Research Report

JON SCHMID

An Open-Source Method for Assessing National Scientific and Technological Standing

With Applications to Artificial Intelligence and Machine Learning

In the context of what the 2018 National Defense Strategy describes as "the re-emergence of long-term, strategic competition between nations" understanding the scientific and technological position of the United States relative to other nation-states is increasingly important (U.S. Department of Defense [DoD], 2018, p. 2). The RAND Corporation was asked to develop open-source methodological approaches for determining national standing within the science and technology (S&T) fields. Understanding the possible evolution of emerging technologies will inform DoD and Intelligence Community officials' decisionmaking on actions to take. As technologies evolve, these officials will make decisions about the implications of those technologies for defensive and offensive missions, investments in research and development (R&D), personnel, procurement, foreign material acquisition, and additional information collection. This report is designed to provide analysts and decisionmakers with a quick-turn and open-source methodology for assessing national scientific and technological standing for a given field. The sponsoring organization is DoD, and the sponsoring office seeks to understand the level and trajectory of output and capacity for various emerging technologies in major powers and potential adversaries.

This report describes the use of four metrics to quickly assess national S&T standing for a given field: high-impact publications, collaborative network density, quality-adjusted patents, and S&T organizational capacity. These metrics were selected according to the extent to which they capture major dimensions of the national S&T endeavor and their transparency, generalizability, and extensibility.

Metric 1 (high-impact publications) is a measure of national scientific output for the field in question. Because scientific journals and conference papers are the primary means by which researchers communicate scientific advancements,[1] scientific publications are a widely used mea-

sure of national scientific output.[2] This report considers only *high-impact scientific publications* (defined as publications that fall into the top decile in terms of citations received) to account for heterogeneity in average publication quality across countries (Schmid and Wang, 2017).

Metric 2 (network density) measures the connectedness of the communities of scientific practice within the focal country for the field in question. Countries with high network density for the focal field are those in which scientific collaboration is relatively prevalent. The empirical literature on scientific impact finds that increased collaboration is associated with greater scientific impact (Wuchty, Jones, and Uzzi, 2007). Thus, this report includes field-specific network density as a means of assessing the relative "health" of a country's scientific network within the field in question.

Metric 3 (quality-adjusted patents) gauges a country's ability to produce new technological inventions within the S&T field under consideration. To receive a patent, an application must document—and convince a patent examiner with subject-matter expertise in the field in question—that the underlying invention is nonobvious, novel, and useful. The set of patent grants developed by inventors from a given country for a given field thus constitutes a measure of a country's nontrivial invention output that, because of the novelty criteria, is free of double-counting. Here, each country's field-specific patent counts are weighted by average patent family size to account for international heterogeneity in patent quality (Schmid and Wang, 2017).

Metric 4 (S&T organizational capacity) assesses the extent of the organizational resources that a country employs to advance the focal S&T area. It is calculated as the sum of the number of organizations to have produced at least one patent in the field in question and the number of organizations to have produced at least one scientific publication in the field. It is thus a proxy for the number of organizations that a country hosts that are active in advancing the scientific or technological frontier for the field under scrutiny.

To best serve the dynamic needs of DoD and Intelligence Community officials, the method proposed here seeks to be transparent, generalizable, and extensible. To assure methodological transparency, several steps are taken. First, rather than building a single composite indicator, each metric is presented as a stand-alone indicator. Composite indicators—generally calculated as a weighted average of various component metrics—provide the apparent benefit of simplicity of interpretation, but they can mask important international differences. Furthermore, scholars have found that the particular approach chosen to calculate the component weights for composite indicators has a significant impact on indicator scores and can result in unstable country rankings (Grupp and Schubert, 2010).

Instead of combining the metrics developed here into a single composite indicator, each metric is presented as a stand-alone indicator for a given concept. Thus, rather than providing an apparently definitive—yet possibly misleading—answer to the question of which country is the global leader in field X, this methodology provides its user with four pieces of evidence that are critical to answering that question. This allows for a more nuanced understanding of the notion of S&T leadership in a given field. For example, suppose that a country is far and away the top producer of high-impact scientific journal articles in a given field, yet the firms from this country file very few patents. A composite indicator might assign the country a middling ranking, which would mask the observed high performance in one area and low performance in another. Furthermore, the composite metric would not point its user toward a potential flaw in the country's system of innovation: an apparent failure to commercialize scientific knowledge.

To allow assessment of a dynamic national security environment, this methodology is also designed

Abbreviations

AI	artificial intelligence
DoD	U.S. Department of Defense
GII	Global Innovation Index
IPC	International Patent Classification
ML	machine learning
R&D	research and development
S&T	science and technology
T&E	testing and evaluation
WOS	Web of Science

to be generalizable. To assure generalizability with regard to country, international data sources are used and adjustments are made to the metrics to account for well-known international differences in patenting and publishing behavior (Fisch, Block, and Sandner, 2016; Michalska-Smith and Allesina, 2017). To assure generalizability with regard to S&T topic area, the topic scoping strategy is defined primarily by using carefully designed keyword searches rather than predefined science or technology categories.[3] For example, a search strategy could be designed by the methodology's user to characterize a broad technology field, such as semiconductors. Similarly, by simply changing the search strategy—rather than the underlying metrics—an analyst can narrow the search to a particular type of semiconductor, such as photovoltaic semiconductors.[4] Such finely tuned scope definition is simply not possible using less descriptively rich data sources, such as R&D expenditure or total factor productivity.[5]

Finally, the proposed methodology is extensible. The primary purpose of the methodology is to inform answers to the question: Who are the global country-level leaders in a given S&T field? However, this is but one question of potential interest to analysts of S&T trends. Additional questions include the following:

- How do organizations collaborate internationally to advance a given field?
- What are the research foci of particular organizations within a country?
- What are the likely applications of the technologies in question in a given country?

Methods for beginning to answer these additional questions are proposed in the final section of this report.

To assure that these questions can be answered, the data-gathering strategy proposed here collects data beyond what is immediately necessary to calculate the four metrics. Specifically, data are collected on the organizations and individuals responsible for advancing the S&T sector, keywords, publications and patent classification codes, coauthorship and copatenting, patent and publication citations, and funding sources. The method also collects text data in the form of publication abstracts and titles and patent abstracts and titles. The analysis of these data using natural language processing methods is a powerful means of gleaning information regarding the particular scientific and technical foci of individuals, organizations, or countries.

Several limitations of this method are worth noting. First, it is important to note that although the methodology proposed here can be applied to most other S&T areas, the measurement of certain S&T areas would benefit from methodological tailoring. For example, the application of this model to S&T areas that are particularly capital-intensive or require extensive testing and evaluation (T&E) infrastructure could benefit from including an additional metric related to a country's T&E infrastructure. For example, if one were to apply the methodology to the field of hypersonic weapons, it might be advisable to include a T&E infrastructure metric based on the number of wind tunnels hosted by each assessed country. Second, the methodology proposed here relies on open-source quantifiable measures of national S&T activity. Although this expedites the analysis process, additional insight for any given S&T field regarding the prominence of a given individual, organization, or country could be gleaned through interviews with subject-matter experts from the field in question. Third, because this methodology relies on open source information, it might not yield accurate results for such S&T areas as nuclear physics and nuclear engineering, for which a large portion of advancement is kept secret.

A final limitation is that the proposed methodology focuses on a country's ability to advance the sci-

The data-gathering strategy proposed here collects data beyond what is immediately necessary to calculate the four metrics.

entific and technological frontier for a given field, not its ability to simply attain the output of the process of advancement. More exactly, the proposed methodology does not consider technology attained through external sources, such as intellectual property theft, espionage, international trade, or state-to-state technology transfer. The international diffusion of atomic weapons and rocket technology following the end of World War II suggests that technology attainment through those sources is likely to be particularly important for a state's capacity to acquire certain military technologies.

The remainder of this report is organized as follows. The next section describes the methodology; it provides definitions for each metric, an explanation of the rationale for the metric's inclusion, and the method of calculation. The section after that applies the methodology to a specific S&T area: artificial intelligence (AI) and machine learning (ML). That section reports the results of collecting data and calculating the four metrics for nine countries: China, France, Germany, India, Japan, South Korea, Russia, the United Kingdom, and the United States.[6] The final section extends the analysis to additional research questions.

The use of highly cited publications, as opposed to all publications in the field, is designed to account for the fact that there is systematic country-level variation in the impact of published scientific journal articles.

An Open-Source Method for Assessing National Scientific and Technological Standing

Metric 1: High-Impact Publications

High-impact publications are determined by calculating the annual number of publications written by authors affiliated with an organization within the country in question that fall into the top decile in terms of citations received. The decile cut-off point is calculated on an annual basis. The "country" field within the Web of Science (WOS) database (a scientific publication database that features over 90 million records) is used to determine the country affiliation associated with a publication. WOS populates the "country" field using the author address that is associated with a given publication. For example, publications for which an author lists an address as "Georgia Institute of Technology, North Ave NW, Atlanta, GA 30332, USA," will be assigned to the United States.

To illustrate the calculation of high-impact publications, consider Germany's 2015 Metric 1 performance in the field of AI/ML. In 2015, Germany published a total of 1,060 publications in the field. In 2015, the global top decile cut point was 18 citations. That is, 10 percent of AI/ML publications from 2015 globally received more than 18 citations as of the date of calculation (August 29, 2019). Of Germany's AI/ML publications, 210 received more than 18 citations. These publications—those within the top 10 percent of publications in terms of citations received—constitute Germany's 2015 Metric 1 calculation.

The use of highly cited publications, as opposed to all publications in the field, is designed to account for the fact that there is systematic country-level variation in the impact of published scientific journal articles. For example, research has found that the provision of direct financial incentives to publish in China appears to have led to a glut of low-quality publishing in that country (Schmid and Wang, 2017). The presence of systematic country-level variation in publication quality suggests that publication-based measures that simply count publications without adjusting for quality will suffer from low cross-country commensurability. Metric 1 accounts for

systematic country-level variation in publication quality by counting only highly cited publications. During the process of documenting scientific results, scientists cite articles that have informed their findings. Thus, the number of times that an article is cited is an indicator of the impact of a given article on subsequent scientific research.

Metric 2: Network Density

Network density is calculated as the number of observed interorganization collaborations that a country's publishing organizations have participated in as a proportion of the maximum possible number of collaborations for the field in question. A collaboration occurs when two researchers from different organizations work together on the same research article. The maximum possible number of collaborations is calculated as $n(n-1)/2$, where n refers to the number of organizations that have contributed to a scientific journal article during the year in question.

Whereas Metric 1 focused on scientific output, network density is a measure of the process by which this output is produced. It is a measure of the connectedness of the communities of scientific practice that produce scientific research. The empirical literature on scientific impact suggests that communities that are characterized by more collaboration are preferable to ones in which researchers are isolated. For example, studies have found that research produced by teams has, on average, greater scientific impact in terms of citations received than research produced by solo authors and that teams are more likely to produce research with exceptional impact (Wuchty, Jones, and Uzzi, 2007).

Metric 3: Quality-Adjusted Patents

Quality-adjusted patents refers to the number of patents granted to organizations from a country discounted by the average annual patent family size of patents from that country.[7] Patents are assigned to the country of origin of the patent's assignee (the patent's owner). For example, patents that are filed at the U.S. Patent and Trademark Office but have a Chinese assignee are allocated to China.

As in the case of scientific publications, there is well-known international heterogeneity in patent quality. For example, multiple recent studies have found that Chinese patents are of lower-than-average quality compared with their international counterparts (Fisch, Block, and Sandner, 2016; Li, 2012). Recent research indicates that U.S. patents are cited 2.25 times more often, on average, than Chinese patents (Schmid and Wang, 2017). Counting all patents equally would thus degrade the cross-country commensurability of the metric. Thus, to account for country-specific heterogeneity in patent quality, Metric 3 adjusts the raw patent count data using the average annual patent family size for the country in question.

The *family size of a patent* is the number of countries in which the patent has been filed. The method leverages the finding that patents filed in multiple jurisdictions have been found to be of higher quality than those filed in a single jurisdiction (Sampat, 2005; Thomson Reuters, 2011). Inventors of low-quality patents are thought to be unlikely to file in multiple countries because of the application cost and the risk of the application being denied (Schmid and Fajebe, 2019). The quality-adjusted patents metric for a given country c at year t is defined as

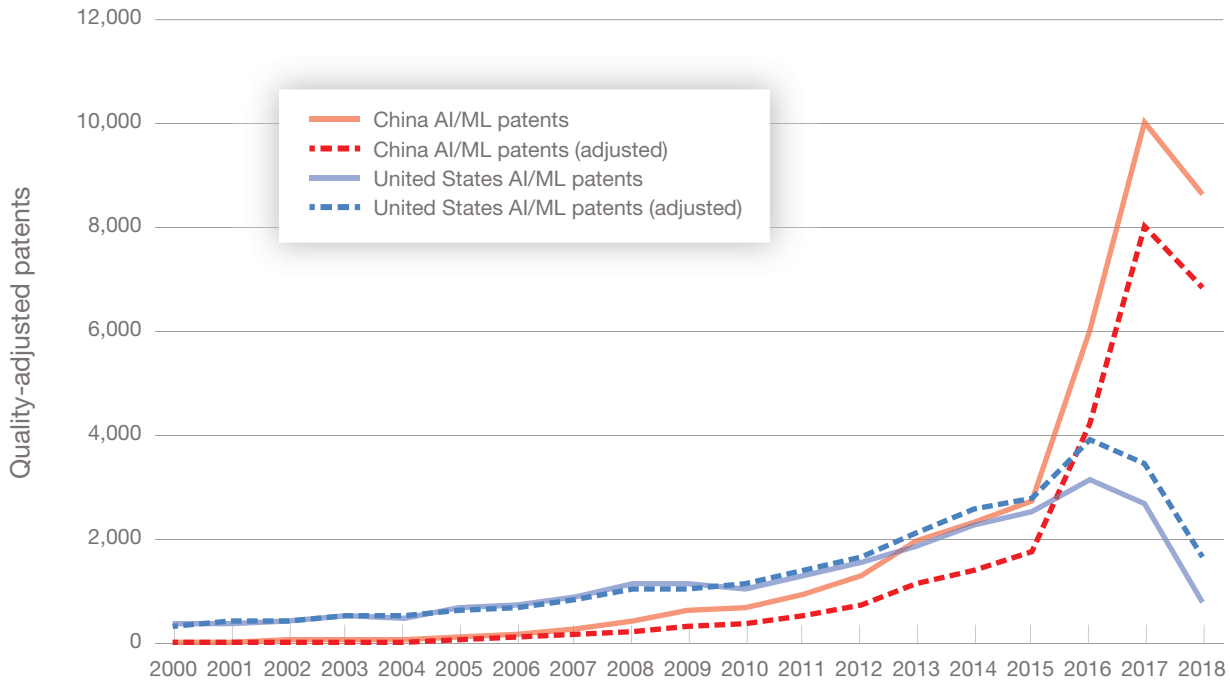
$$\frac{\text{average family size}_{t,\,c}}{\text{average family size}_{t,\,\text{global}}} \times \text{patents}_{t,\,c}\,.$$

In essence, Metric 3 is a country's patents in a given field adjusted for quality. In the formula provided, a country's quality proxy is calculated relative to the global average, so when the country's patents are of higher quality than the global average, the quality factor inflates that country's patent count. When a country's patents are of lower quality than the global average, the ratio is less than one, and the country's annual raw patent count is deflated.

Figure 1 demonstrates the effect of this quality adjustment on patent counts. The solid lines depict the annual unadjusted patent counts fitting the AI/ML patent search. Dotted lines depict the quality-adjusted patent counts. For most of the time series shown, the quality adjustment factor—the ratio of a country's average patent family size to the global average—results in an increase in the quality-adjusted metric

FIGURE 1

Quality-Adjusted Artificial Intelligence Patents, China and United States, 2000–2018



for the United States and a decrease for China. This is consistent with other research finding that Chinese patents tend to be on average lower quality than their U.S. counterparts (Fisch, Block, and Sandner, 2016).

Metric 4: Science and Technology Organizational Capacity

S&T organizational capacity is calculated as the annual number of organizations that have produced either a patent or a publication in the field in question. It is simply the sum of a country's patent assignees and its authors' affiliations. Assignment of nationality for publications and patents follows the same process as in previous metrics: For publications, the author's country field is used; for patents, the country of origin of the patent assignee is used.

Including a measure of organizational capacity seeks to capture the importance of organizational infrastructure in determining national S&T output. Holding other factors constant, this metric assumes that more organizations patenting and publishing in a given field is preferable to fewer organizations doing so. The presence of S&T organizations—such as firms, universities, and research government laboratories—has been empirically linked to national S&T output (Etzkowitz and Zhou, 2017).

It is worth noting that the S&T organizational capacity and network density metrics entail a normative claim about optimal organizational arrangement. S&T organizational capacity assumes that, holding other factors constant, more organizations participating in a sector is preferable to fewer organizations doing so. The network density metric assumes that, holding other factors constant, more collaboration is preferred to less. These normative assumptions are based on empirical scholarship linking organizations and collaboration to S&T output (Etzkowitz and Zhou, 2017; Wuchty, Jones, and Uzzi, 2007).

The Methodology in Context: Other Approaches to Assessing National Science and Technology Standing

The methodology proposed here was developed to be applied to a particular S&T field defined by the methodology's user. The four metrics used here—high-impact publications, network density, quality-adjusted patents, and S&T organizational capacity—can be calculated for the vast majority of S&T areas.[8] The

methodology proposed is thus distinct from methodologies that rank countries according to aggregate S&T capacity or overall innovation performance. It is also distinct from methodologies that are designed to rank countries for a single S&T field. This section locates the approach proposed in this report within the context of other S&T measurement approaches.

The Global Innovation Index (GII), produced annually by the World Intellectual Property Organization, is a prominent example of a measurement approach that seeks to measure aggregate national innovation performance (Dutta, Lanvin, and Wunsch-Vincent, 2019). The GII is made up of 80 indicators that are divided into seven major categories: institutions, human capital and research, infrastructure, market sophistication, business sophistication, knowledge and technology outputs, and creative outputs. Given that its focus is aggregate national innovation performance, the GII uses country-level variables (meaning variables that describe the national innovation environment irrespective of S&T sector). However, approaches that focus on country-level variables (such as domestic economic institutions and infrastructure) are not conducive to sector-specific international comparison.[9] For example, the GII uses Wikipedia page edits as one of its metrics of online creativity. This metric and the underlying phenomenon that it purports to measure are not equally relevant for all S&T fields. Online creativity can be a meaningful variable for internet-enabled fields (such as AI), but it is less clear that the rate at which a country's citizens edit Wikipedia would be relevant to fields that are less dependent on online user communities (such as nuclear power generation or drug discovery). Thus, because the objective of the methodology developed here is to assess national S&T standing for a given S&T area of interest, only metrics that can be calculated at the level of the S&T area are used.

In addition to multimetric approaches to measuring aggregate national S&T capacity, scholars have developed sector-specific national ranking methodologies. However, these approaches often cannot be applied to fields beyond the one for which they were designed. In 2019, for example, Tortoise Media produced a "Global AI Index," an AI-focused national ranking that uses more than 100 metrics to assign an AI Capacity score to 54 countries (Haynes and Gbedemah, 2019). One of the study's metrics of AI talent is the number of Python package downloads within a given country. Given the prominence of the Python programming language in the AI community at this point in time, this metric is a sound and creative way of measuring country-level AI talent. However, the use of field-specific metrics limits the generalizability of this approach—the use of Python package downloads might not be appropriate for other S&T fields. A user of the methodology seeking to apply the approach to a different S&T field would be required to identify appropriate proxies on a sector-by-sector basis. Thus, because the methodology developed here seeks generalizability across S&T areas, sector-idiosyncratic metrics are forgone.

The GII and the Global AI Index are composite measures.[10] Given the large number of metrics that make up these indexes, the calculation of a summary measure is an understandable way of facilitating interpretation. The method proposed here, however, refrains from including a composite measure because of observed instability in composite measures. Sensitivity analysis of innovation indexes has shown that the chosen weighting method can have a considerable effect on composite scores and country rankings (Grupp and Schubert, 2010; Grupp and Mogee, 2004). Specifically, commonly used weighting approaches, such as weighting via principal component analysis and equal-metric weighting, were found by Grupp and Mogee (2004) and Grupp and Schubert (2010) to produce significant variation in final country rankings.

The downside to not computing a composite indicator is the proposed method's inability, in some cases, to yield apparently definitive declarations regarding global leadership. That is, in cases for which no single country is ranked first in all measures, it is impossible, without the analyst using discretion, to claim that a given country is the definitive global leader for a given S&T field. For example, as discussed in the next section, the United States appears to be the global leader for three of the four metrics in the field of AI/ML. The long-winded conclusion—that the United States is the global leader in the field of AI/ML in terms of high-impact publications, network density, and S&T organizational capacity—is preferable to a more definitive but possibly less precise conclusion about overall global

AI/ML leadership that would result from computing a composite measure.

Finally, the metrics proposed here are not independent from one another. For example, national patent output is likely to be driven by a country's S&T organizational capacity. Neither are they independent from national contextual factors, such as how countries organize their overall S&T enterprise. Countries that organize their S&T enterprise in a fundamentally different way might be underranked using these metrics. Given this, absent further analysis, the metrics should not be used as statistical determinants (for example, as regressors in a regression model of national S&T capacity).

Applying the Methodology: Artificial Intelligence and Machine Learning, 2017–2018

Data and Search Strategy

In applying the measurement approach, this report uses patent data from the Derwent Innovation Index and publication data from the WOS (Clarivate; undated-a; Clarivate, undated-b). The period of analysis is the two-year period from 2017 to 2018. To produce a data set of AI/ML publications, a keyword search was relied on.[11] After removing such results as news items and commentaries that do not represent contributions to the scientific literature, the final sample consisted of 40,988 journal articles.[12] To arrive at a data set comprising AI/ML patents, a search strategy was used that combines keywords and international patent classification (IPC) codes. This patent search approach is modeled on a search strategy developed by the Intellectual Property Office of the United Kingdom (Economics, Research and Evidence Team, 2019). Performing searches over the full 2017–2018 period yielded 48,981 patents.[13] For both the publication and patent data set, additional data elements—such as keywords, titles, and abstracts—were collected to allow the extended analysis presented in the final section of this report.

Metric 1: High-Impact Publications

The plot in Figure 2 depicts the frequency distribution for all publications in the AI/ML sample over the 2017–2018 period. During that time, there were 40,988 articles matching the search criteria. Of these, 19,718 (48 percent) had not been cited at the date of calculation. The high-impact publications metric is concerned with only the top 10 percent of publications. When ranking articles by citations, the cut point for determining the top 10 percent of publications was found to be seven citations. Each country's high-impact publication measure represents the number of AI/ML publications to the right of the cut point that are produced by organizations from that country.

Over the 2017–2018 period of analysis, the United States is the global leader in terms of high-impact AI/ML scientific publications, producing 2,005 publications (41 percent) within the top decile in terms of citations received. During this period, China ranks second, having produced 1,033 (21 percent) high-quality AI/ML scientific publications. Figure 3 presents the Metric 1 results for the full sample of countries.

Metric 2: Network Density

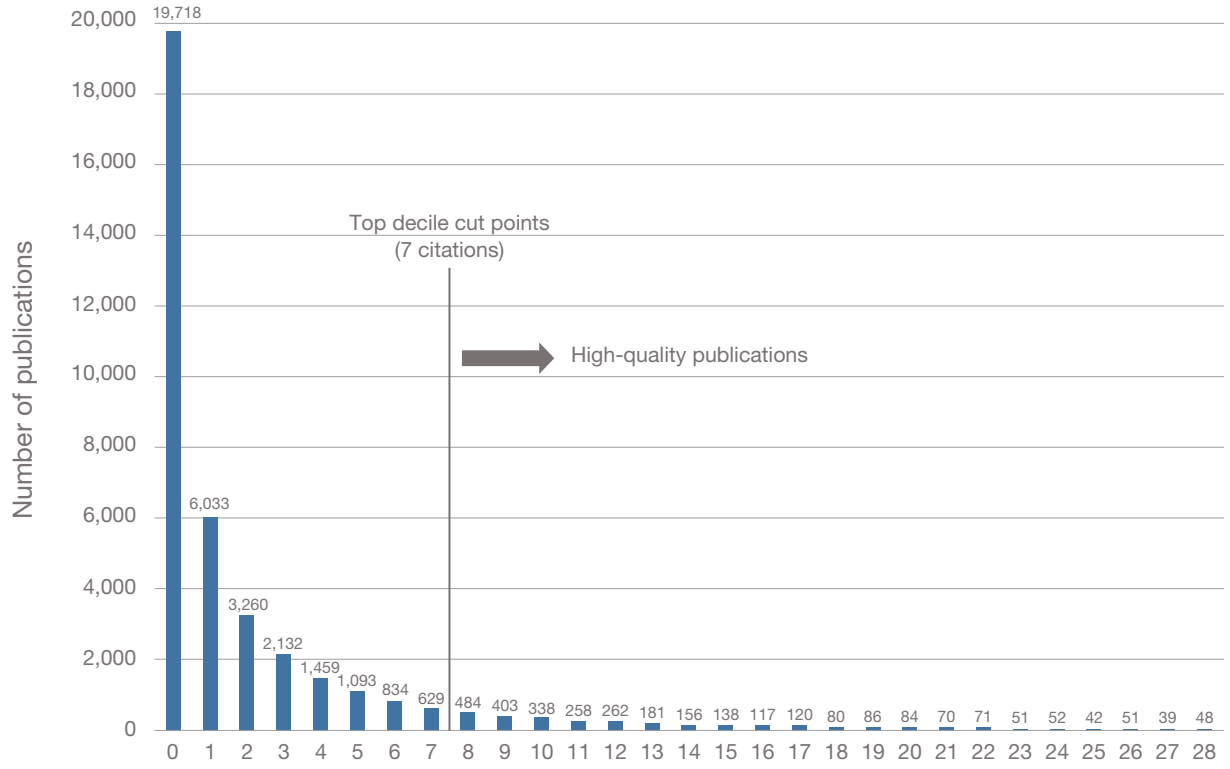
For the 2017–2018 period of analysis, the United States had the highest network density (0.027 percent) of countries in the sample. That is, relative to the other countries examined, the U.S. research community exhibits a high degree of connectedness. Table 1 depicts the performance of the nine countries in the sample in terms of network density over the combined 2017–2018 period.[14]

Metric 3: Quality-Adjusted Patents

Whereas the United States was the global leader for Metrics 1 and 2, China predominates AI/ML patenting. Raw, unadjusted patent counts indicate that China produced 18,646 patents (49.7 percent of the global total) within the AI/ML patent data set over the period of analysis. Even after adjusting for quality, China's patent output greatly exceeds that of other countries. Table 2 provides the raw patent counts, the adjustment factor, and the adjusted figures.

FIGURE 2

Frequency Distribution of Citations, Artificial Intelligence and Machine Learning Publications, 2017–2018



Metric 4: Science and Technology Organizational Capacity

The United States was the global leader for the 2017–2018 period of analysis for Metric 4. That is, the United States hosted more organizations to have produced either a scientific journal article or a patent in the field of AI/ML than any other country in the sample. The U.S. position of leadership is driven by the large number of organizations publishing in the field. In terms of patenting organizations, China is the global leader. Figure 4 depicts the sample countries' performance on this metric.

Conclusion: Final Rankings

Table 3 summarizes the national rankings based on the four metrics calculated here. Using the proposed methodology for assessing national S&T standing, the United States ranks first in three of the four measures. These results suggest that a strong case can be made that the United States is the global leader for AI/ML.

The summary table also illustrates the role that the proposed methodology can play in providing policy insight. For example, given almost any weighting system, creating a composite indicator from these metrics would rank China very highly.[15] However, considering the composite number alone would mask the fact that the collaboration network density in China is not particularly high. Failing to aggregate metrics allows analysts to identify portions of national techno-innovation systems that might warrant policy intervention. For example, empirical research on the collaboration in innovation finds that firms receiving government financial support are more likely to engage in collaboration (Mohnen and Hoareau, 2003; Bayona Sáez, Garcia Marco, and Huerta Arribas, 2002).

Besides allowing for assessment of national standing with regard to an S&T field, this methodology was designed to be extensible. That is, it was designed to collect sufficiently detailed data regarding the selected S&T sector to allow for further analysis. The next section provides sample analyses to illustrate possible

FIGURE 3

High-Impact Artificial Intelligence and Machine Learning Publications, 2017–2018

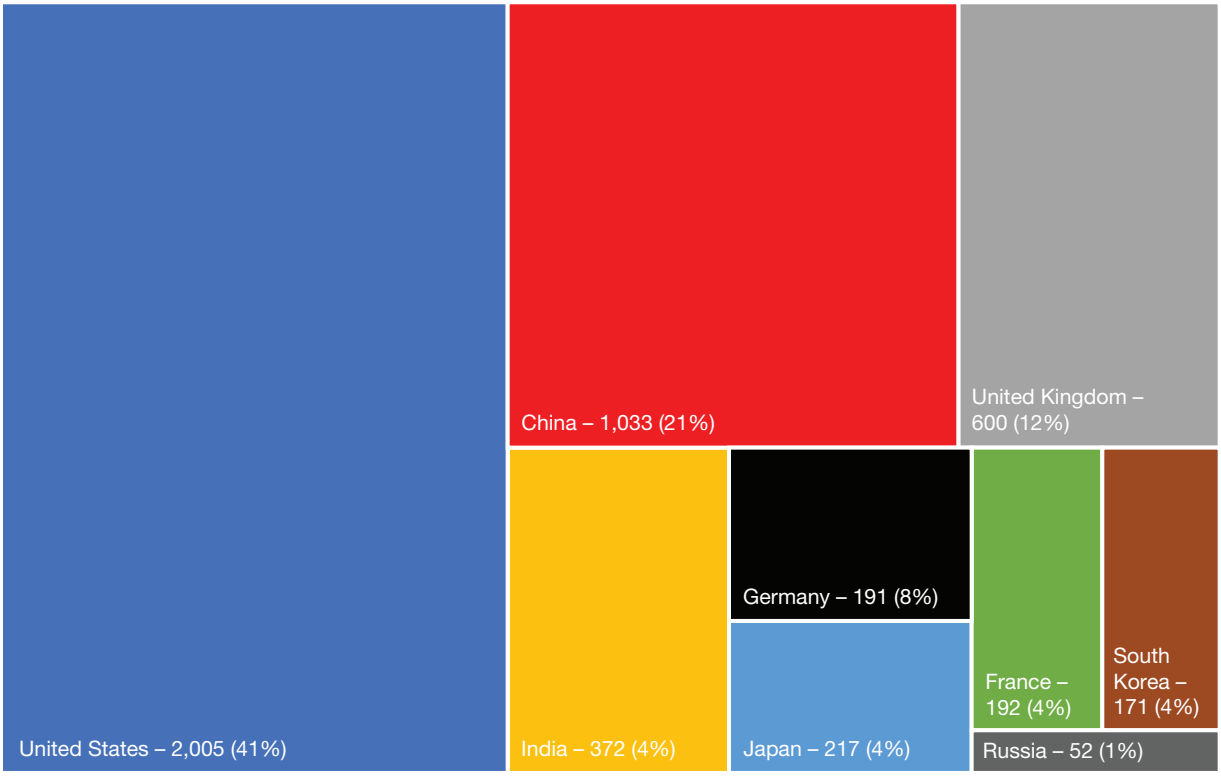


TABLE 1

Network Density for Artificial Intelligence and Machine Learning Publication Network

Country	AI/ML Collaborations (Ties)	AI/ML Articles	Network Density (%)
United States	173,624	13,000	0.027
United Kingdom	128,540	2,701	0.020
Germany	122,748	2,195	0.019
France	112,218	1,363	0.018
China	108,316	6,590	0.017
Russia	86,460	688	0.014
South Korea	85,140	1,238	0.013
India	63,890	2,805	0.010
Japan	28,540	1,613	0.005

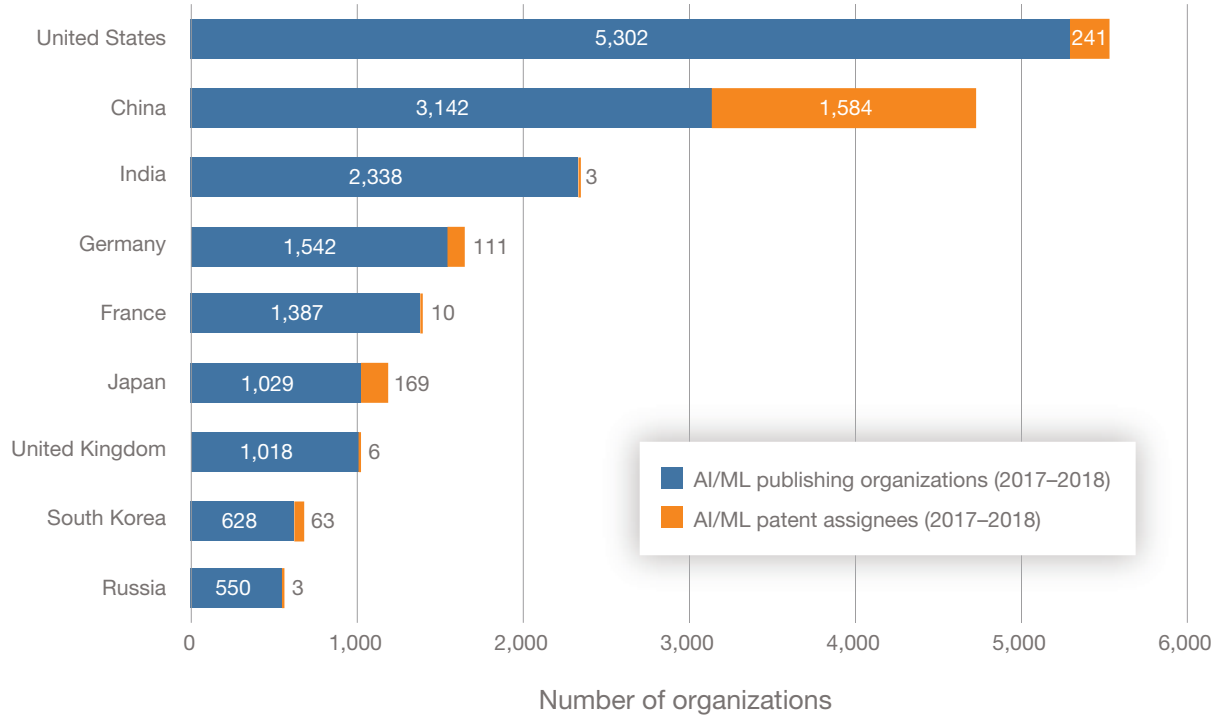
TABLE 2

Quality-Adjusted Artificial Intelligence and Machine Learning Patents

Country	AI/ML Patents	Average Family Size	Quality-Adjusted AI/ML Patents
China	18,646	1.1	14,875
United States	3,442	2.3	5,091
Japan	1,042	2.4	1,570
South Korea	558	1.9	760
Germany	377	2.5	629
United Kingdom	80	2.6	140
India	59	4.5	123
France	32	3.3	71
Russia	5	1.0	4

FIGURE 4

Science and Technology Organizational Capacity, 2017–2018



means by which the data collected here can be used to answer other research questions of interest.

Extending the Analysis

The data gathered to calculate the four metrics presented here contain additional information that can be used to glean insight into other features of S&T activity in the field of AI/ML. This section briefly examines three topics of potential interest: international patterns of collaboration, the role and research foci of particular organizations, and a particular application area of the field.

It should be noted that this section does not present a comprehensive catalogue of scientometric- or bibliometric-based analysis techniques; rather, it

TABLE 3
Final Rankings in Science and Technology, 2017–2018

Country	Metric 1 (High-Impact Publications)	Metric 2 (Network Density)	Metric 3 (Quality-Adjusted Patents)	Metric 4 (S&T Organizational Capacity)
United States	1	1	2	1
China	2	5	1	2
Germany	5	3	5	4
United Kingdom	3	2	6	7
India	4	8	7	3
Japan	6	9	3	6
France	7	4	8	5
South Korea	8	7	4	8
Russia	9	6	9	9

provides a sample of additional methods of analysis. Other areas of potential interest to the defense and intelligence communities that can be explored using these data include principal component analysis of the corpus of text and of keywords to identify the underlying intellectual structure; social network analysis to identify highly impactful researchers or research teams; and analysis of the subset of patents that are overtly meant for military, weapons, or intelligence purposes.[16]

International Patterns of Collaboration

Patterns of international scientific collaboration can be of interest to the defense and intelligence community for several reasons. Research suggests that international collaboration on scientific research is positively related to research impact (Guerrero Bote, Olmeda-Gómez, and de Moya-Anegón, 2013). However, international scientific collaboration, especially between potential adversaries, could represent a means by which intellectual property is unintentionally exported abroad.

Figure 5 depicts the coauthorship network for the 3,000 most highly cited journal articles in the AI/ML sample. Nodes are sized using *betweenness centrality*, a measure of how central a node is in the network. Nodes with a high betweenness centrality are important if information has to be passed from one part of a network to another. Colors are assigned according to a Louvain clustering algorithm that aims to detect communities of collaboration. Edges are weighted according to the number of collaborations. Intriguingly, the most common bilateral collaborative relationship is the one between the United States and China. Figure 6 uses the same data but depicts patterns of collaboration geographically.

The Role and Research Foci of Particular Organizations

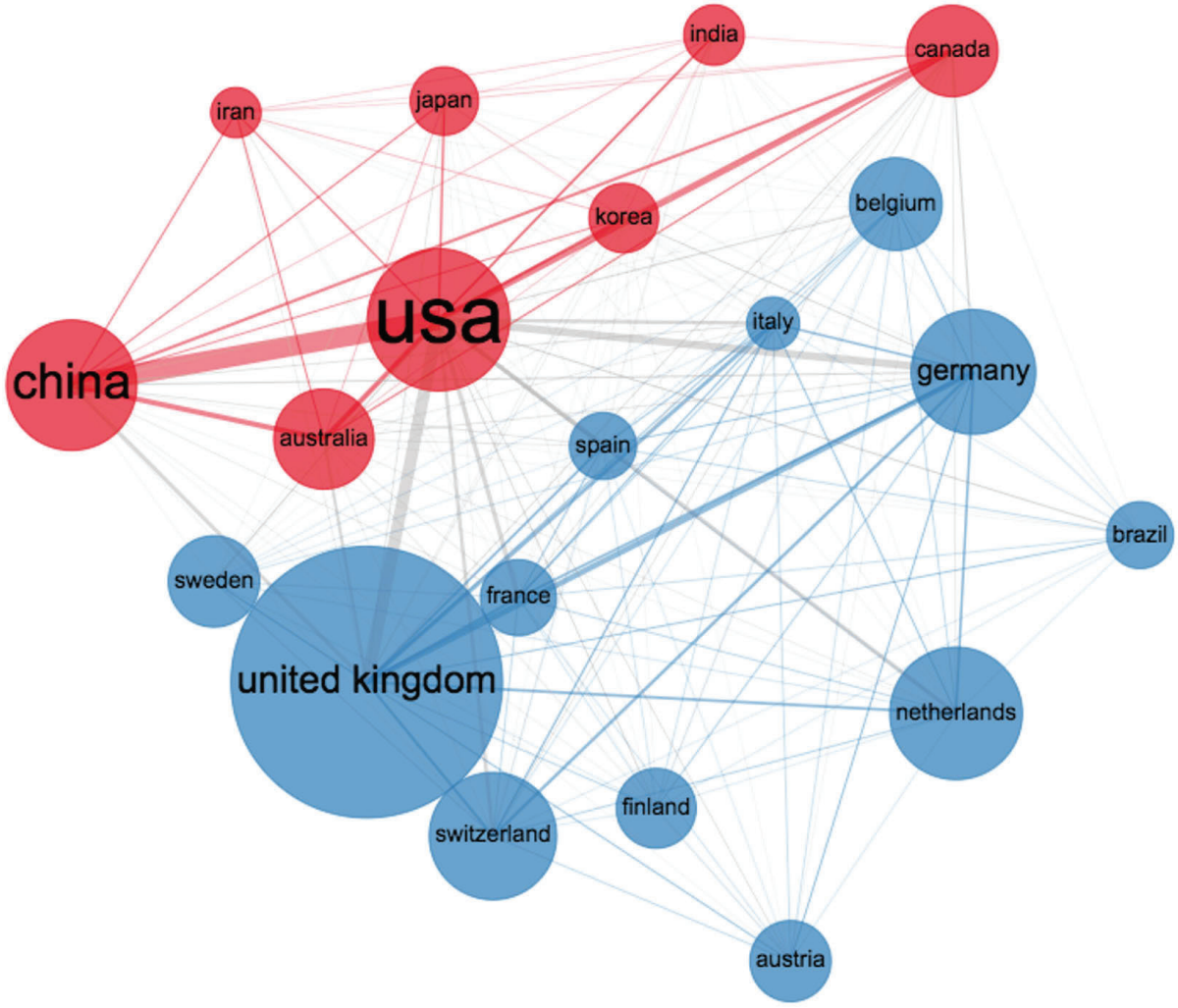
Another topic of potential interest to the defense and intelligence community is how particular research foci are distributed across organizations and countries. The topic of AI/ML is sufficiently broad that increasing the granularity of analysis to the level of keywords can be desirable. Figure 7 depicts how AI/ML publications in the high-impact subsample are allocated across country, organization, and keywords. For example, the figure shows that only two universities (Stanford and the University of Illinois) within the subsample are conducting high-impact research in the field of remote sensing. The diagram also shows that no single university dominates publishing in the field; the size of the university "bars" is not highly variable.

Deep Dive into an Application Area (Artificial Intelligence, Machine Learning, and Cyber)

The data gathered to calculate the metrics proposed here also can be used to examine a particular appli-

FIGURE 5

Collaborative Network, Artificial Intelligence and Machine Learning Publications



NOTES: Members of two distinct communities of scientific collaboration are distinguished by red and blue nodes; edges that connect the two are colored in gray to enhance the graph's clarity. Node colors are assigned according to a Louvain community detection algorithm. The thickness of edges is determined by the number of collaborations between the linked nodes. To more clearly depict major collaborative relationships, the edges shown represent collaborative dyads for which at least ten collaborations have occurred within the subsample. To show major relationships, only a subsample (the top 3,000 observations in terms of citations received) of the full sample is depicted. To increase the clarity of the graph, only the top 20 publishing countries in the sample are displayed in the network graph.

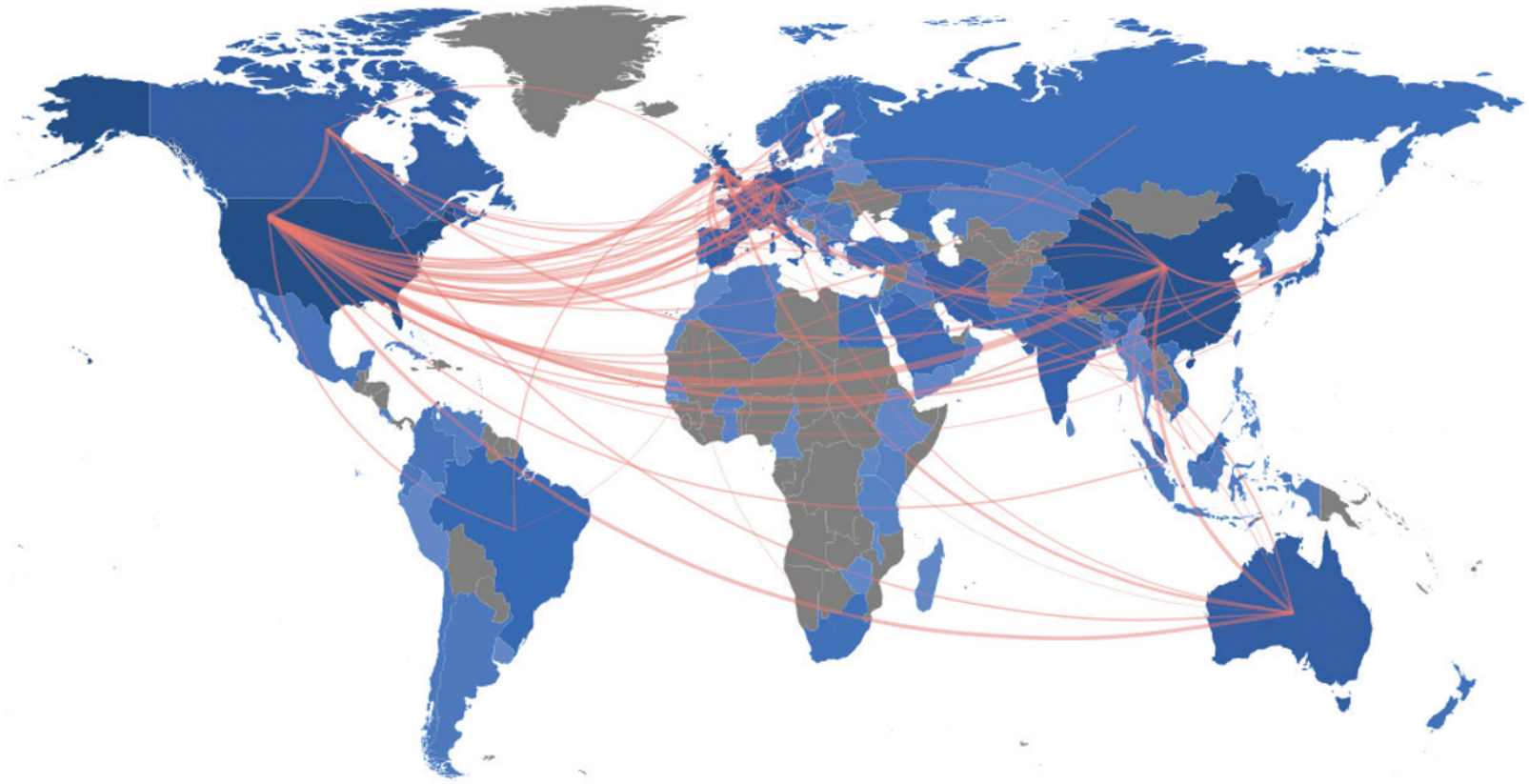
cations area. Recently, AI/ML have been used to detect and respond to cyberattacks.[17] However, AI also poses a means of increasing the efficacy of cyber offense (Brundage et al., 2018). To illustrate how a particular applications area can be analyzed using the data elements collected here, a subset of the AI/ML patent data related to cybersecurity was selected.

Construction of this data set began with the full AI/ML patent data set for the 2000–2019 period. For each of the 106,740 patents in this database, the patent abstract and patent title were searched for a series of key terms related to cybersecurity.[18] When a match was found, these patents were added to the new "AI + cyber" data set. Thus, the new "AI + cyber" data set represents the intersection of the AI/ML data set and the cyber keyword search.

Table 4 depicts the frequency for the cybersecurity terms for U.S. and Chinese patents. The numbers in cells refer to the number of patents that contain both the cyber keyword and one of the AI keywords. The shaded cells depict the difference (United States minus China) in the number of pat-

FIGURE 6

International Collaboration, Artificial Intelligence and Machine Learning Publications



NOTE: Countries shaded in gray have not produced any AI/ML publications. Countries shaded in blue have produced at least one AI/ML publication; darker blue reflects a larger number of AI/ML publications produced. The red lines reflect international collaborations between the linked countries.

FIGURE 7

Research Foci by Country and Organization, Artificial Intelligence and Machine Learning Publications

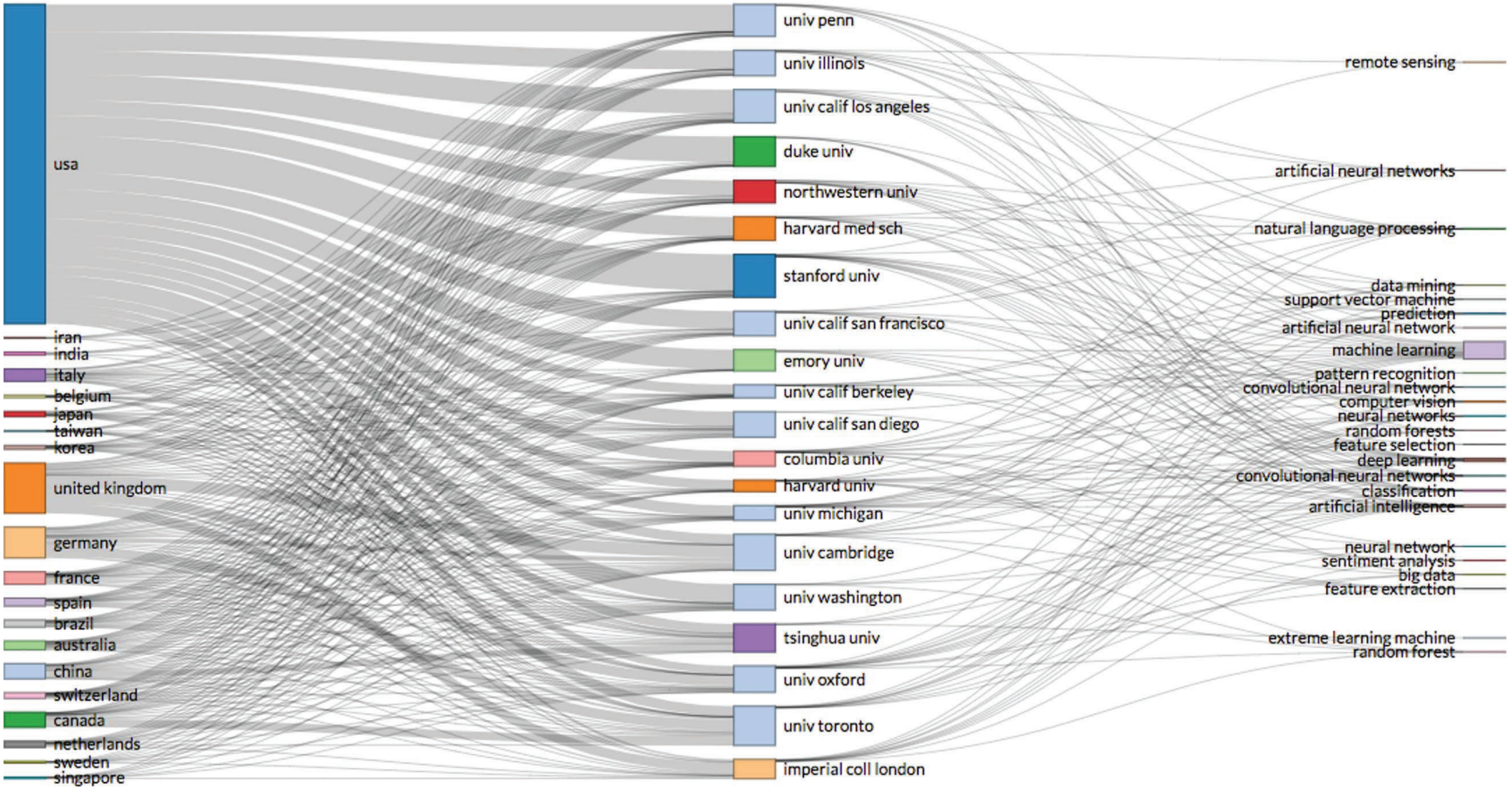


TABLE 4

Term Analysis of Artificial Intelligence and Cyber Application Domain, United States and China

	2010–2019			2015–2019		
Search term	China	United States	U.S. advantage	China	United States	U.S. advantage
AI + malware	13	66	53	12	48	36
AI + anomaly detection	26	73	47	24	55	31
AI + bot	2	33	31	2	31	29
AI + cyber	2	35	33	2	24	22
AI + Firewall	6	27	21	3	17	14
AI + breach	4	17	13	4	12	8
AI + cyber attack	0	9	9	0	7	7
AI + botnet	5	8	3	4	8	4
AI + data integrity	7	13	6	4	7	3
AI + Adware	0	4	4	0	3	3
AI + Spoof	6	8	2	6	8	2
AI + Handshak*	3	7	4	2	4	2
AI + cybersecurity	0	3	3	0	2	2
AI + cyber threat	0	2	2	0	2	2
AI + computer virus	1	2	1	0	1	1
AI + change point detection	0	3	3	0	1	1
AI + hack*	5	12	7	5	5	0
AI + Zero-Day	1	2	1	1	1	0
AI + data injection	1	1	0	1	1	0
AI + Rootkit	0	1	1	0	0	0
AI + data security	11	13	2	10	9	–1
AI + traffic classification	5	5	0	5	4	–1
AI + Sniff*	4	2	–2	3	2	–1
AI + cascading failure	5	0	–5	2	0	–2
AI + scada systems	10	2	–8	6	1	–5
AI + Phishing	18	4	–14	9	3	–6
AI + DDOS	11	6	–5	10	2	–8
AI + penetration	27	13	–14	22	11	–11
AI + access control	39	37	–2	27	15	–12
AI + Trojan horse	21	3	–18	15	2	–13
AI + information security	28	5	–23	24	5	–19
AI + vulnerability	39	14	–25	33	8	–25
AI + Encryption	110	71	–39	86	43	–43
AI + intrusion detection	86	24	–62	69	13	–56
AI + Hash	143	87	–56	122	38	–84

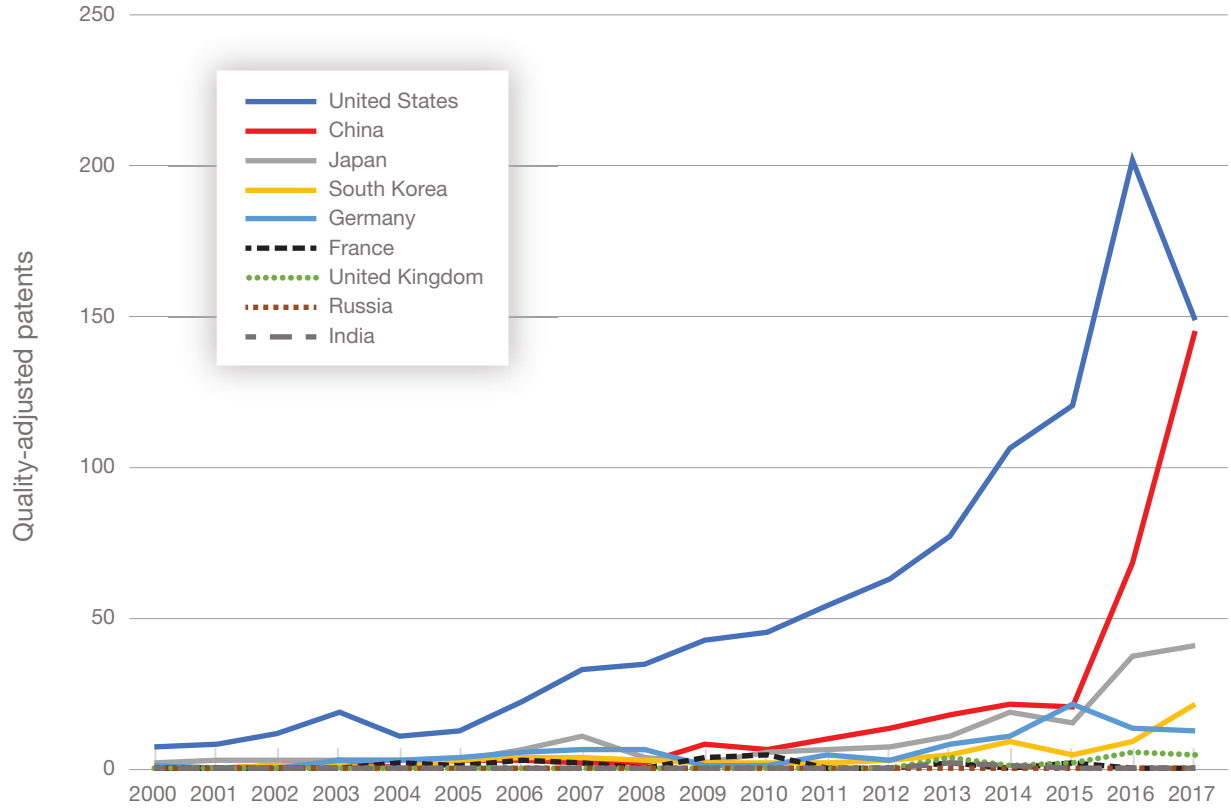
NOTE: The keyword analysis presented here is sensitive to the particular language choices of the authors of the patent documents. In some cases, variation in countries' totals might reflect intercountry variation in language preferences rather than true differences in innovation. Blue cells reflect positive values; red cells reflect negative values.

ents for each keyword combination. Cells shaded in blue represent those in which the United States has an advantage in terms of patents produced in the area in question. Cells shaded in red (those with negative values) represent those in which China has an advantage.

Figure 8 depicts the annual adjusted patent counts for the AI + cyber subset. For this data series, we applied the quality adjustment methodology used to discount Metric 3. The data in Figure 8 are truncated for 2017 because the Derwent Innovation Index patent data are not complete for 2018. The plot indicates that the United States was the early leader in AI + cyber patents, but Chinese patenting on this topic accelerated dramatically beginning in 2015. By 2017, the United States and China were producing roughly the same number of quality-adjusted patents in the AI + cyber field.

FIGURE 8

Artificial Intelligence and Cyber Patents (Quality-Adjusted), 2000–2017



Notes

[1] In certain fields, such as computer science, conferences are the primary means of communicating advancements. The high-impact publications metric proposed here thus includes conference presentations.

[2] The Nature Index, published annually by Nature Research, is one example of a prominent national scientific output ranking based on scientific publications (Nature Research, 2021).

[3] As will be described in the next section ("An Open-Source Method for Assessing National Scientific and Technological Standing"), the patent search strategy employs a dual approach, combining keyword and patent classification code searches.

[4] For example, to define the search strategy for photovoltaic semiconductors, the original semiconductor search could be combined with a series of photovoltaic keywords (such as *active layer*, *amorphous silicon*, *light absorption*, *photoelectric conversion efficiency*, *solar cell*, *radiation*, and *lead frame*).

[5] Patent and publication data are particularly effective in satisfying the generalizability criteria because these data sources have descriptively rich fields—such as abstracts, titles, keywords, and patent classification codes—that can be used to precisely define a topic of interest.

[6] These countries were selected at the request of the sponsor of the research.

[7] Eusebi and Silberglitt (2014) propose an innovative method for assessing patent maturity via cross-domain linkages (meaning co-occurrence of international patent classification codes on a given patent). Eusebi and Silberglitt argue that, as a technology area matures, patents will be linked to more technological domains. If an analyst were interested in weighting patents (or a portfolio of patents) for maturity, this approach could be applied in a manner similar to the weighting approach proposed here (that is, weighting the patents by their cross-domain linkages relative to the global average number of cross-domain linkages).

[8] Exceptions are S&T fields that protect intellectual property via secrecy or those in which the publications and patents are not publicly released.

[9] Other examples of composite metrics that use country-level variables are the European Innovation Scoreboard (produced by the European Commission) and the International Innovation Index (produced by Boston Consulting Group, National Association of Manufacturers, and The Manufacturing Institute). Importantly, the purpose of these aggregate indexes is not to assess sector-specific S&T standing but to rank countries in the aggregate.

[10] Both indexes provide access to the composite metrics. The Global AI Index provides a sensitivity analysis in its weighting approach and found high stability in the rankings based on the chosen weighting strategy.

[11] Search performed on the WOS Core Collection. The exact search was TS=("Machine Learning" OR "Artificial Intelligence")

[12] The document types that were removed were editorial materials, book reviews, letters, corrections, news items, reprints, software reviews, biographical items, bibliographies, retracted publications, corrections or additions, film reviews, items about individuals, database reviews, retractions, and poetry.

[13] The exact search performed was: IP=(G06F-019/24 OR G06N-003* OR G06N-005* OR G06N-007/02 OR G06N-007/04 OR G06N-007/06 OR G06N-020* OR G06T-001/40 OR G16B-040/20 OR G16B-040/30 OR G16C-020/70) OR TS=("ant-colony" OR "factorization machin*" OR "high-dimensional* feature*" OR "particle-swarm*" OR "bee-colony" OR "factorisation machin*" OR "factorization machin*" OR "high-dimensional* input*" OR "pattern-recogni*" OR "fire-fly" OR "feature engineer*" OR "k-means" OR "policy-gradient method" OR "adversar* network*" OR "feature extract*" OR "kernel learn*" OR "q-learn*" OR "artificial*-intelligen*" OR "feature select*" OR "latent-variable*" OR "random-forest*" OR "association rule" OR "fuzzy-c" OR "link* predict*" OR "recommender system*" OR "auto-encod*" OR "fuzzy environment*" OR "machine intelligen*" OR "reinforc* learn*" OR "autonom* comput*" OR "fuzzy logic*" OR "machine learn*" OR "sentiment* analy*" OR "back-propagat*" OR "fuzzy number*" OR "map-reduce" OR "sparse represent*" OR "back-propogat*" OR "fuzzy set*" OR "memetic algorithm*" OR "sparse*-code*" OR "cognitiv* comput*" OR "fuzzy system*" OR "multi* label* classif*" OR "spectral cluster*" OR "collaborat* filter*" OR "gaussian mixture model" OR "multi*-objective* algorithm*" OR "stochastic*-gradient*" OR "deep-belief network*" OR "gaussian process*" OR "multi*-objective* optim*" OR "*supervis* learn*" OR "deep-learn*" OR "genetic program*" OR "natural-gradient" OR "support-vector machine" OR "differential*-evol* algorithm*" OR "genetic* algorithm" OR "neural-turing" OR "swarm behav*" OR "dimensional*-reduc*" OR "high-dimensional* data" OR "*neural-network*" OR "swarm intell*" OR "ensemble-learn*" OR "high-dimensional* model*" OR "neuro-morph comput*" OR "transfer-learn*" OR "evolution* algorithm*" OR "high-dimensional* space*" OR "non-negative matri* factor*" OR "variation*-infer*" OR "evolution* comput*" OR "high-dimensional*system*" OR "object-recogni*" OR "vector-machine*")

[14] Over the period in question, 35,539 organizations produced at least one patent in the field. Thus, there are 631,492,491 [(35,539 * (35,539 – 1)/2] possible ties in the network. The network density metric is the ratio of actual ties to the number of possible ties. The U.S. network density metric is 0.027 percent (173,624 ÷ 631,492,491).

[15] An equal weighting scheme would, for example, rank China second among the nine countries considered here.

[16] An approach to identifying and analyzing military patents is provided in Schmid (2018).

[17] It is worth noting that some portion of patents and scientific publication on AI and cyber is likely to be classified. As the method proposed here seeks to rely exclusively on publicly available databases, these data are considered out of the scope of research.

[18] The process for extracting cybersecurity keywords was as follows. The WOS was searched for cybersecurity. All keywords from the top 200 most-cited journal articles and the 200 most-recent journal articles were then extracted (literature review). Journal articles, compared with patents, contain standardized keywords provided by the authors. Additional cybersecurity terms were added to the list based on a cybersecurity literature review. The terms used to define the cybersecurity portion of the AI + cyber data set are provided. The source of the term is provided in parentheses. Search terms are access control (WOS), adequacy assessment (WOS), Adware (literature review), anomaly detection (WOS), bad data detection (WOS), bot (literature review), botnet detection (WOS), breach announcements (WOS), cascading failures (WOS), cash out problem (WOS), change point detection (WOS), collaborative intrusion detection (WOS), common vulnerability (WOS), computer virus (WOS), computing security (WOS), cyber (WOS), cyber attacks (WOS), cyber conflict (WOS), cyber physical security (WOS), cyber physical systems (WOS), cyber security (WOS), cyber threat intelligence (WOS), cyber warfare (WOS), cyberattack (literature review), cybersecurity (WOS), cyberwar (WOS), data injection attacks (WOS), data integrity (WOS), data security (WOS), ddos (literature review), detection systems (WOS), encryption (literature review), encryption (literature review), energy theft detection (WOS), false data injection (WOS), firewall (literature review), hackers (WOS), hacking (WOS), handshaking (literature review), hashing (literature review), information security (WOS), intrusion detection (WOS), key agreement protocol (WOS), keylogger (literature review), malware (WOS), malware (literature review), malware detection (WOS), microcystins (WOS), penetration (literature review), pharming (literature review), phishing (literature review), protection motivation (WOS), protection motivation theory (WOS), ransomare (literature review), trojan horse (literature review), rootkit (literature review), scada systems (WOS), sniffing (literature review), spoof (literature review), traffic classification (WOS), trojans (WOS), virus (literature review), vulnerability (WOS), vulnerability assessment (WOS), zero-day (literature review).

References

Bayona Sáez, Cristina, Teresa Garcia Marco, and Emilio Huerta Arribas, "Collaboration in R&D with Universities and Research Centres: An Empirical Study of Spanish Firms," *R&D Management*, Vol. 32, No. 4, 2002, pp. 321–341.

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, arXiv, 2018.

Clarivate Analytics, "Web of Science Platform," database, undated-a. As of August 19, 2021: https://clarivate.com/webofsciencegroup/solutions/web-of-science

Clarivate Analytics, "Web of Science Platform: Derwent Innovation Index," database, undated-b. As of August 19, 2021: https://clarivate.libguides.com/webofscienceplatform/dii

Dutta, Soumitra, Bruno Lanvin, and Sacha Wunsch-Vincent, eds., *Global Innovation Index 2019: Creating Healthy Lives—The Future of Medical Innovation*, 12th ed., Ithaca, N.Y.: Cornell University, INSEAD, and the World Intellectual Property Organization, 2019.

Economics, Research and Evidence Team, *Artificial Intelligence: A Worldwide Overview of AI Patents and Patenting by the UK AI Sector*, Newport, South Wales: Intellectual Property Office, June 18, 2019.

Etzkowitz, Henry, and Chunyan Zhou, *The Triple Helix: University–Industry–Government Innovation and Entrepreneurship*, London: Routledge, 2017.

Eusebi, Christopher A., and Richard Silberglitt, *Identification and Analysis of Technology Emergence Using Patent Classification*, Santa Monica, Calif.: RAND Corporation, RR-629-OSD, 2014. As of August 18, 2021: https://www.rand.org/pubs/research_reports/RR629.html

Fisch, Christian O., Joern H. Block, and Philipp Sandner, "Chinese University Patents: Quantity, Quality, and the Role of Subsidy Programs," *Journal of Technology Transfer*, Vol. 41, No. 1, 2016, pp. 60–84.

Grupp, Hariolf, and Mary Ellen Mogee, "Indicators for National Science and Technology Policy: How Robust Are Composite Indicators?" *Research Policy*, Vol. 33, No. 9, 2004, pp. 1373–1384.

Grupp, Hariolf, and Torben Schubert, "Review and New Evidence on Composite Innovation Indicators for Evaluating National Performance," *Research Policy*, Vol. 39, No. 1, 2010, pp. 67–78.

Guerrero Bote, Vicente P., Carlos Olmeda-Gómez, and Félix de Moya-Anegón, "Quantifying the Benefits of International Scientific Collaboration," *Journal of the American Society for Information Science and Technology*, Vol. 64, No. 2, 2013, pp. 392–404.

Haynes, Andrew, and Luke Gbedemah, *The Global AI Index: Methodology*, Tortoise Media, December 2019.

Li, Xibao, "Behind the Recent Surge of Chinese Patenting: An Institutional View," *Research Policy*, Vol. 41, No. 1, 2012, pp. 236–249.

Michalska-Smith, Matthew J., and Stefano Allesina, "And, Not Or: Quality, Quantity in Scientific Publishing," *PLoS ONE*, Vol. 12, No. 6, 2017.

Mohnen, Pierre, and Cathy Hoareau, "What Type of Enterprise Forges Close Links with Universities and Government Labs? Evidence from CIS 2," *Managerial and Decision Economics*, Vol. 24, Nos. 2–3, 2003, pp. 133–145.

Nature Research, "Nature Index," Current Index database, 2021. As of January 2021: https://www.natureindex.com/country-outputs/generate/All/global/All/score

Sampat, Bhaven, *Determinants of Patent Quality: An Empirical Analysis*, New York: Columbia University, 2005. As of January 2021: http://immagic.com/eLibrary/ARCHIVES/GENERAL/COLUMBIA/C050902S.pdf

Schmid, Jon, "The Diffusion of Military Technology," *Defence and Peace Economics*, Vol. 29, No. 6, 2018, pp. 595–613.

Schmid, Jon, and Ayodeji Fajebe, "Variation in Patent Impact by Organization Type: An Investigation of Government, University, and Corporate Patents," *Science and Public Policy*, Vol. 46, No. 4, August 2019, pp. 589–598.

Schmid, Jon, and Fei-Ling Wang, "Beyond National Innovation Systems: Incentives and China's Innovation Performance," *Journal of Contemporary China*, Vol. 26, No. 104, 2017, pp. 280–296.

Thomson Reuters, *Top 100 Global Innovators*, 2011.

U.S. Department of Defense, *Summary of the 2018 National Defense Strategy: Sharpening the American Military's Competitive Edge*, Washington, D.C., 2018.

Wuchty, Stefan, Benjamin F. Jones, and Brian Uzzi, "The Increasing Dominance of Teams in Production of Knowledge," *Science*, Vol. 316, No. 5827, May 2007, pp. 1036–1039.

About This Research Report

This report describes a multimethod analytic approach to assessing national standing in an emerging technological field. It illustrates how patent and publication data can provide insight regarding the evolution of technologies of relevance to decisionmakers in the Department of Defense and the Intelligence Community.

The approach described in this report will be employed to evaluate emerging commercial technologies, identified by the Department of Defense sponsor, that might have implications for military and intelligence operations. The sponsor directed the research team to use open sources of scientific and technological activity to produce an easy-to-employ and generalizable method for assessing technological developments in the commercial domain that could affect national security.

This report complements a set of focused short analyses of critical emerging commercial technologies. The set covers the following topics: the use of patent data to assist in understanding global trends in emerging technologies, quantum technology, technologies for enhancing human performance, semiconductor technology, the intersection of artificial intelligence and cybersecurity, and deepfake generation and detection technology. These focused assessments have two objectives. First, they are designed to provide open source background for Department of Defense and Intelligence Community officials on commercially developed technologies that could have an impact on future military and intelligence operations. Second, these focused assessments are designed to provide background for the multimethod analysis approach using a combination of data on patent filings, citations, scientific collaborations, and organizational capacity.

At the sponsor's direction, the RAND Corporation will employ this approach to examine emerging commercial technology developments in foreign countries as a way of anticipating technological developments that could affect military and intelligence operations. Defense and Intelligence Community officials will benefit from this approach as one source of validated data to inform decisionmaking about investments in research and development, collection priorities, collection targeting, foreign material acquisition, and demand for scientific talent. The research reported here was completed in July 2021 and underwent security review with the sponsor and the Defense Office of Prepublication and Security Review before public release.

RAND National Security Research Division

This research was sponsored by the Department of Defense and conducted within the Cyber and Intelligence Policy Center of the RAND National Security Research Division (NSRD), which operates the National Defense Research Institute (NDRI), a federally funded research and development center sponsored by the Office of the Secretary of Defense, the Joint Staff, the Unified Combatant Commands, the Navy, the Marine Corps, the defense agencies, and the defense intelligence enterprise.

For more information on the RAND Cyber and Intelligence Policy Center, see www.rand.org/nsrd/intel or contact the director (contact information is provided on the webpage).

For more information on this publication, visit www.rand.org/t/RRA1482-3.

www.rand.org

RR-A1482-3