For the purposes of this report, we explored whether we could narrow each machine learning topic to a single cluster from which we could extract metadata. These single clusters then became the

core of our subsequent bibliometric analysis. However, adjustments can be made to this method depending on the distribution of seed papers across clusters (i.e., research concentration), ambiguity of the topic, or quality of the data. For example, the bibliometric analysis could include multiple clusters if papers related to a topic are highly distributed. We must stress that individual clusters in the Map of Science represent related research (via citations between papers) and are not explicitly structured to encompass a topic. Therefore, in various cases, a multi-cluster analysis will be optimal or necessary.

Performance Metrics Assessment: In machine learning research, emerging areas of study often become focused around a performance metric or group of metrics, which tends to spur further investment in the field. The clusters used in the bibliometric analysis provide relevant research to look for such metrics. These metrics can supplement and corroborate the insights from the metadata analysis. By manually reviewing the literature within a research cluster, we can identify when a performance metric for the field first emerged and how top-level performance on that metric has changed over time. This can provide policymakers with a better sense of how rapidly capabilities in a given area are improving and where such developments are occurring. A performance assessment works well when there are existing common benchmarks to evaluate but can be more challenging if there are no relevant metrics available or they are outdated. However, by identifying a lack of common metrics, our methodology can illuminate knowledge gaps in areas that are relevant to policymakers.

Figure 1. Bibliometric Clustering, Bibliometrics Analysis, and Performance Metrics Assessment



Note: See Appendix A for a more detailed description of bibliometric clustering via the Map of Science, bibliometric analysis, and performance metrics assessment. Source: CSET.

By applying this methodology, we can measure the publication growth of specific machine learning–related topics, where it is occurring, what organizations and individuals are involved, and when significant technical advancements occur. This can be adjusted or augmented with additional processes, but we believe these three elements establish an effective baseline methodological structure that is applicable to a range of topics and users.

It should be noted that this approach is not perfect. Clustering tools like the Map of Science do not necessarily encompass the entirety of a given subject area, and they may often contain overlapping, incomplete, or unreliable metadata that must be manually reviewed. Bibliometrics also do not include private or closed-source research, and instead rely exclusively on published scientific literature. Finally, both the selection of seed papers and the identification of performance metrics are subjective and often involve heuristic decisions that cannot be automated easily (at present).[8]

Case Studies: Re-Identification, Speaker Recognition, and Image Synthesis

In this section, we apply the methodology to gain insights into the state of research in two emerging machine learning–enabled capabilities: re-identification and speaker recognition. We selected these topics because their emergence may have significant policy implications. We compare the results of this analysis to a more mature subfield of machine learning technology: image synthesis. To demonstrate this methodology clearly and without excessive bibliometric analytics, we narrowed the scope of analysis to a single research cluster for each capability (topic); however, this approach can be altered to include multiple clusters.

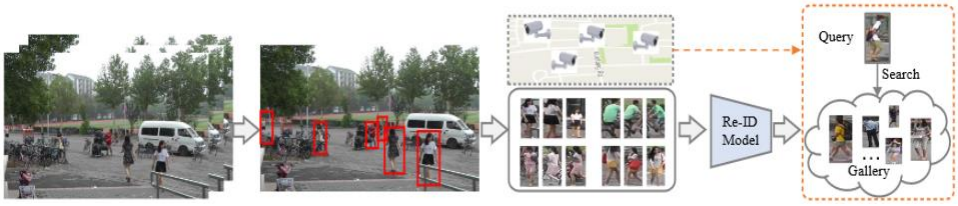
The results described below provide meaningful visibility into the rate, scale, distribution, and improvement of re-identification and speaker recognition capabilities across countries, organizations, and individuals. However, our particular application of the methodology—using one cluster per topic—was unsuccessful with image synthesis, primarily because research papers related to this topic are distributed across multiple clusters.

Topic Overview

Re-Identification

Re-identification involves integrating computer vision techniques with networked surveillance systems to identify and distinguish objects, track their locations, and correlate their movements (around obstacles and amidst appearance changes). The input to these models is typically CCTV footage, and the underlying methods can be adapted to track people, vehicles, or objects. Recent progress in re-identification was enabled by major breakthroughs in computer vision between 2010 and 2015, which made use of deep neural networks to rapidly improve on image recognition datasets like ImageNet.[9] (See Appendix C for additional details.)

Figure 2. Re-Identification Process



Source: Mang Ye et al.[10]

The development of re-identification systems has important implications for policymakers: these systems could potentially offer extremely useful tools to law enforcement, making it cheaper and easier to surveil large populations, apply other machine learning models to the resulting data, and subsequently uncover patterns of activity and association amongst individuals of interest.[11] These capabilities exhibit the potential to augment or replace humans in the manual analysis of CCTV footage and dramatically alter the process of imagery intelligence in public surveillance. Entities in both authoritarian and non-authoritarian regimes have incentives to develop re-identification for various purposes—such systems can equally help retailers understand patterns of traffic in shopping malls as they could help police track and identify criminal or dissident behavior.

Speaker Recognition

Speaker recognition is an area of research that focuses on using machine learning to recognize the voices of specific speakers in audio data.[12] Speaker recognition systems can be adapted to perform a number of more specific tasks, such as verifying the identity of a known speaker, identifying audio from an unknown speaker in a group of known speakers, or partitioning audio into segments depending on when each speaker is talking.[13] As with re-identification, current progress in speaker recognition is heavily dependent on previous advances in audio speech recognition during the 2010–2015 era, as measured by performances on datasets like Switchboard.[14]

Speaker recognition models can be integrated into a range of different systems that use audio data, including trivial applications

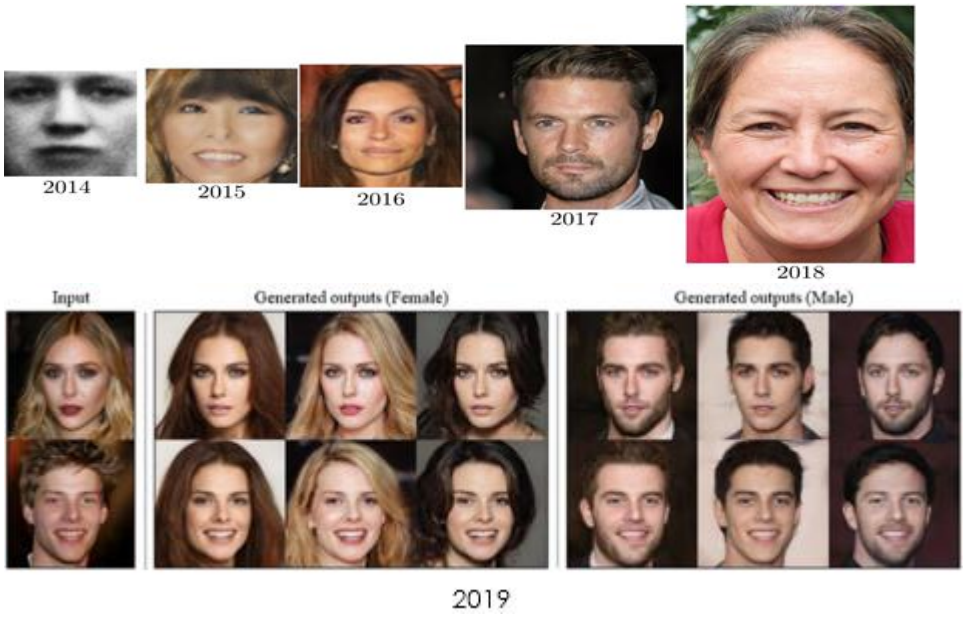
that identify celebrities' voices in YouTube videos and security programs that verify voice identities to grant physical access to sensitive locations. Improvements in speaker recognition merit attention from policymakers: these systems could enable improved intelligence processes that automate the analysis of raw phone audio data at unprecedented scales. (See Appendix C for additional details.)

Image Synthesis

Image synthesis is a subfield of machine learning research that focuses on generating realistic-looking images.[15] It has been closely linked to the development of generative adversarial networks (GANs) since 2014.[16] This topic includes the notorious world of "deepfakes," or hyper-realistic images of human beings that have already been used in Hollywood special effects, political campaigns, and online propaganda. By comparing the fields of re-identification and speaker recognition to image synthesis, we can assess how our method of analysis performs on both specific subfields, as well as more generic ones like image synthesis.

Image synthesis presents a host of issues with policy implications. In 2019, the president of Gabon was accused of using synthetic human imagery (or more prosaically, a body double) to stage a video meant to reassure civilians in the country.[17] In India, a politician made himself appear to give a speech in a language he did not speak, while propagandists in West Papua used synthetic facial imagery for fake accounts that aimed to undercut the independence movement.[18] By 2021, image synthesis was widely used for a variety of legitimate and illegitimate purposes, ranging from entertainment to making disinformation more convincing.

Figure 3. Image Synthesis Progression, 2014-2019



Sources: Ian J. Goodfellow et al.; Alec Radford, Luke Metz, and Soumith Chintala; Ming-Yu Liu and Oncel Tuzel; Tero Karras et al.; Tero Karras, Samuli Laine, and Timo Aila; and Yunjey Choi et al.[19]

Results and Analysis

In what follows, we focus first on re-identification and speaker recognition. We then show how our attempt to replicate the methodology for image synthesis was unsuccessful when using a single-cluster analysis because many relevant papers were distributed across multiple clusters. To baseline our bibliometric results, we compare the growth of re-identification and speaker recognition papers against changes in the composition of all AI papers, as identified by CSET's Map of Science. We also baseline relevant performance metrics by comparing them to two broader, but related, topics of computer vision and natural language processing.

1. Key Takeaways

Overall:

(1) We identified two respective research clusters for re-identification and speaker recognition and are confident that they encompass most publications on the topics. From the papers in these clusters, we extracted metadata and performance metrics.[20]

(2) Publications related to re-identification and speaker recognition increased substantially between 2010–2020, and their performance metrics improved significantly. A continuation of this trend will likely have political, security, and commercial consequences that warrant policymaker attention.

(3) Growth of re-identification and speaker recognition publications occurred alongside the discovery and proliferation of many key machine learning subcomponents, including deep learning–based techniques, large datasets, and increased computational capacity.

(4) Most organizations affiliated with the authors of papers for both topics were educational, although speaker recognition had a higher affiliation with private industry.

Re-Identification	Speaker Recognition
(5) Re-identification publications in cluster #1419 had an average annual growth rate of 40 percent between 2010–2020. Publications increased substantially between 2015–2018, then plateaued in 2020.	**(5)** Speaker recognition publications in cluster #6855 had an average annual growth rate of 33 percent between 2010–2020. Publications increased substantially around 2017, which continued through to 2020.
(6) There were 67 author-affiliated countries in the re-identification cluster, of which China was the most prominent by a large margin, followed by the United States, the UK, and Australia.	**(6)** There were 64 author-affiliated countries in the speaker recognition cluster, of which the top two were the United States and China, which had relatively equal affiliation, followed by India, the Czech Republic, and Japan.
(7) There were over 600 author-affiliated organizations in the re-identification cluster, most of which are in the education sector. The top three were the Chinese Academy of Sciences, Queen Mary University of London, and Sun Yat-sen University.	**(7)** There were over 400 author-affiliated organizations in the speaker recognition cluster, most of which are in the education sector. The top four were Johns Hopkins University, MIT, Brno University of Technology, and SRI International.
(8) The performance of re-identification models improved significantly in recent years, nearly doubling in performance against our selected benchmark since 2018. Multiple organizations—predominantly from China—made progress on one of the most challenging testing datasets.	**(8)** The performance of speaker recognition models improved significantly in recent years. Improvements were driven by researchers at the University of Oxford (UK) who designed the metrics (i.e., dataset and challenge), Ghent University (Belgium), and most recently SpeakIn Technologies (China).

2. Bibliometric Analysis

From our bibliometric analysis we identified and extracted metadata from two respective research clusters for re-identification and speaker recognition. These clusters encompass most, but not all, papers on the topics, which makes an analysis of the clusters' metadata akin to an analysis of the topics. For re-identification, 12 of our 14 seed papers fell into a single cluster (#1419), with the other two falling into a cluster related to object detection and tracking that was relatively light on machine learning–related papers. Even more promisingly, all 15 of our seed papers for speaker recognition fell within a single cluster (#6855). After a spot-check confirming the clusters' relevance, we felt comfortable making the focus of our analysis a single cluster for each topic—most subsequent metadata in this report is that of the papers in these clusters.

Here we begin with an analysis of publication and citation growth within each cluster (per topic), then proceed to investigate country affiliation, organization affiliation, and authors. To provide context, we compare the growth of our two topic clusters with the growth of AI broadly (i.e., all publications related to AI). However, unlike with re-identification and speaker recognition, the AI data does not stem from clusters. Instead, we used a trained classifier model to identify and extract all "AI-related" publications from the merged corpus,[21] ultimately giving us approximate figures on the entire field. Summary statistics for the two topic clusters, as well as AI publishing overall, are displayed in Table 1. As we describe in more detail below, image synthesis is omitted from this section as it required a multi-cluster analysis, while our focus in this paper is on single-cluster bibliometric analysis.

A. Publications and Citations

Table 1. Papers, Citations, and Growth Rates

	Re-Identification (Cluster #1419)	Speaker Recognition (Cluster #6855)	AI
Total Papers (2010-2020)	3,282	1,938	2,015,541
Total Paper Citations (2010-2020)[22]	96,462	29,096	24,896,447
Average Annual Paper Growth Rate (2010-2020)	40%	33%	27%
Average Annual Paper Citation Growth Rate (2010-2020)	21%	15%	23%

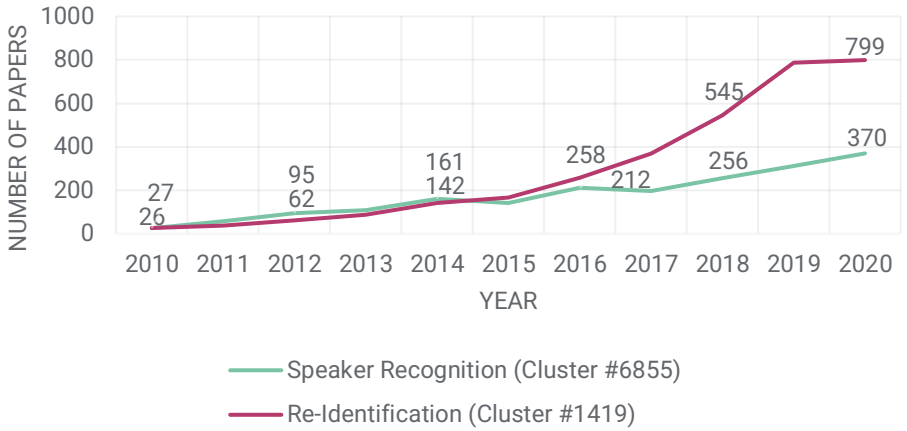
Source: CSET Map of Science derived from CSET's research clusters and merged corpus of scholarly literature, as of March 2021.

Research into both re-identification and speaker recognition has grown more quickly than overall AI research. AI grew at an annual rate of 27 percent from 2010–2020, while the number of papers related to re-identification and speaker recognition grew at an average rate of 40 percent and 33 percent per year, respectively. Although the citation numbers in Table 1 may make it appear that citations for re-identification and speaker recognition papers have not grown as quickly as citations for AI papers overall, this is likely an artifact, in part, of the relatively later increase in attention to these sub-topics—especially re-identification—around 2015, as shown in Figures 4 and 5. Such recent papers can take time to accumulate citations, which helps explain why the citation numbers do not appear to have grown as rapidly in the two subfields.

Figure 4 shows a divergence in growth between the topics around 2015, with re-identification publications quickly outpacing speaker recognition. We suspect that this divergence is related to the fact that computer vision has been a larger research area with a

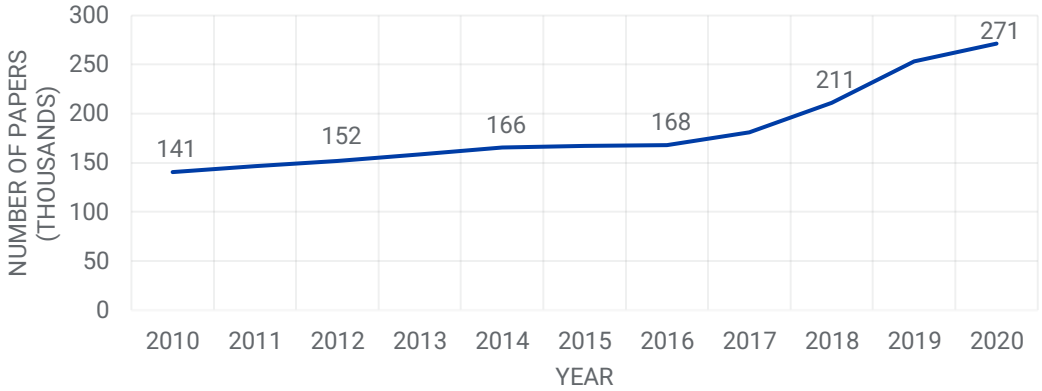
broader problem space than voice biometry. However, we are ultimately uncertain of its cause and further investigation is warranted, as many trends of the early 2010s—such as the development of deep neural networks, increased computing capabilities, and a growth in available data—enabled both types of research.

Figure 4. Machine Learning Topic Papers Published per Year, 2010–2020



Note: For a more detailed analysis of re-identification and speaker recognition publication progression, see Appendix D. Source: CSET Map of Science, as of March 2021.

Figure 5. AI Papers Published per Year, 2010–2020



Source: CSET Map of Science, as of March 2021.

Image Synthesis—Topic Necessitates a Multi-Cluster Investigation: We were unable to identify a single cluster for image synthesis research because relevant publications were distributed across multiple clusters. This largely stems from the fact that clusters in the Map of Science are representations of related research and are not explicitly intended to represent one topic in one cluster. Subsequently, topics that have more distributed research (e.g., image synthesis) require analyzing multiple clusters, while topics with more concentrated research (e.g., re-identification) can be analyzed through a single cluster. Following an extensive spot-check of image synthesis publications across clusters, we decided to omit this topic from our single-cluster focus, as we were unable to get reliable metadata on this topic from one cluster. Future research could explore this topic through a multi-cluster analysis.

Here we briefly address three related limitations that stem from restricting this methodology to a single cluster per topic and from using a cluster of papers (based on citations) to investigate a topic:

- We identified one cluster (#1220) that contained many publications related to image synthesis; however, it did not include five of our twelve image synthesis seed papers, which suggests that it did not sufficiently encompass publications directly related to the topic.

- Many publications related to image synthesis are dispersed across research clusters. In our initial attempt to identify a single cluster for this topic, we had to omit clusters that, while slightly relevant on a technical level, contained many papers and research topics not directly related to image synthesis as a capability.[23] These topics include facial recognition, point clouds, neural style transfer, and deepfake detection. This differs from the previous investigations, during which we identified single clusters that encompassed most papers directly related to re-identification and speaker recognition.

- Cluster #1220 included many publications explicitly about GANs and various extraneous topics (e.g., generative

grammar). To avoid having to manually omit data, and because GANs are deeply intertwined with the advancement of image synthesis, we considered partially incorporating GANs into the investigation and attempted to analyze the entire cluster. However, using the whole cluster and expanding the scope of the topic made the investigation too imprecise, as GANs are used for many purposes beyond image synthesis, and we could not readily discern the topics within the cluster metadata. This issue can arise when a topic is not sufficiently distinct, or it encompasses technologies with dual-use applications.

B. Countries

CSET's Map of Science provides country affiliation metadata, which is derived from organizational affiliations provided by the authors of papers. The denominator for this country analysis is not the quantity of papers in the clusters—it is the total country affiliations of the papers in the clusters.[24] Extracting this metadata from our re-identification and speaker recognition clusters allows us to see which countries were most affiliated with paper authors and which accrued the most paper citations (i.e., how many times a country-affiliated paper is cited). This is displayed in Tables 2 and 3, which show the top five countries in terms of author country affiliations and paper citations.[25]

Table 2. Top 5 Author Country Affiliations, per Topic

	AI		Re-Identification		Speaker Recognition	
1.	China	27%	China	58%	United States	24%
2.	United States	16%	United States	13%	China	22%
3.	India	5%	UK	6%	India	6%
4.	UK	5%	Australia	5%	Czech Republic	4%
5.	Japan	4%	Italy	4%	Japan	4%

Source: CSET Map of Science, as of March 2021.

Table 3. Top 5 Countries' Share of Paper Citations, per Topic

	AI		Re-Identification		Speaker Recognition	
1.	United States	35%	China	60%	United States	48%
2.	China	26%	United States	24%	Canada	15%
3.	UK	9%	UK	14%	France	14%
4.	Germany	7%	Australia	13%	China	11%
5.	Canada	6%	Italy	8%	UK Source:	7%

CSET Map of Science, as of March 2021.

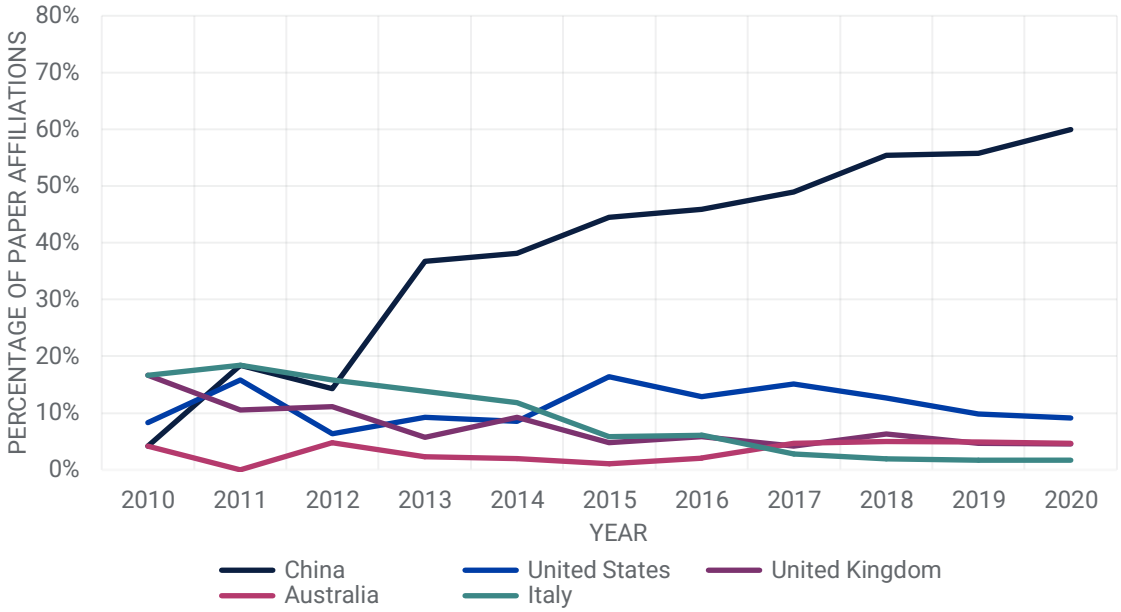
Compared to AI research generally, China was affiliated with a slightly smaller percentage of papers in speaker recognition but a substantially larger percentage of re-identification papers. This finding suggests that Chinese entities may have a particular interest in re-identification. Conversely, the United States was affiliated with a slightly smaller percentage of re-identification papers and a slightly larger percentage of speaker recognition papers than would be proportionate to its share of all AI research. It is, however, important to emphasize that much research in each of these fields is highly international; the top-cited re-identification paper between 2010–2020, for instance, was written by researchers at the University of San Antonio, Microsoft, and Tsinghua University.[26]

United States-affiliated papers for both topics tended to garner disproportionately more citations relative to the number of publications the country is affiliated with. China-affiliated speaker recognition papers were cited disproportionately little relative to the country's affiliated publication output but roughly on par with its affiliated publication output in re-identification and AI overall.

Notably, the composition of the countries affiliated with re-identification papers changed between the years 2010–2020. This is displayed in Figures 6 and 7, which show the share of paper affiliations and paper citations for the top five countries per year (from the re-identification cluster).[27] While no single country

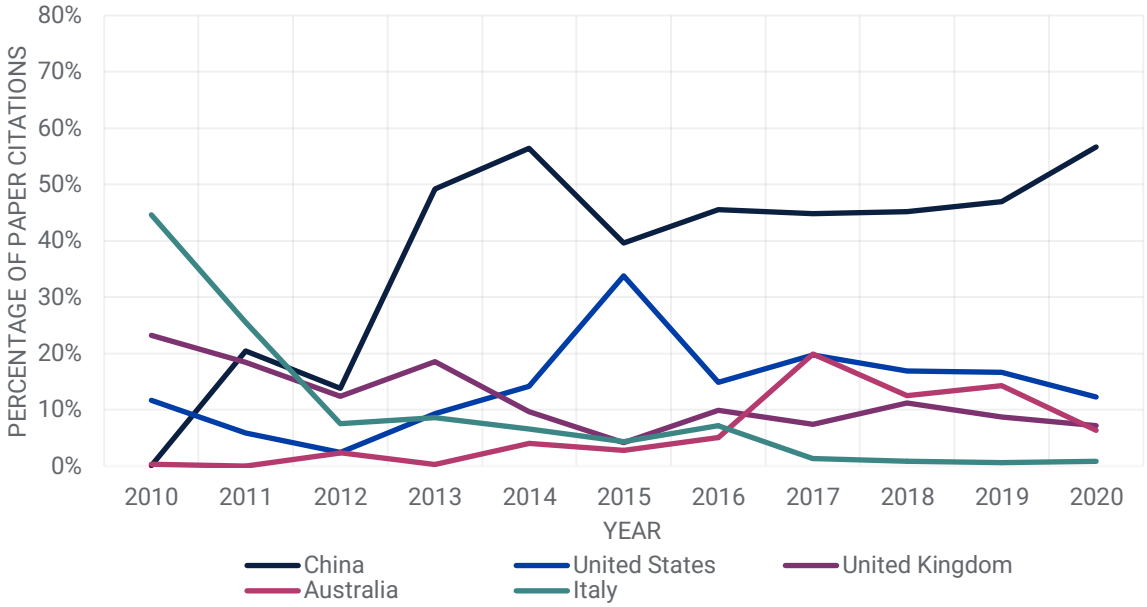
initially appeared dominant, China has been affiliated with most papers in this cluster since 2015.

Figure 6. Country Share of Re-Identification Paper Affiliations, 2010–2020



Source: CSET Map of Science, as of March 2021.

Figure 7. Country Share of Re-Identification Paper Citations, 2010–2020



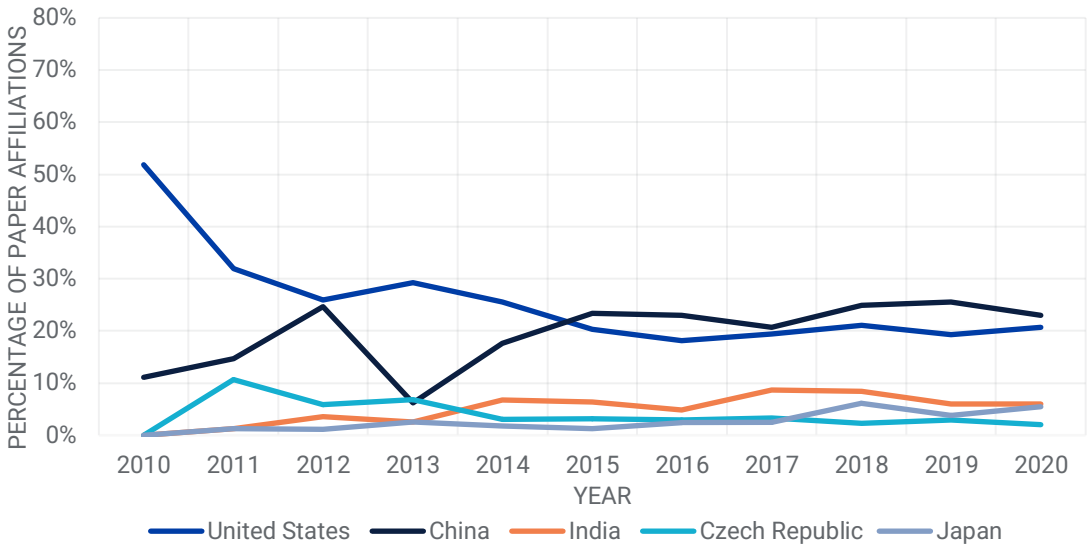
Source: CSET Map of Science, as of March 2021.

China-affiliated papers garnered 60 percent of the citations in our re-identification cluster,[28] while United States-affiliated papers garnered 24 percent of the citations. UK and Australia-affiliated papers had relatively low citations of 14 and 13 percent, respectively. Although China was still the country most affiliated with papers and paper citations, the citation distribution suggests that many entities from the other top countries are relatively more prominent than the paper distribution alone suggests.

Researchers affiliated with Italy, the UK, and the United States were involved in re-identification publications in our cluster in 2010 (the beginning of our analysis). This was prior to the development of many key components used in emerging re-identification systems. By 2013, China leaped ahead of all other countries,[29] maintaining a lead in affiliated publications to 2020.[30] Overall, these trends suggest that China is far more active in published re-identification research than any other nation.

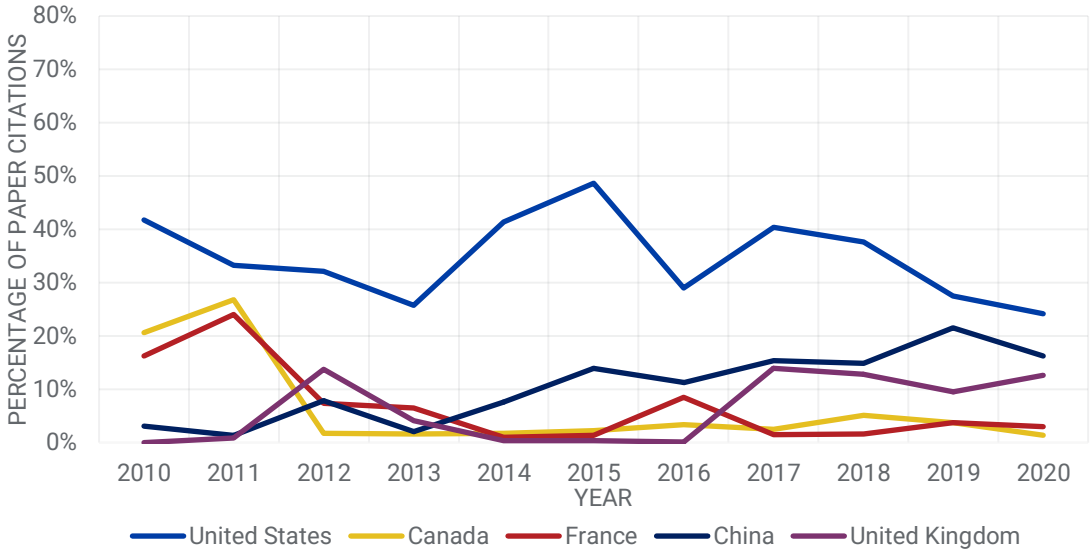
In the speaker recognition cluster, by contrast, we see that the United States was initially the country most affiliated with publications, but by 2015 the proportion of papers from Chinese entities increased and remained roughly on par with publications by American researchers—though the United States maintained an edge in its share of paper citations. This is displayed in Figures 8 and 9, which show the share of paper affiliations and paper citations for the top five countries per year, for the speaker recognition cluster.[31]

Figure 8. Country Share of Speaker Recognition Paper Affiliations, 2010–2020



Source: CSET Map of Science, as of March 2021.

Figure 9. Country Share of Speaker Recognition Paper Citations, 2010–2020



Source: CSET Map of Science, as of March 2021.

Among speaker recognition paper citations, the United States garnered 48 percent of the citations in our speaker recognition cluster. This was followed by Canada and France-affiliated papers at 15 and 14 percent citations, respectively. Although China was ranked second for paper affiliations, its affiliated papers only garnered 10 percent of the citations in the cluster. Similarly, although India was ranked third for paper affiliation, its affiliated papers only garnered 3 percent of the citations.

With speaker recognition, researchers affiliated with the United States, China, and the Czech Republic were involved in publications between 2010–2011, as illustrated in Figures 8 and 9. As with re-identification, the growth rate of speaker recognition papers affiliated with the United States and China increased substantially around 2016. The United States was affiliated with papers that accrued more citations during most years, even though China was affiliated with more publications between 2015–2020. Notably, recent speaker recognition growth appears more internationally distributed than re-identification. For example, China and the United States were collectively affiliated with approximately 75 percent of the re-identification cluster's publications in 2019 but only affiliated with 58 percent of the speaker recognition cluster's publications the same year. Explaining the differences in the top affiliated countries between the two topic clusters is an area for additional research.

C. Organizations

Organizational affiliation metadata is based upon information provided by authors in their publications and provides a more granular level of visibility into machine learning publications trends. Of the over 600 author-affiliated organizations in the re-identification cluster,[32] 76 percent were universities, 9 percent are government organizations, and 6 percent are private companies.[33] Of the over 400 author-affiliated organizations in the speaker recognition cluster, 65 percent were universities, 11 percent are private companies, 5 percent are government organizations, and 4 percent are in the non-profit sector. Table 4 displays the ten organizations with the largest publication output for both re-identification and speaker recognition.[34] Although the data further

demonstrates the dominance of Chinese organizations in re-identification research, the results were (marginally) less Chinese-centric when looking at paper citation counts. Microsoft-affiliated authors of re-identification papers, for instance, accounted for 5 percent of all citations in the cluster. And yet, even here there is significant Chinese involvement: Microsoft conducts much of its re-identification research through the Beijing-based Microsoft Research Asia, its largest research lab outside of the United States.[35]

Table 4. Top Organizations Affiliated with Papers (Per Topic)

	Top 10 Organizations Affiliated with Papers	Number of Papers	Share of Topic Papers
Re-Identification	**Australia**		
	University of Technology, Sydney	71	2%
	China		
	Beihang University	70	2%
	Beijing University of Posts and Telecommunications	71	2%
	Chinese Academy of Sciences	253	8%
	Peking University	85	3%
	Shanghai Jiao Tong University	104	3%
	Sun Yat-sen University	113	3%
	Tsinghua University	70	2%
	Wuhan University	82	2%
	UK		
	Queen Mary University of London	85	3%

Speaker Recognition	**China**		
	Chinese Academy of Sciences	71	4%
	Hong Kong Polytechnic University	46	2%
	Tsinghua University	69	4%
	Czech Republic		
	Brno University of Technology	62	3%
	India		
	Indian Institutes of Technology System (IIT System)	40	2%
	United States		
	IBM	39	2%
	Johns Hopkins University	71	4%
	Massachusetts Institute of Technology	53	3%
	SRI International	51	3%
	University of Texas at Dallas	55	3%

Note: For a more detailed analysis of re-identification and speaker recognition organization affiliations, see Appendix E. Source: CSET Map of Science, as of March 2021.

D. Authors

Author metadata provides the most granular level of bibliometric visibility into machine learning topics, enabling the identification of prominent individuals involved in open-source publications. In science, certain authors tend to serve as important "datapoints" in their own right—for instance, certain authors may run a lab exclusively dedicated to research in one area, so surfacing them can help governments identify figures with significant influence in a given field. Conversely, other authors may be multidisciplinary and multi-institutional, serving as the connective tissue between disparate sets of authors and research specialties. As with organizations, we only intend to illustrate the potential for insight,

and thus the analysis here is limited. However, the data creates opportunities for deeper analysis, such as identifying prominent authors of interest for research collaboration and policy formulation, as well as government investigation into foreign authors researching machine learning capabilities that could have security implications.

The data on the authors' current country locations was manually collected, as opposed to extracted from the clusters, because at the time of our analysis such information was not readily available in the prototype Map of Science.[36] Additionally, many authors have diverse backgrounds that include multiple organizations and work locations over time, much of which is not portrayed through the metadata. Notwithstanding these shortcomings, the data can be a starting point for a deeper analysis of individual researchers, the organizations they work for, and the professional networks in which they are embedded.

Of the top twenty authors that published re-identification papers in cluster #1419, almost half were located in China, with most others dispersed across the UK, Australia, Italy, and Japan. Notably, only one of the top authors was located in the United States despite its relatively high affiliation to papers. Sixteen authors were affiliated with organizations in the education sector, primarily public universities, while four work in private industry at Huawei (China), the Inception Institute of Artificial Intelligence (UAE), the Samsung AI Center (UK), and Google (United States). The prominence of authors in the education sector is influenced by the fact that they publish their work, but we cannot determine to what extent this skews the data toward education and away from other sectors.

Of the top twenty authors that published speaker recognition papers in cluster #6855, five were located in the United States and China, with the remaining dispersed across Singapore, the Czech Republic, Sri Lanka, Italy, India, Canada, Australia, and Argentina. Sixteen top authors were affiliated with organizations in the education sector, primarily public research universities, while two were in private industry at Amazon and Apple (United States), two in the nonprofits SRI International (United States) and the

Computer Research Institute of Montréal (Canada), and only one in government (Singapore's Agency for Science Technology and Research). As previously noted, many authors have diverse backgrounds that include multiple organizations and work locations, much of which is not portrayed through the data.

Table 5. Locations of Top 20 Paper Authors

	Country	Number of Top 20 Authors
Re-Identification	China	9
	UK	3
	Australia	2
	Italy	2
	Japan	2
	United Arab Emirates	1
	United States	1
Speaker Recognition	United States	5
	China	5
	Singapore	2
	Czech Republic	2
	Canada	1
	Australia	1
	Sri Lanka	1
	India	1
	Argentina	1
	Italy	1

Source: CSET Map of Science, as of March 2021.

3. Performance Metrics Assessment

Although bibliometric tools like CSET's Map of Science provide a useful means of exploring the growth, geographic distribution, and affiliations of research in emerging areas, bibliometrics alone do not say anything about the quality of a growing body of research. To assess whether the growth in re-identification and speaker recognition research has generated corresponding improvements in performance, we need to find useful performance metrics from within those fields to track. Fortunately, after reviewing papers in the clusters we identified two such metrics: MSMT17 for re-identification and VoxCeleb for speaker recognition, which are described in Table 6. However, if no such metrics existed, we could scan publications to identify datasets and performance evaluation methodologies shared across multiple areas, then manually compile the data to give performance metrics. Most all papers used to acquire performance metrics are from the same two clusters used in the bibliometric analysis.[37] Lastly, we must note that experts in the field often disagree over what datasets to measure progress on, so a more rigorous application of this methodology could involve tracking progress on different models across varying datasets to get a more nuanced picture of development.

Table 6. Performance Metrics and Datasets

	Dataset	Dataset Description	Performance Metric	Metric Description
Re-Identification	MSMT17	MSMT17 is a multi-scene, multi-time, person re-identification dataset. It consists of 180 hours of videos, captured by 12 outdoor cameras and 3 indoor cameras during 12 time slots.	mean Average Precision (mAP)	mAP calculates the accuracy of a model's imagery detection. A higher mAP score signifies greater accuracy.
Speaker Recognition	VoxCeleb (VoxCeleb1; VoxCeleb2)	VoxCeleb1 is an audio-visual dataset for speaker recognition containing hundreds of thousands of "real world" utterances from more than 1,000 celebrities. VoxCeleb2 contains over a million utterances from more than 6,000 speakers.	Equal Error Rate (EER)	EER calculates the likelihood of a biometric system incorrectly classifying the subject of interest (e.g., speaker audio) via false positives or false negatives. A lower EER score signifies less error and greater accuracy.

Source: See Appendices B and G.

The MSMT17 dataset was introduced in 2017 by researchers at Peking University and the University of Texas at San Antonio with the intention of providing a challenging test relative to existing benchmarks.[38] Because of its recency, as well as the fact that the best-performing attempts have tended to make use of some of the latest innovations in computer vision, we are confident that performance on the MSMT17 dataset is reflective of meaningful progress.[39] Several publications advanced performance on the MSMT17 dataset, of which we selected five.[40] The improvements in performance, as measured by mean Average Precision, have doubled since 2018. This is displayed in Figure 10. Researchers driving this progress correlate with those we would expect to see from our bibliometric analysis—three of the performance data points were created by China-affiliated researchers and organizations, while two were affiliated with the United States. This corroborates the insights derived from our bibliometric

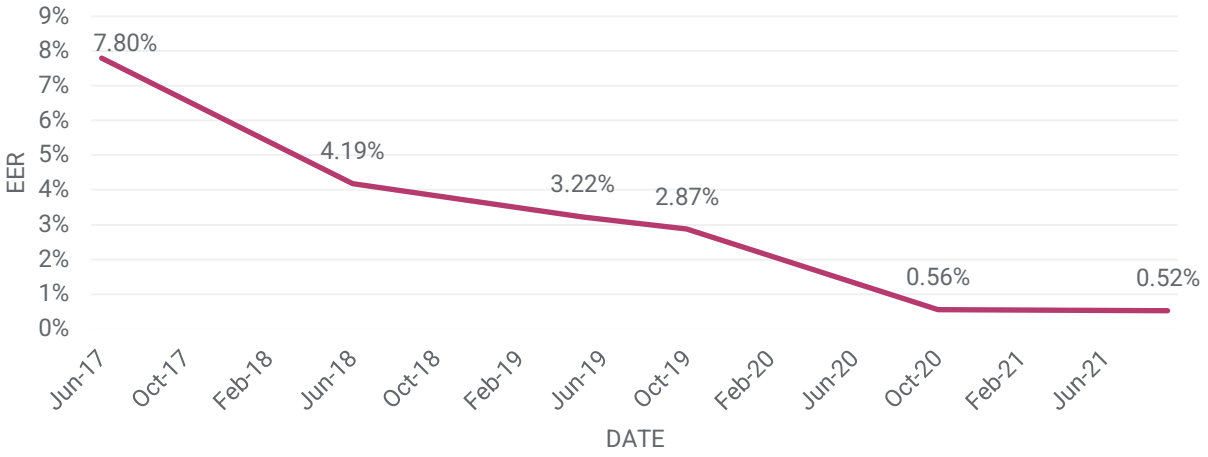
analysis of re-identification, which indicated that both are the most prominent countries affiliated with re-identification publications and citations.

Figure 10. Re-Identification Performance, 2018–2021

Source: See Appendix G.

The VoxCeleb dataset, which we used to assess improvements in speaker recognition, was introduced in 2017 by researchers at the University of Oxford.[41] Unlike MSMT17, VoxCeleb is also accompanied by an annual competition which sees multiple groups compete to test the performance of their systems against a standardized test. We selected six publications that advanced performance on the VoxCeleb dataset since 2017. These improvements, as measured by reductions in competitors' equal error rate, have been significant and sustained since 2017. This is displayed in Figure 11. However, as of 2020, the Equal Error Rates of models tested against VoxCeleb had limited room for improvement, so testers must design more difficult benchmarks to gauge performance of newer models. Improvements in state-of-the-art performance on VoxCeleb have primarily been driven by a research group at the University of Oxford (UK) that created the dataset and challenge, Ghent University (Belgium), and most recently SpeakIn Technologies (China). Compared to re-identification, these speaker recognition metrics correlated less with the distribution of the top country affiliations that were analyzed in the bibliometric analysis—the UK and Belgium were not amongst the most prominent countries affiliated with speaker recognition publications.

Figure 11. Speaker Recognition Performance, 2017–2021

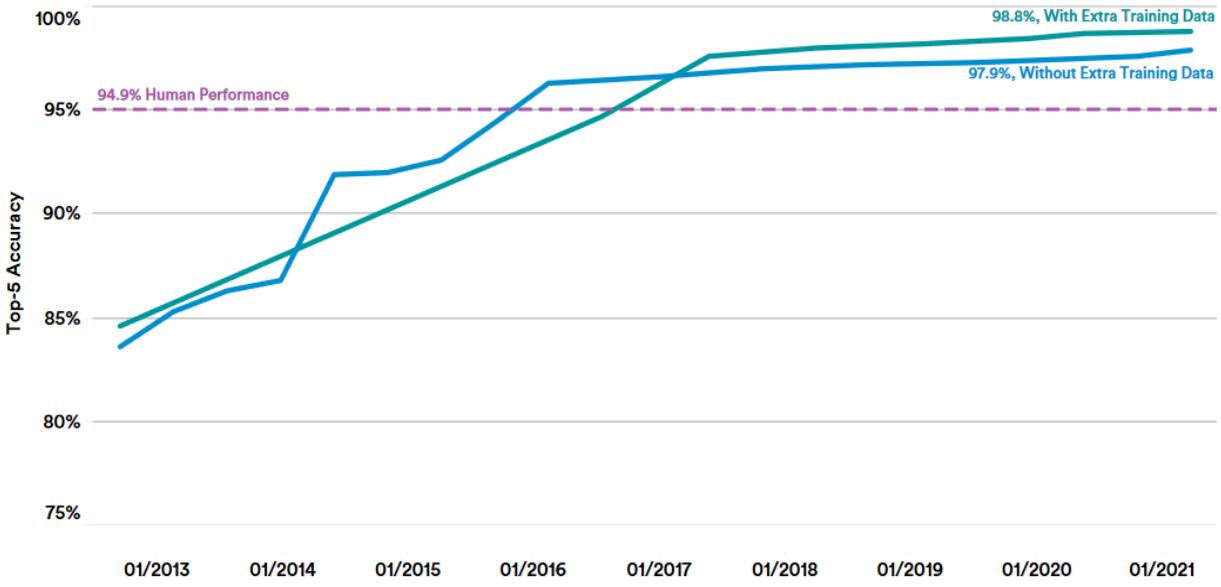


Source: See Appendix G.

In both research areas, we can see improvements between the period 2017–2020. Performance on the VoxCeleb dataset improved so dramatically—the last record-breaking model had only a 0.52 percent equal error rate—that researchers are currently trying to develop harder benchmarks to measure continued progress. These results indicate that the bibliometric findings of rapid growth in both research areas have in fact translated to major gains in performance.[42]

The performance improvements illustrated above are similar to those seen in other AI areas such as computer vision and natural language processing. As displayed in Figure 12, the performance of computer vision systems against ImageNet (a prominent image classification benchmark dataset used to test a range of models) advanced rapidly between 2013–2015, then exceeded human performance in 2016.[43] By 2018, improvement rates plateaued as it neared the upper stratum of performance (97–98 percent), with later performance increases coming from the addition of even larger datasets and more powerful computational resources. This is higher than re-identification performance (~69 mAP score), likely because re-identification is a more complicated task.

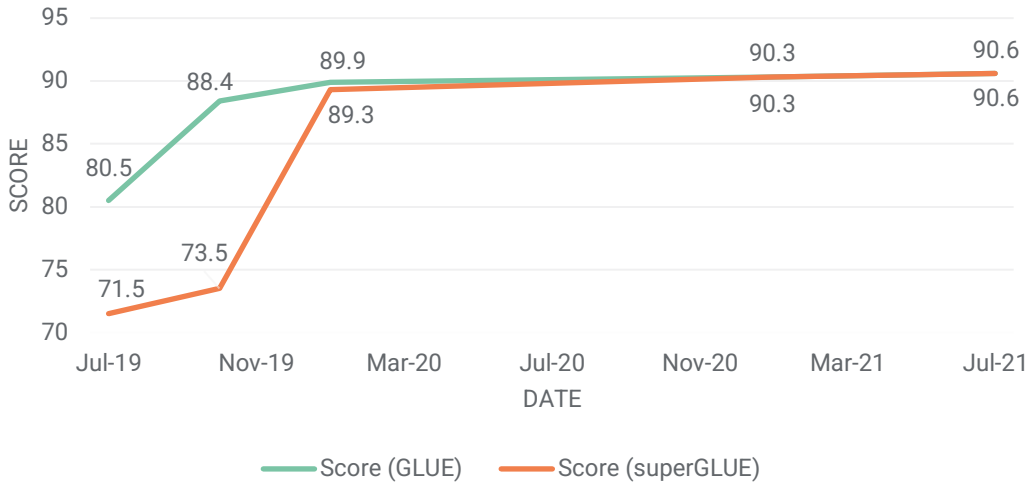
Figure 12. Computer Vision Performance, 2013–2021 (ImageNet Challenge: Top-5 Accuracy)



Source: Papers with Code, 2020; AI Index, 2021 | Chart: 2021 AI Index.

Natural language processing capabilities show similar improvements in two prominent benchmarks: "GLUE" and the updated "superGLUE."[44] These benchmarks only cover the years 2019–2021 because natural language processing only recently emerged as an area of rapid, competitive AI development using contemporary techniques.[45] GLUE was an attempt to create a benchmark that captured progress in natural language processing systems broadly by testing single systems against multiple suites of tests. However, as displayed in Figure 13, GLUE performance was rapidly saturated, forcing the creators to make superGLUE, a harder variant. Even this metric has become saturated recently. In this way, the replacement of benchmarks highlights the progress happening in the field and serves as a signifier for recent rapid progress.

Figure 13. Natural Language Processing Performance, 2019–2021



Source: Gluebenchmark "Leaderboard" and "Leaderboard Version: 2.0."

Leveraging the Methodology

The previous sections described a bibliometric and technical performance-based measurement approach that policymakers can use to understand changes in AI. A key question remains—how could the U.S. government leverage such a process to support policymaking? Determining the optimal degree of government involvement is outside the scope of this report, but we believe that it should, at the very least, play a leading role in organizing and resourcing private and academic entities that wish to participate, as well as steering research efforts toward areas relevant to policymakers. Notwithstanding the degree of involvement, we offer three overarching suggestions for how this type of methodology might be utilized and further improved with government support.

1. *Augment Technology Assessment with More Continuous Analysis*

Many attempts to measure AI involve periodic or ad-hoc assessments, rather than more continuous analyses. There has been a sustained surge in the publication of AI papers, and the popularity of preprint repositories like arXiv means the publishing cycles within this domain have accelerated. The methodology outlined in the previous sections can help the U.S. government take advantage of the information surfaced from these AI papers and provide a way to generate a more continuous stream of information about the state of particular AI subfields.

As an example for how this approach could complement ad-hoc assessments, consider NIST's Face Recognition Vendor Test (FRVT). NIST conducts periodic assessments of deployed facial recognition systems' state-of-the-art technical capabilities. To enhance these assessments, NIST could leverage the bibliometric and technical performance-based methodology outlined above to gain a more continuous understanding of facial recognition research, instead of relying solely on intermittent assessments. This process could be conducted biannually for bibliometrics to incorporate new publications and quarterly for performance metrics to incorporate new or improved benchmarks (in addition to

NIST-created benchmarks). Such an approach could also help policymakers understand the gap between capabilities being deployed in the world and those being demonstrated by researchers, and might provide an early warning system to flag major progress on a technical metric.

2. *Create Shared Infrastructure*

The bibliometrics analysis in the prior sections relies on CSET's Map of Science. But it is easy to imagine that a similar type of resource, housed within or resourced by the United States government, could become a useful tool for policymakers across multiple agencies. We could imagine such a tool being used by agencies such as the National Science Foundation, the National Institutes of Health, DARPA, and IARPA to surface trends in research activity. All these users will have different needs, but a shared infrastructure could allow multiple agencies to coordinate on areas of interest. It could also create multiple "demand signals" for the developers of such a bibliometric system. A shared platform that enables the collection, cleaning, and sharing of bibliometrics data provides an important resource for federal agencies and independent researchers to perform their own analysis.[46]

The performance metrics assessment in the prior section relied on the availability of common benchmarks. However, there are machine learning topics that lack reliable metrics altogether, and there is no single resource to readily attain existing metrics. This is an area that could benefit from a shared repository for capturing available metrics—just as bibliometrics are resourced into a single database (e.g., the Map of Science).[47] Such an infrastructure could be continuously updated to incorporate new metrics and identify short-lived or saturated benchmarks. The repository could also identify policy-relevant machine learning technologies that currently lack performance benchmarks. Candidates to support and maintain the repository include a government organization like NIST or a consortium of public and private participants. Ultimately, this repository could be integrated with a bibliometrics infrastructure, creating opportunities to apply methodologies like the one described in this report.[48]

3. *Enable Open-Source Analysis*

Governments periodically convene technical experts to provide advice on matters of technical change, opportunity, and hazard. The methodology outlined in this report provides a means to enable more nuanced policymaker-advisor communication and deeper insight. Governments could construct or support optimized research publication infrastructures, maintain or resource a database of performance metrics, pair it with periodic assessment of bibliometric trends, and then provide the system to technical experts. By supporting the construction and sustainment of such a tool and then sharing access to it with technical experts, governments could more easily receive expert insights on an as-needed basis, enable the development of low-cost public "insight competitions" where members of the public could compete to generate meaningful data from the tool, and share information with the public to provide a sense of progress in a given area.

Authors

Jack Clark is a non-resident research fellow with CSET, where Kyle A. Miller is a student research analyst, and Rebecca Gelles is a data scientist.

Acknowledgments

The authors are grateful to C. Raymond Perrault at SRI International, Kuansan Wang at Microsoft, and Robert Stojnic at Meta AI for reviewing the report and providing feedback and suggestions. We also thank our CSET colleagues John Bansemer and Catherine Aiken for their guidance and comprehensive reviews, Ben Buchanan (currently on leave), Igor Mikolic-Torreira, Micah Musser, and Drew Lohn for their support in the research and writing process, as well as Heeu Millie Kim for assistance with data editing. Melissa Deng and Alex Friedland provided editorial support.

Document Identifier: doi: 10.51593/20210008

Appendices

Appendix A: Methodology (Detailed)

Database—CSET's Map of Science (Bibliometric Clustering): The Map of Science is an aggregation of multiple bibliometrics datasets that encompasses most scientific literature and is updated quarterly to incorporate new publications and accrued citations, thus enabling regular assessments. The literature is grouped into research clusters, which are collections of research publications linked by citations. Clustering creates representations of related research, but the nature of the relationships within or between clusters is indeterminate and requires manual evaluation.[49] The database is considered curated because it calculates indicators that can inform analysis, such as a cluster's relation to AI, computer vision, and natural language processing. While the Map of Science is currently under development and has limitations, this report seeks to demonstrate its potential to automate and expedite the identification and extraction of publication metadata on specific machine learning–related topics. Metadata provides a range of information, including keywords and phrases, sources and citations, dates, authors, country and organization affiliations, publication mediums, and subject tags. Once fully developed, the database will likely include functions that automate much of the manual data structuring required in this report.

Sub-Method—Bibliometric Analysis: This sub-method is used to identify, extract, and analyze metadata from the Map of Science, enabling the measurement of trends in specific machine learning–related topics. The first step involves selecting a distinct topic to investigate (before using the Map of Science), framing what it encompasses, reviewing publications related to it on open-access repositories (e.g., arXiv), and collecting 10–20 "seed papers" that encapsulate it. This requires a degree of domain expertise, as the classification of "quality" seeds varies depending on the topic. Seed papers should include highly cited works from prominent authors that were published over different years and must encompass all key elements of the topic being investigated—which often requires "incorporating" any highly relevant terms or

topics into the investigation to establish sufficient distinction and framing.

Step two involves correlating the seed papers with clusters in the Map of Science, identifying the cluster or clusters that sufficiently encompass the topic, scanning neighboring clusters to corroborate topic distinction and provide context, omitting extraneous information, and then analyzing the extracted metadata for strategic insight. Adjustments can be made to the bibliometric analysis depending on the ambiguity of the topic, distribution of seed papers across clusters, quality of the data, or expertise of the user. For example, instead of collecting seed papers at the outset, non-experts could use keyword searches in the Map of the Science to identify relevant clusters, leverage automated data analytics functions in the Map of Science to identify prominent (e.g., highly cited) papers in the clusters, and then manually inspect the papers. This sub-method can be further optimized, such as by recruiting domain experts to tailor versions of bibliometric analysis for different machine learning research areas and policy-related inquiries, or incorporating natural language processing models to automate and scale the analysis of full-body publication text. These improvements are a vector for further research, as they are outside the scope of this report.

Sub-Method—Performance Metrics Assessment: This sub-method is used to assess the technical performance of a given machine learning–enabled system or model, corroborate the insight from bibliometric analysis, provide decisionmakers with information that relates to the capabilities of a particular aspect of machine learning, and more directly aid in identifying significant changes and advancements. The process revolves around manually tracking the technical performance metrics in publications (from clusters identified during the bibliometric analysis). Metrics are designed by researchers to assess particular machine learning systems and solve specific problems, and are often released publicly to be used and updated by the community. This allows us to select a distinct topic for investigation, manually collect and compare particular metrics in different publications across repositories like arXiv and Papers with Code, chart a timeline of

performance development, and assess current performance benchmarks—if the metrics are available and reliable. A performance assessment works well when there are existing benchmarks with track records of testing and evaluation, but can be more challenging if there are no relevant metrics available or they are outdated (i.e., saturated). However, identifying topics that lack metrics can help illuminate knowledge gaps in policy-relevant areas that require further research. Although this process requires extensive domain expertise and manual inspection of publications, it could be more streamlined if a maintained database of categorized metrics was available.

Appendix B: Seed Papers

Re-Identification Seed Papers:

- Rodolfo Quispe and Helio Pedrini, "Top-DB-Net: Top DropBlock for Activation Enhancement in Person Re-Identification," arXiv preprint arXiv:2010.05435v1 (2020), https://arxiv.org/abs/2010.05435v1.
- Xingyang Ni, Liang Fang, and Heikki Huttunen, "AdaptiveReID: Adaptive L2 Regularization in Person Re-Identification," arXiv preprint arXiv:2007.07875v1 (2020), https://arxiv.org/abs/2007.07875v1.
- Changxing Ding et al., "Multi-task Learning with Coarse Priors for Robust Part-aware Person Re-identification," arXiv preprint arXiv:2003.08069v1 (2020), https://arxiv.org/abs/2003.08069v1.
- S.V. Aruna Kumar et al., "P-DESTRE: The P-DESTRE: A Fully Annotated Dataset for Pedestrian Detection, Tracking, Re-Identification and Search from Aerial Devices," arXiv preprint arXiv:2004.02782 (2020), https://arxiv.org/abs/2004.02782.
- Jianyang Gu et al., "Long-Term Cloth-Changing Person Re-identification," arXiv preprint arXiv:2012.13498 (2020), https://arxiv.org/abs/2005.12633.
- Jianyang Gu et al., "1st Place Solution to VisDA-2020: Bias Elimination for Domain Adaptive Pedestrian Re-identification," arXiv preprint arXiv:2012.13498 (2020), https://arxiv.org/abs/2012.13498.
- Ammarah Farooq et al., "A Convolutional Baseline for Person Re-Identification Using Vision and Language Descriptions," arXiv preprint arXiv:2003.00808 (2020), https://arxiv.org/abs/2003.00808.
- Zhedong Zheng et al., "Joint Discriminative and Generative Learning for Person Re-Identification," arXiv preprint

arXiv:1904.07223v1 (2019), https://arxiv.org/abs/1904.07223v1.

- Guan'an Wang et al., "Color-Sensitive Person Re-Identification," 2019, https://www.ijcai.org/Proceedings/2019/131.
- Hui Li et al., "Pedestrian re-identification based on Tree branch network with local and global learning," arXiv preprint arXiv:1904.00355 (2019), https://arxiv.org/abs/1904.00355.
- Shi-Zhe Chen, Chun-Chao Guo, and Jian-Huang Lai, "Deep Ranking for Person Re-Identification via Joint Representation Learning," arXiv preprint arXiv:1505.06821 (2016), https://arxiv.org/abs/1505.06821.
- Brian H. Wang et al., "Deep Person Re-identification for Probabilistic Data Association in Multiple Pedestrian Tracking," arXiv preprint arXiv:1810.08565 (2018), https://arxiv.org/abs/1810.08565.
- Jianfu Zhang, Naiyan Wang, and Liqing Zhang, "Multi-shot Pedestrian Re-identification via Sequential Decision Making," arXiv preprint rXiv:1712.07257v1 (2017), https://arxiv.org/abs/1712.07257v1.
- Ejaz Ahmed, Michael Jones, and Tim K. Marks, "An Improved Deep Learning Architecture for Person Re-identification," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, https://ieeexplore.ieee.org/document/7299016.

Speaker Recognition Seed Papers:

- Xugang Lu, Peng Shen, Yu Tsao, and Hisashi Kawai, "Integrating a Joint Bayesian Generative Model in a Discriminative Learning Framework for Speaker Verification," arXiv preprint arXiv:2101.03329 (2021), https://arxiv.org/abs/2101.03329.

- Jixuan Wang et al., “Speaker Attribution with Voice Profiles by Graph-based Semi-supervised Learning,” arXiv preprint arXiv:2102.03634 (2021), https://arxiv.org/abs/2102.03634.
- Wei Yao et al., “Multi-stream Convolutional Neural Network with Frequency Selection for Robust Speaker Verification,” Version 2, arXiv preprint arXiv:2012.11159v2 (2021), https://arxiv.org/abs/2012.11159v2.
- Wei Yao et al., “Multi-stream Convolutional Neural Network with Frequency Selection for Robust Speaker Verification,” Version 1, arXiv preprint arXiv:2012.11159v1 (2020), https://arxiv.org/abs/2012.11159v1.
- Arsha Nagrani et al., “Voxceleb: Large-scale speaker verification in the wild,” October 2019, https://www.robots.ox.ac.uk/~vgg/publications/2019/Nagrani19/nagrani19.pdf.
- Hemlata Tak et al., “End-to-end anti-spoofing with RawNet2,” arXiv preprint arXiv:2011.01108 (2020), https://arxiv.org/abs/2011.01108.
- Deep CNNs With Self-Attention for Speaker Identification, (2019), https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8721628.
- VoxCeleb2: Deep Speaker Recognition, (2018), https://www.robots.ox.ac.uk/~vgg/publications/2018/Chung18a/chung18a.pdf.
- Weicheng Cai, Jinkun Chen, and Ming Li, “Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System,” arXiv preprint arXiv:1804.05160 (2018), https://arxiv.org/abs/1804.05160.
- David Snyder et al., “X-Vectors: Robust DNN Embeddings for Speaker Recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, https://www.semanticscholar.org/paper/X-

Vectors%3A-Robust-DNN-Embeddings-for-Speaker-Snyder-Garcia-Romero/389cd9824428be98a710f5f4de67121a70c15fd3.

- Suwon Shon, Hao Tang, and James Glass, "Frame-Level Speaker Embeddings for Text-Independent Speaker Recognition and Analysis of End-to-End Model," arXiv preprint arXiv:1809.04437 (2018), https://arxiv.org/abs/1809.04437.

- D Snyder et al., "Deep Neural Network Embeddings for Text-Independent Speaker Verification," (2017), INTERSPEECH, http://refhub.elsevier.com/S0885-2308(19)30271-2/sbref0055.

- Gautam Bhattacharya, Jahangir Alam, and Patrick Kenny, "Deep Speaker Embeddings for Short-Duration Speaker Verification," INTERSPEECH, August 2017, https://www.isca-speech.org/archive/Interspeech_2017/pdfs/1575.PDF.

- Mitchell McLaren et al., "The Speakers in the Wild (SITW) Speaker Recognition Database," INTERSPEECH, 2016, https://www.semanticscholar.org/paper/The-Speakers-in-the-Wild-(SITW)-Speaker-Recognition-McLaren-Ferrer/3fe358a66359ee2660ec0d13e727eb8f3f0007c2.

- L. Burget et al, "Discriminatively Trained Probabilistic Linear Discriminant Analysis for Speaker Verification," *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, https://www.semanticscholar.org/paper/0cf0a9f92e252f10bb31a56da19fbe3b71bf1151.

Image Synthesis Seed Papers:

- Simranjeet Singh, Rajneesh Sharma, and Alan F. Smeaton, "Using GANs to Synthesise Minimum Training Data for Deepfake Generation," arXiv preprint arXiv:2011.05421 (2020), https://arxiv.org/abs/2011.05421.

- Shilong Shen, “Correspondence Learning for Controllable Person Image Generation,” arXiv preprint arXiv:2012.12440 (2020), https://arxiv.org/abs/2012.12440.
- Janet Rafner et al., “The Power of Pictures: Using ML Assisted Image Generation to Engage the Crowd in Complex Socioscientific Problems,” arXiv preprint arXiv:2010.12324 (2020), https://arxiv.org/abs/2010.12324.
- Xiangrui Xu, Yaqin Li, and Cao Yuan, “Conditional Image Generation with One-Vs-All Classifier,” arXiv preprint arXiv:2009.08688 (2020), https://arxiv.org/abs/2009.08688.
- Thanh Thi Nguyen et al., “Deep Learning for Deepfakes Creation and Detection: A Survey,” arXiv preprint arXiv:1909.11573v1 (2019), https://arxiv.org/abs/1909.11573v1.
- Sakshi Agarwal and Lav R. Varshney, “Limits of Deepfake Detection: A Robust Estimation Viewpoint, arXiv preprint arXiv:1905.03493 (2019), https://arxiv.org/abs/1905.03493.
- Andrew Brock, Jeff Donahue, and Karen Simonyan, “Large Scale GAN Training for High Fidelity Natural Image Synthesis,” arXiv preprint arXiv:1809.11096 (2019), https://arxiv.org/abs/1809.11096.
- Jun-Yan Zhu et al., “Visual Object Networks: Image Generation with Disentangled 3D Representation,” arXiv preprint arXiv:1812.02725 (2018), https://arxiv.org/abs/1812.02725.
- Pavel Korshunov and Sebastien Marcel, “DeepFakes: A New Threat to Face Recognition? Assessment and Detection,” arXiv preprint arXiv:1812.08685 (2018), https://arxiv.org/abs/1812.08685.
- Mevlana Gemici, Zeynep Akata, and Max Welling, “Primal-Dual Wasserstein GAN,” arXiv preprint arXiv:1805.09575 (2018), https://arxiv.org/abs/1805.09575.

- Alexey Dosovitskiy and Thomas Brox, “Generating Images with Perceptual Similarity Metrics based on Deep Networks,” arXiv preprint arXiv:1602.02644 (2016), https://arxiv.org/abs/1602.02644.
- Alec Radford, Luke Metz, and Soumith Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,” arXiv preprint arXiv:1511.06434 (2015), https://arxiv.org/abs/1511.06434.

Appendix C: Notes on Terminology

Re-Identification Terminology: Researchers use a variety of terms to discuss re-identification (RE-ID), as is the case with many emerging machine learning–enabled capabilities and systems. Due to this varying taxonomy, we must "incorporate" any highly relevant terms or topics into the investigation to maintain sufficient distinction and framing. In addition to the term RE-ID, we incorporated seven related terms into the investigation: Person RE-ID, Pedestrian RE-ID, Object RE-ID, Vehicle RE-ID, Human RE-ID, Pedestrian Tracking, and Pedestrian Retrieval. For the context of this report, we consider all of them synonymous and refer to them collectively through the term "re-identification." Additionally, we must note that RE-ID is distinguished from "gait recognition," which is the task of identifying individuals based on their unique manner of moving and walking (although both capabilities can be integrated). Domain expertise and an analysis of prominent RE-ID papers informed the selection of these terms.

Table 7. Re-Identification Cluster #1419 Topic Relevance Spot-Check (via Keywords in Publication Titles)

RE-ID (Cluster #1419) Terminology	Keyword Search Matches
Re-Identification	2,563
Person Re-Identification	2,128
Pedestrian Re-Identification	72
Object Re-Identification	5
Vehicle Re-Identification	157
Human Re-Identification	25
Pedestrian Tracking	4
Pedestrian Retrieval	9
Total Matches	4,963

Source: CSET Map of Science, as of March 2021.

Speaker Recognition Terminology: In addition to the term speaker recognition (SR), we incorporated four related labels into the investigation: Speaker Verification, Speaker Identification, Voice Identification, and Speaker Diarization. We must note that there are specific elements that distinguish these terms. "SR" broadly relates to identifying speakers, "Speaker Verification" is verifying a speaker identity that was previously learned, "Speaker Identification" is identifying an unknown speaker that is in a group of known speakers, "Voice Identification" is identifying a known voice, and "Speaker Diarization" is a process for partitioning audio to identify who spoke when.[50] For the context of this report, we refer to all these labels collectively through the term "SR." Additionally, we must note that SR is distinguished from "speech recognition," which is the task of identifying and converting words from audio to text. As with RE-ID terminology, domain expertise and the analysis of SR papers informed the selection of these terms.

Table 8. Speaker Recognition Cluster #6855 Topic Relevance Spot-Check (via Keywords in Publication Titles)

SR (Cluster #6855) Terminology	Keyword Search Matches
Speaker Recognition	539
Speaker Verification	653
Speaker Identification	97
Voice Identification	0
Speaker Diarization	46
Total Matches	1,335

Source: CSET Map of Science, as of March 2021.

Appendix D: Progression of Key Machine Learning Technologies

Re-Identification Progression: The early years of 2010–2013 had steady growth rates, but the overall quantity of papers and citations was rather limited. During that period the key components (e.g., cost and availability of data and computational capacity) and model architectures (e.g., deep learning approaches to computer vision) for re-identification were in their technological infancy. Significant advancements in machine learning computer vision occurred around 2012, which was followed by increased growth in re-identification publications between 2015–2018. This six-year period saw the discovery and proliferation of many key machine learning subcomponents that enable re-identification, including deep convolutional neural networks (CNN), residual networks, and general techniques like pre-training, where you train a general-purpose computer vision neural network on a large amount of generic data, then fine-tune it on more specific datasets. Following these advancements, the research community went from producing dozens of re-identification papers over multiple years to hundreds annually (although growth started to plateau in 2020).[51]

Speaker Recognition Progression: The early years of 2010–2014 had steady growth rates, but the overall quantity of papers and citations was limited, a dynamic that parallels re-identification development. Many of the same machine learning advancements that enabled the proliferation of re-identification also enabled speaker recognition, including deep CNNs and increased data and compute capacity. Developments in natural language processing learning techniques, such as Transformers, played a significant role in the accelerated speaker recognition growth after 2017. During this timeframe, the research community went from producing dozens of speaker recognition papers to producing hundreds annually. Alongside, and intertwined with, these advancements was the proliferation of "smart" devices (e.g., smartphones), programs (e.g., Apple's Siri), and services (e.g., audio recording) that could use speaker recognition.[52]

Appendix E: Organizational Affiliation Data (Detailed)

Re-Identification Organizational Affiliations: There were over 600 organizations affiliated with an author of at least one re-identification-related publication, of which approximately 76 percent were in the education sector, 9 percent in government, and 6 percent in private industry. We identified six "primary organizations," meaning they were ranked in the top ten for both publication and citation count: (1) Chinese Academy of Sciences, (2) Queen Mary University of London, (3) Sun Yat-sen University, (4) University of Technology, Sydney, (5) Peking University, and (6) Tsinghua University. All these entities are educational and highly affiliated with their respective governments.

China was home to the most organizations affiliated with the authors of re-identification papers in our cluster. Of the top 50 organizations affiliated with re-identification publications, 35 were located in China, five in the United States, and two in both Australia and Italy. Of the top 50 cited organizations, 22 were located in China, twelve in the United States, three in both France and Italy, and two in Australia, Austria, and Singapore.

Most of the top organizations were public research universities that have significant government affiliations, along with varying affiliations with private entities. For example, the Chinese Academy of Sciences is a national academy fully supported by Chinese government entities. When we focus on the rankings amongst private industry organizations, the data indicates that Microsoft (United States), Alibaba Group (China), Tencent (China), Vision Semantics Limited (UK), and SenseTime (China) were the most affiliated with re-identification papers. Notably, Microsoft conducts much of its re-identification-related research through the Beijing-based "Microsoft Research Asia," the company's largest research lab outside of the United States. This information can aid in a range of governmental regulatory and collaborative efforts, as well as inform security considerations regarding the international proliferation of technologies designed by domestic companies (e.g., Microsoft).

Speaker Recognition Organizational Affiliations: There were over 400 organizations affiliated with an author of at least one speaker recognition publication, of which approximately 65 percent were in the education sector, 11 percent in private industry, 5 percent in government, and 4 percent non-profit. This is a similar dynamic to re-identification, although with fewer total organizations and greater private affiliation. Speaker recognition also has different primary organizations ranked in the top 10 for both paper affiliations and citations of affiliated papers: (1) Johns Hopkins University, (2) MIT, (3) Brno University of Technology, and (4) SRI International. In contrast to re-identification, speaker recognition primary organizations were a mix of private, public, and nonprofits, most of which are not directly affiliated with their respective governments.

The geographic distribution of speaker recognition papers was more heterogeneous than re-identification. Of the top 50 organizations affiliated with authors of speaker recognition papers, 13 were located in the United States, 11 in China, three in Canada, and two in Switzerland, Spain, South Korea, Singapore, India, Australia, and Argentina. Of the top 50 organizations ranked by the citation counts of their affiliated papers, 17 were in the United States, six in China, four in Spain and Canada, and three in the UK.

Most of the top organizations were public universities with significant government affiliations, notwithstanding outliers such as the private Johns Hopkins University (most publications) and the nonprofit Computer Research Institute of Montreal (most-cited). For example, Tsinghua University, the University of Texas at Dallas, and Hong Kong Polytechnic University were each public universities in the top ten for speaker recognition publications. When focused on rankings among private industry organizations, we see that IBM, Microsoft, NEC Corporation, Tencent, and Google were the most affiliated with speaker recognition.

Appendix F: Metrics Assessment (Detailed)

Re-Identification Metrics Assessment: The assessment of re-identification concentrates on performance against the MSMT17 dataset using the mean Average Precision (mAP) metric. We chose this dataset because: it is relatively new (introduced in 2017 by researchers at Peking University and the University of Texas at San Antonio), has not yet been saturated, involves some of the latest techniques, and at the time of its creation was designed to be challenging relative to existing benchmarks.[53] Therefore, we are confident that the performance on MSMT17 is reflective of real progress, rather than progress on an easy benchmark.

Five publications informed our assessment of re-identification advancement between 2018–2021 (see Appendix G for details), four of which were affiliated with China, two with the United States, and one with both France and Australia. These were chosen because the researchers tested their systems against MSMT17.

We extracted two insights from the progress of re-identification models' mAP performance against the MSMT17 dataset:

1. Re-identification performance has improved significantly, with a high rate of progress between 2018–2019 but slower improvements between 2020–2021. Notably, the metric improvements correlated with the increased rate of re-identification publication, as seen through the bibliometric metadata analysis. Finding out where further improvements will come from could be important, as subsequent investigations of yet-to-exist results will tell us whether performance growth is stagnating or if re-identification is able to successfully harness further breakthroughs from computer vision writ large to further expand performance (both in test environments and its real-world application). This insight warrants attention from policymakers as a continuation of this trend will see re-identification systems advance to a degree that makes them more efficient and reliable—increasing incentives for both government and private entities to adopt the technology.

2. Researchers driving this progress correlate with those we would expect to see from our bibliometric analysis—four of the performance data points were created by China-affiliated researchers and organizations, while two were affiliated with the United States. This corroborates the insights derived from our bibliometric analysis of re-identification, which indicated that both are the most prominent countries affiliated with re-identification publications and citations. This also suggests that studying technical metrics can "sanity check" the insights derived from bibliometrics. However, a more comprehensive application of the performance metrics assessment is necessary to determine the extent to which metrics affiliations correlate with the distribution of country affiliations seen in the bibliometrics metadata.

Speaker Recognition Metrics Assessment: The assessment of speaker recognition concentrates on performance against the VoxCeleb dataset using the Equal Error Rate (EER) metric. We chose VoxCeleb because, like MSMT17, it is relatively new (introduced in 2017 by researchers at the University of Oxford). Unlike MSMT17, VoxCeleb is also accompanied by an annual competition that sees multiple groups compete to test the performance of their systems against a standardized test (in many ways, VoxCeleb is to speaker recognition what ImageNet is to object detection)—this gives us a pre-structured way to gather data for the technical analysis and provides some inbuilt auditing of technical results by competition administrators.

Six publications informed our assessment of speaker recognition advancement between 2017–2020 (see Appendix G for details), four of which were affiliated with the UK and one with both Belgium and China. These were chosen because they were the highly ranked solutions for the VoxCeleb competitions.

We extracted two insights from the progress of speaker recognition models' EER performance against the VoxCeleb dataset:

1. Speaker recognition performance has improved significantly, with the EER decreasing (i.e., increasing model precision) from ~8 percent in 2017 to ~0.5 percent in 2021. This performance is so good that VoxCeleb evaluations have, in recent years, moved to testing against harder and more complicated underlying datasets. This reflects a similar pattern of development that occurred in computer vision where, after results started to stagnate on ImageNet, researchers developed harder variants of the dataset to test against (e.g., ImageNet Adversarial).[54] Notably, some of these improvements occurred alongside the development of key natural language processing learning techniques (e.g., using Transformers) and correlated with the increase in speaker recognition publications, as seen through our bibliometric analysis. This insight warrants attention from policymakers, as a continuation of this trend will see speaker recognition systems advance to a degree that could supplement intelligence capabilities in profound ways (e.g., the recognition of millions of speakers in perpetuity).

2. Improvements in state-of-the-art performance on VoxCeleb have primarily been driven by a research group at the University of Oxford (UK) that created the dataset and challenge, Ghent University (Belgium), and most recently SpeakIn Technologies (China). Compared to re-identification, these speaker recognition metrics did not correlate as much with the distribution of the top country affiliations that were analyzed in the bibliometric analysis—the UK and Belgium were not amongst the most prominent countries affiliated with authors of speaker recognition publications. This is an area for further analysis, as it neither corroborates nor invalidates the insight from publication metadata. However, it does illustrate how metadata alone cannot capture important technical elements of machine learning research and further highlights how pairing technical metrics with bibliometric analysis can "sanity check" the findings of different analyses.

Further metrics assessment is needed to better determine country affiliation. First, the results might superficially indicate that the UK and Belgium are leading performance on speaker recognition applied to VoxCeleb1, but if we analyze more results and adopt a cohort-based analysis technique—that is, look at the people that came in second and third place in the VoxCeleb challenge each year—we can get a sense of the broader landscape. Here, teams associated with the United States and China appeared within the competition as well. This suggests that, although a performance assessment of state-of-the-art speaker recognition models can yield insight, a more comprehensive application of the metrics assessment is necessary to get a clearer signal and context. Second, researchers from various countries may be testing out their systems against VoxCeleb without publishing the results, due to the commercial and security value of speaker recognition capabilities.

Appendix G: Performance Metrics Data and Sources

Table 9. Re-Identification Performance Metrics Data and Sources

Date	mAP	Method	Organization(s)	Country Affiliation	Publication Title
3/26/2021	69.4	TransReID	Alibaba Group; Zhejiang University	China	TransReID: Transformer-based Object Re-Identification
2/18/2021	67.3	Deep Miner	Digeiz AI Lab; Ecole Polytechnique	France	Deep Miner: A Deep and Multi-branch Network which Mines Rich and Diverse Features for Person Re-identification
9/21/2020	62.7	Multi-task Part-aware Network (MPN)	South China University of Technology; University of Sydney	China; Australia	Multi-task Learning with Coarse Priors for Robust Part-aware Person Re-identification
8/9/2019	60.8	ABD-Net	Texas A&M University; University of Science and Technology of China; Walmart Technology; Wormpex AI Research	United States; China	ABD-Net: Attentive but Diverse Person Re-Identification
9/13/2018	34	GLAD	Peking University, University of Texas at San Antonio	China; United States	Person Transfer GAN to Bridge Domain Gap for Person Re-Identification

Source: Chuting He et al.; Abdallah Benzine et al.; Changxing Ding et al.; Tianlong Chen et al.; Wei et al. [55]

Table 10. Speaker Recognition Performance Metrics Data and Sources

Date	EER	Method	Organization	Country Affiliation	Publication Title
9/5/2021	0.52%	QMF	SpeakIn Technologies Co. Ltd.	China	The SpeakIn System for VoxCeleb Speaker Recognition Challenge 2021
10/20/2020	0.56%	Speech duration QMF	Ghent University	Belgium	The IDLAB VoxSRC-20 Submission: Large Margin Fine-Tuning and Quality-Aware Score Calibration in DNN Based Speaker Verification
10/16/2019	2.87%	Ours + Relation Module	University of Oxford	UK	Voxceleb: Large-Scale Speaker Verification in the Wild
5/17/2019	3.22%	GhostVLAD	University of Oxford	UK	Utterance-level Aggregation for Speaker Recognition In The Wild
6/27/2018	4.19%	ResNet50	University of Oxford	UK	VoxCeleb2: Deep Speaker Recognition
6/26/2017	7.80%	CNN256D-Embedding	University of Oxford	UK	VoxCeleb: A Large-Scale Speaker Identification Dataset

Note: A lower EER score signifies less error and greater accuracy.
Source: Nagrani et al.; Joon Son Chung, Arsha Nagrani, and Andrew Zisserman; Weidi Xie et al.; Arsha Nagrani et al.; Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck; Miao Zhao et al.[56]

Endnotes

[1] For more literature on the field of bibliometrics, see the following sources: Office of Management, "Learn More About Bibliometrics," National Institutes of Health (NIH), https://www.nihlibrary.nih.gov/services/bibliometrics/learn-more-about-bibliometrics; Ashok Agarwal et al., "Bibliometrics: Tracking Research Impact by Selecting the Appropriate Metrics," *Asian Journal of Andrology* 18, no. 2 (January 2016): 296-309, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4770502/; Anthony van Raan, "Measuring Science: Basic Principles and Application of Advanced Bibliometrics," in *Springer Handbook of Science and Technology* Indicators (2019): 237-280, https://link.springer.com/chapter/10.1007/978-3-030-02511-3_10; Bibliometrics involves the "quantitative evaluation of scientific articles and other published works, including the authors of articles, the journals where the works were published, and the number of times they are later cited," A.W. Jones, "Forensic Journals: Bibliometrics and Journal Impact Factors," *Encyclopedia of Forensic and Legal Medicine* (Second Edition) (2016), https://www.sciencedirect.com/science/article/pii/B9780128000342001816#.

[2] This is not an attempt to forecast future developments, an in-depth analysis of a particular aspect of AI, nor a comprehensive review or critique of other research efforts to measure AI. Rather, it is an outline of a systematic process to measure AI developments and a demonstration that we hope will inspire policymakers to implement and improve the methodology. Additionally, though none of the elements of the methodology are novel, by integrating them together we can outline a process to improve measurement.

[3] National Institute for Standards and Technology (NIST), "The EMNIST Dataset," U.S. Department of Commerce, https://www.nist.gov/itl/products-and-services/emnist-dataset; Defense Advanced Research Projects Agency (DARPA), "The DARPA Grand Challenge: Ten Years Later," March 13, 2014, https://www.darpa.mil/news-events/2014-03-13; Defense Advanced Research Projects Agency (DARPA), "DARPA Robotics Challenge (DRC)," https://www.darpa.mil/program/darpa-robotics-challenge.

[4] The efforts of this report are similar to parts of the Intelligence Advanced Research Projects Activity's (IARPA) Foresight and Understanding from Scientific Exposition (FUSE) program. The FUSE program ran from 2010–2017 and was intended to "enable reliable, early detection of emerging scientific and technical capabilities across disciplines and languages found within the full-text content of scientific, technical, and patent literature" and "discover patterns of emergence and connections between technical concepts at a speed, scale, and comprehensiveness that exceeds human capacity." Like this report, FUSE emphasized the need for a methodology that incorporated continuous assessment of bibliometrics over ad-hoc assessment. However, the FUSE

program intended to design a fully refined and automated methodology that involved six "research thrusts" (theory development, document features, indicator development, nomination quality, evidence representation, and system engineering), while this report's objectives are far more limited, as we only intend to demonstrate the potential for insight through a prototype methodology that can be adopted and updated by different end users. Additionally, the methodology in this report incorporates performance metrics assessment and an AI-curated corpus of scientific literature, which are elements not included in IARPA's FUSE program. For more details, see Intelligence Advanced Research Projects Activity (IARPA), "Foresight and Understanding from Scientific Exposition (FUSE)," Office of the Director of National Intelligence, https://www.iarpa.gov/index.php/research-programs/fuse.

[5] The CSET Map of Science is a merged corpus of Digital Science's Dimensions, Clarivate's Web of Science, Microsoft Academic Graph (MAG), China National Knowledge Infrastructure (CNKI), arXiv, and Papers with Code. As of September 2021, it included approximately 130 million papers, 1.4 billion citation linkages, and 123 thousand research clusters. See also the Map of Science user interface, Jennifer Melot and Ilya Rahkovsky, "CSET Map of Science" (Center for Security and Emerging Technology, October 2021), https://cset.georgetown.edu/publication/cset-map-of-science/. We must note that the public user interface was released in October 2021 but the data in this report was extracted from the Map of Science in March 2021, therefore the clusters addressed in this report do not reflect clusters in the current iteration of the Map of Science; Autumn Toney, "Creating a Map of Science and Measuring the Role of AI in it" (Center for Security and Emerging Technology, June 2021), https://cset.georgetown.edu/publication/creating-a-map-of-science-and-measuring-the-role-of-ai-in-it//.

[6] For more information on the methodology behind CSET's research clusters, see Ilya Rahkovsky et al., "AI Research Funding Portfolios and Extreme Growth" (Center for Security and Emerging Technology, April 2021), https://www.frontiersin.org/articles/10.3389/frma.2021.630124/full; For an explanation of the methodology for classifying papers as AI-relevant, see James Dunham, Jennifer Melot, and Dewey Murdick, "Identifying the Development and Application of Artificial Intelligence in Scientific Text," arXiv preprint arXiv:2002.07143 (2020), https://arxiv.org/abs/2002.07143; For more information on clustering methodology and the clustering model, see Kevin W. Boyack, Caleb Smith, and Richard Klavans, "A detailed open access model of the PubMed literature," *Scientific Data* 7 (November 2020), https://www.nature.com/articles/s41597-020-00749-y; For a "Data Snapshot" that explores the underlying data and analytic utility of the CSET Map of Science, see Toney, "Creating a Map of Science and Measuring the Role of AI in it."

[7] We primarily use the term "machine learning" instead of "artificial intelligence" for the remainder of the report because AI is too broad (e.g., this investigation does not encompass expert systems).

[8] There are preliminary systems being built that help automate this, ranging from expert-led syntheses of the field (e.g., the AI Index) to community-curated resources aided by automated extraction of performance metrics from papers (e.g., Papers with Code). Additionally, CSET is currently working to extract more data from papers and clusters in the Map of Science, including but not limited to compute use, libraries use, source code, and datasets—all of which could further automate and improve the methodology presented in this report.

[9] "ImageNet," https://www.image-net.org/.

[10] Mang Ye et al., "Deep Learning for Person Re-identification: A Survey and Outlook," arXiv preprint arXiv:2001.04193 (2020), https://arxiv.org/abs/2001.04193.

[11] Re-identification systems can track individuals in anonymous and non-anonymous capacities. There is also a growing field of research into unsupervised re-identification—that is, systems that automatically identify and track individuals seen in streaming video data without needing a prior label and image.

[12] More specifically, speaker recognition is an emerging voice biometry capability that is enabled by deep machine learning models trained on labeled and unlabeled audio data, while speaker recognition systems are designed to recognize the voices of specific speakers in an audio sample or live feed. The speaker recognition process is first focused on training the model through ingesting labeled data, such as the voice audio of a speaker that is linked to said speaker's name or some other identifier. This process "enrolls" a speaker's voice into a unique signature that the model can recognize. Once speaker signatures are learned to an ascertainable degree, speaker recognition systems can then be deployed in a variety of capacities to identify audio inputs and correlate them to known individuals. For more information, see "What is Speaker Recognition (Preview)?" Microsoft, November 2021, https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/speaker-recognition-overview.

[13] Note that speaker recognition differs from speech recognition, which focuses on identifying and converting words from audio to text.

[14] J.J. Godfrey, E.C. Holliman, and J. McDaniel, "Switchboard: Telephone Speech Corpus for Research and Development," *ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1992, https://www.semanticscholar.org/paper/SWITCHBOARD%3A-telephone-

speech-corpus-for-research-Godfrey-Holliman/d80000d84223e177d070a01a734dba56d5f5c069.

[15] More specifically, image synthesis is a matured computer vision capability enabled by deep generative machine learning models and frameworks (e.g., GANs and variational autoencoders) that are trained on large imagery datasets. It is a relatively broad term that encompasses many techniques and systems for creating computer-generated images. The general purpose of this capability contrasts with re-identification and speaker recognition, which involves the discrimination, identification, and recognition of phenomena—not the creation of synthetic media.

[16] A GAN is an architecture of two "competing" neural networks, a generator and a discriminator, that can be used for many purposes. In image synthesis, the model is trained on labeled and unlabeled imagery data, which is fed into a generator neural network that extracts features and learns to generate synthetic images based on them. These generated images are iteratively tested against a second neural network that is designed to discriminate between synthetic and organic media and acts as a feedback mechanism to optimize the generator.

[17] Sarah Cahlan, "How Misinformation Helped Spark an Attempted Coup in Gabon," *The Washington Post*, February 13, 2020, https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/.

[18] Charlotte Jee, "An Indian Politician Is Using Deepfake Technology to Win New Voters," *MIT Technology Review*, February 19, 2020, https://www.technologyreview.com/2020/02/19/868173/an-indian-politician-is-using-deepfakes-to-try-and-win-voters/; Benjamin Strick, "West Papua: New Online Influence Operation Attempts to Sway Independence Debate," *Bellingcat*, November 11, 2020, https://www.bellingcat.com/news/2020/11/11/west-papua-new-online-influence-operation-attempts-to-sway-independence-debate/.

[19] Ian J. Goodfellow et al., "Generative Adversarial Networks," arXiv preprint arXiv:1406.2661 (2014), https://arxiv.org/abs/1406.2661; Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," arXiv preprint arXiv:1511.06434 (2015), https://arxiv.org/abs/1511.06434; Ming-Yu Liu and Oncel Tuzel, "Coupled Generative Adversarial Networks," arXiv preprint arXiv:1606.07536 (2016), https://arxiv.org/abs/1606.07536; Tero Karras et al., "Progressive Growing of GANs for Improved Quality, Stability, and Variation," arXiv preprint arXiv:1710.10196 (2017), https://arxiv.org/abs/1710.10196; Tero Karras, Samuli Laine, and Timo Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," arXiv preprint arXiv:1812.04948 (2018), https://arxiv.org/abs/1812.04948; Yunjey Choi et al., "StarGAN v2: Diverse

Image Synthesis for Multiple Domains," arXiv preprint arXiv:1912.01865 (2019), https://arxiv.org/abs/1912.01865.

[20] All re-identification and speaker recognition publication metadata displayed in this report was queried via BigQuery and extracted from the Map of Science clusters in March 2021. Subsequent updates to the Map of Science after this date will likely alter the clusters. All papers published before January 2010 and after December 2020 were omitted from the bibliometric analysis.

[21] The model determines which publications are "related" to AI broadly by predicting the categories that authors assign to their publications. It is "a strategy for identifying the universe of research publications relevant to the application and development of artificial intelligence. The approach leverages the arXiv corpus of scientific preprints, in which authors choose subject tags for their papers from a set defined by editors. We compose a functional definition of AI relevance by learning these subjects from paper metadata, and then inferring the arXiv-subject labels of papers in larger corpora: Clarivate Web of Science, Digital Science Dimensions, and Microsoft Academic Graph," Dunham et al., "Identifying the Development and Application of Artificial Intelligence in Scientific Text."; The merged corpus includes more papers than the Map of Science. Papers in the merged corpus that lack citations or references are not assigned to a research cluster and therefore are not included in the papers represented in the Map of Science.

[22] The amount of times a given set of papers was cited by other papers.

[23] Four clusters (#91424, #57343, #95593, and #14681) were omitted.

[24] Papers can have multiple country affiliations if the authors indicate more than one affiliation in a paper. Therefore, there are more country paper affiliations than papers (in the clusters).

[25] Table 2 shows the distributions of AI, re-identification, and speaker recognition paper author affiliations of the top 5 countries (by paper affiliations). Table 3 shows the citations of AI, re-identification, and speaker recognition papers of the top 5 countries (by citations of the country-affiliated papers).

[26] The most cited re-identification paper in our cluster is Liang Zheng et al., "Scalable Person Re-identification: A Benchmark," in *2015 IEEE International Conference on Computer Vision (ICCV)*, December 2015, https://ieeexplore.ieee.org/document/7410490; Additionally, the most cited speaker recognition paper in our cluster was produced by authors in French academia (School of Engineers in Computer Intelligence), United States academia (MIT), and Canadian academia (Universities of Quebec) and nonprofits (Computer Research Institute of Montréal). The paper is Najim Dehak et al., "Front-End Factor Analysis for Speaker Verification," IEEE/ACM Transactions on

Audio, Speech, and Language Processing 19, no. 4 (May 2011), https://ieeexplore.ieee.org/document/5545402.

[27] Percentages in Figures 6 and 7 are out of the total re-identification paper affiliations and paper citations of all countries in cluster #1419.

[28] China-affiliated re-identification paper citations may be moderately undercounted because papers from the CNKI dataset, which is a Chinese publication repository that CSET merged into the Map of Science, can lack citation data. There were at least 88 CNKI-sourced papers in cluster #1419 that had null citation values, although many of them were recently published. For the purpose of this report, we assess that the quantity of missing citations does not skew the data to a problematic degree.

[29] The sudden and sustained rise in China-affiliated re-identification publication is tied to many variables, but we assess that it largely stems from the proliferation of deep learning architectures, datasets, and compute resources, combined with political, security, and commercial incentives associated with China's rapid modernization and urbanization.

[30] The disparity in re-identification paper affiliations between China and other countries widened rapidly after 2016, with the ratio of U.S.-to-China affiliated papers being over 1:3 in 2017, 1:4 in 2018, and 1:5 in 2019. The affiliated paper citation disparity was 1:2 in 2017 and 1:3 in 2018, which is less stark but still significant. However, Chinese-affiliated papers sometimes lack data on citations because they are sourced from the CNKI dataset (which can lack data on citations). Therefore, the lower citations disparity is somewhat skewed, but it is not highly problematic.

[31] Percentages in Figures 8 and 9 are out of the total speaker recognition paper affiliations and paper citations of all countries in cluster #6855. Note that the top countries affiliated with papers are not always the same as the top countries affiliated with papers that are more frequently cited.

[32] The figure on the number of organizations is an approximation based on MAG affiliation IDs. Although reaching an exact figure is possible, it can require extensive manual data cleaning in the current version of the Map of Science.

[33] Approximately 20 percent of re-identification publications in cluster #1419 and 25 percent of speaker recognition publications in cluster #6855 are not linked to "organization type" metadata. A fully developed and operational Map of Science will likely address this issue, but it must be recognized that there are currently blind spots within the data. Areas that lack such visibility are readily identifiable and can often be addressed manually.

[34] Much of the data displayed in Table 4 was manually organized due to current limitations of the CSET Map of Science metadata (e.g., there were often overlapping names for the same entities).

[35] A comprehensive organizational analysis is outside the scope of this report, as we primarily intend to demonstrate a process for obtaining the information that enables subsequent policy inquiries. However, the metadata opens multiple vectors for further analysis and insight. This includes identifying key clusters of research among organizations (along with their geographic distributions), fostering collaborative research initiatives with universities and research centers for specific machine learning cross-sector collaboration (e.g., FFRDCs) and tracking organizational affiliations with foreign regimes.

[36] The public Map of Science will provide data on individual author countries (in cases where such information is available).

[37] We must note that two of the papers used to acquire more recent metrics were published after March 2021 (the month we extracted metadata from the clusters), therefore they are not included in our version of the clusters but would be if we repeated the methodology using updated clusters from the most recent version of the Map of Science.

[38] For more information on the MSMT17 dataset, see Longhui Wei et al., "Person Transfer GAN to Bridge Domain Gap for Person Re-Identification," arXiv preprint arXiv:1711.08565 (2018), https://arxiv.org/abs/1711.08565.

[39] For example, the use of large-scale "pre-training" on much broader vision recognition datasets.

[40] See Appendix G for details on performance metrics data and sources.

[41] For more information on the VoxCeleb dataset, see Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," arXiv preprint arXiv:1706.08612 (2018), https://arxiv.org/abs/1706.08612.

[42] For a more detailed assessment of re-identification and speaker recognition performance metrics, see Appendix F. For more details on metrics data and sources, see Appendix G.

[43] Figure 12 is sourced from Daniel Zhang et al., "Artificial Intelligence Index Report 2021" (Stanford University Human-Centered Artificial Intelligence (HAI), March 2021), https://hai.stanford.edu/research/ai-index-2021.

[44] "Leaderboard Version: 2.0," Gluebenchmark, https://super.gluebenchmark.com/leaderboard/; "Leaderboard," Gluebenchmark, https://gluebenchmark.com/leaderboard; "The GLUE benchmark, introduced a

little over one year ago, offers a single-number metric that summarizes progress on a diverse set of such tasks, but performance on the benchmark has recently surpassed the level of non-expert humans, suggesting limited headroom for further research;" SuperGLUE "is a new benchmark styled after GLUE with a new set of more difficult language understanding tasks, a software toolkit, and a public leaderboard … As with GLUE, we seek to give a sense of aggregate system performance over all tasks by averaging scores of all tasks. Lacking a fair criterion with which to weigh the contributions of each task to the overall score, we opt for the simple approach of weighing each task equally, and for tasks with multiple metrics, first averaging those metrics to get a task score," Alex Wang et al., "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems," arXiv preprint arXiv:1905.00537v1 (2019), https://arxiv.org/abs/1905.00537v1.

[45] GLUE and SuperGLUE consist of multiple natural language processing-related tasks that models are tested against, the results of which are aggregated into a single score. They were developed because of the emergence of a new class of highly capable natural language processing models that leveraged techniques that had previously improved computer vision (specifically, large-scale pre-training and the use of Transformer architecture models).

[46] By having multiple types of users, it will become easier to identify and triage areas for further investment and investigation. For instance, significant demand for greater country-level granularity regarding research paper affiliation could spur governmental efforts to develop the capability. Having such demand signals provides a cheap way to identify areas of shared need.

[47] Papers with Code is an example of such a resource.

[48] For example, one could identify a cluster or clusters related to a topic, extract all relevant metrics from the cluster publications, ingest them into the repository, then use both elements (the cluster bibliometrics and the repository metrics) to analyze the topic more holistically.

[49] "The cluster-level view in the Map of Science is the most granular aggregation of scientific research publications that provides insight into specific areas of research without analyzing publications individually," 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, 2021), https://cset.georgetown.edu/publication/comparing-the-united-states-and-chinas-leading-roles-in-the-landscape-of-science/. They are also assigned broad subject areas and specific research fields using MAG.

[50] "A typical speaker diarization system usually contains multiple steps. First, the non-speech parts are filtered out by voice activity detection. Second, the speech

parts are split into small homogeneous segments either uniformly or according to the detected speaker change points. Third, each segment is mapped into a fixed dimensional embedding. Finally, clustering methods or end-to-end approaches are applied to generate the diarization results," Jixuan Huang et al., "Speaker Diarization with Session-Level Speaker Embedding Refinement Using Graph Neural Networks," arXiv preprint arXiv:2005.11371 (2020), https://arxiv.org/abs/2005.11371.

[51] Regarding the mediums through which re-identification papers are published (i.e., document type metadata) and distributed (i.e., database metadata): Most re-identification papers in our cluster came from conferences between 2010–2018, but journal publications have increased significantly since 2019. Re-identification publications' growth likely began to stagnate in 2020 because COVID-19 disrupted publishing by forcing changes in the dates of conferences, through which approximately 41 percent of the re-identification papers in our cluster have been released. Additionally, this metadata was extracted from the Map of Science in March 2021, so the datasets may not be fully updated to reflect all the papers published (and cited) in 2020.

[52] Approximately 19 percent of the papers in cluster #6855 have null citations metadata. Although some data is missing, many null values are likely because approximately 12 percent of the papers were published since the year 2019 and have had limited time to accrue citations that would be reflected in the Map of Science.

[53] For example, systems that do well on MSMT17 tend to do large-scale "pre-training" on non-re-identification datasets, mirroring the performance trends in computer vision more broadly.

[54] Dan Hendrycks et al., "Natural Adversarial Examples," arXiv preprint arXiv:1907.07174 (2019), https://arxiv.org/abs/1907.07174.

[55] Chuting He et al., "TransReID: Transformer-Based Object Re-Identification," arXiv preprint arXiv:2102.04378 (2021), https://arxiv.org/abs/2102.04378; Abdallah Benzine et al., "Deep Miner: A Deep and Multi-Branch Network which Mines Rich and Diverse Features for Person Re-identification," arXiv preprint arXiv:2102.09321 (2021), https://arxiv.org/abs/2102.09321; Changxing Ding et al., "Multi-Task Learning with Coarse Priors for Robust Part-Aware Person Re-Identification," arXiv preprint arXiv:2003.08069 (2021), https://arxiv.org/abs/2003.08069; Tianlong Chen et al., "ABD-Net: Attentive but Diverse Person Re-Identification," arXiv preprint arXiv:1908.01114v3 (2019), https://arxiv.org/abs/1908.01114v3; Wei et al., "Person Transfer GAN to Bridge Domain Gap for Person Re-Identification," arXiv preprint arXiv:1711.08565 (2018), https://arxiv.org/abs/1711.08565.

[56] Nagrani et al., "VoxCeleb: A Large-Scale Speaker Identification Dataset"; Joon Son Chung, Arsha Nagrani, and Andrew Zisserman., "VoxCeleb2: Deep Speaker Recognition," arXiv preprint arXiv:1806.05622 (2018), https://arxiv.org/abs/1806.05622; Weidi Xie et al., "Utterance-Level Aggregation for Speaker Recognition in the Wild," arXiv preprint arXiv:1902.10107 (2019), https://arxiv.org/abs/1902.10107; Arsha Nagrani et al., "Voxceleb: Large-Scale Speaker Verification in the Wild," *Computer Speech & Language* 60 (March 2010), https://www.sciencedirect.com/science/article/pii/S0885230819302712; Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck, "The IDLAB VoxSRC-20 Submission: Large Margin Fine-Tuning and Quality-Aware Score Calibration in DNN Based Speaker Verification," arXiv preprint arXiv:2010.11255v1 (2020), https://arxiv.org/abs/2010.11255v1; Miao Zhao et al., "The Speakln System for VoxCeleb Speaker Recognition Challenge 2021," arXiv preprint arXiv:2109.01989 (2021), https://arxiv.org/abs/2109.01989.