Data Brief

The Inigo Montoya Problem for Trustworthy AI

The Use of Keywords in Policy and Research

Authors
Emelia S. Probasco
Autumn S. Toney
Kathleen T. Curlee

CSET CENTER for SECURITY and EMERGING TECHNOLOGY

June 2023

Executive Summary

While top-level principles regarding trustworthy, ethical, and responsible artificial intelligence and machine learning (ML) are critical to the formation of international norms, so too is the detailed work of the academic and research communities in establishing precise framings, techniques, and tools that will help create or assess trustworthy AI. An obvious interest among policymakers, therefore, is to understand and assess where the technical community may be making progress that can be harnessed, and where policymakers would do well to support or otherwise incentivize more activity.

Understanding where progress is being made in developing trustworthy AI is complicated. First, the field of AI/ML is rapidly advancing, with new tools and techniques emerging in rapid succession. Second, trustworthy AI is a nascent, multifaceted concept that is hard to bound. And third, there is the possibility that policymakers and technical researchers may be talking past each other, at least in the published literature, by using the same key terms to describe trustworthy AI—but with different meanings ascribed to these terms.

This paper aims to assist technology policymakers interested in trustworthy AI by examining the use of trustworthy AI keywords in AI research publications and whether or not that use overlaps with how the research and development community uses the same terms. Drawing on the National Institute of Standards and Technologies' AI Risk Management Framework (NIST AI RMF), a set of terms related to trustworthy AI is defined, and 2.3 million AI-related research publications between 2010 and 2021 are analyzed, with the following findings:

- Roughly 14 percent of AI papers between 2010 and 2021 include at least one of 13 trustworthy AI keywords (322,209 keyword papers). The growth in the number of publications using these terms exceeds the growth of research on AI generally in the past five years.

- A review of the titles and abstracts of the most cited papers with a trustworthy AI keyword in 2021, reveals that researchers are using most of the keywords in ways that align with the intent of the NIST AI RMF. However, tracking trends in trustworthy AI research through keywords can be misleading, because not all papers that use a keyword for trustworthy AI actually discuss that subject, and some keywords are used in different contexts more often than others. For example:

- The keywords *reliability* and *robustness* are the most frequently mentioned trustworthy AI terms in publications, and most of the titles and abstracts reviewed for this study indicate that the terms are used in ways that align with NIST's AI RMF. These terms may appear frequently in part because they are generally expected evaluation metrics used widely in AI research. This trend was noted in a review of titles and abstracts. However, in the case of *reliability*, a significant minority of papers using the term do so in the context of research on how AI could improve the reliability of a non-AI system.

- Like *reliability*, *safety*, *security*, and *resilience* are also terms frequently used with varying meanings. While most of the titles and abstracts reviewed for this study use these terms in ways that align with NIST's AI RMF definitions, a significant minority use them in research on how AI could improve the reliability, safety, security, and/or resilience of a non-AI system.

- While the keyword *bias* is frequently used in policy conversations in the context of mitigating or avoiding the harmful effects of discrimination, in AI publications it has two main uses, one technical to describe meaningful components of an algorithm, and the other to describe unfair discrimination. NIST's definition accounts for both, though it focuses on harmful bias mitigation in the sense of unfair discrimination. Researchers are evenly split between these two options in how they use the word *bias*.

- Many publications that use the terms *explainability*, *interpretability*, *transparency*, and *accountability* are referencing how to develop AI models and systems that an end-user can trust, specifically in the context of the Explainable AI (XAI) research area. This is interesting, because, while trustworthy AI is not currently considered a research area, XAI has developed into one. Although the terms *explainability* and *interpretability* can be confusing to non-experts, they appear to be distinct and core to the XAI area of research.

Background

Top-level policy statements have emphasized the desire for the development of trustworthy, ethical, and responsible AI norms and regulations. The United States, United Kingdom, China, and Japan, to name a few, have all stated their interest in engaging internationally to "promote a shared understanding of responsible AI design, development, deployment, and use through domestic and international engagements."[1] Multilateral organizations have also published guidance documents, from the United Nations Educational, Scientific, and Cultural Organization (UNESCO), to the European Union, to the Organization for Economic Cooperation and Development (OECD), to the Group of 20.[2] These statements are encouraging and could galvanize action, but developers and users may be at a loss to realize those goals until specific processes, frames, standards or technical solutions are developed.

If the convergence of policy goals and technical approaches is needed to promote international norms and standards for trustworthy AI, then policymakers can benefit from an understanding of the state of the research around these technical and socio-technical concepts. For example, government policies on privacy should be informed by and co-evolve with computer science privacy-enhancing techniques, ranging from anonymization to federated learning. On bias, as another example, policymakers have voiced concern about the negative effects of bias in official statements, but data scientists are still debating what constitutes bias that leads to discrimination and how to develop algorithms with biases and weights that avoid discrimination.

Knowing where researchers may be focusing their efforts might help guide funding and investment decisions, appropriately evolve policies, and possibly lend insights into the relative focus areas of different institutions or even nations—but knowing where researchers are making progress is challenging. First, because research and development in artificial intelligence and machine learning (AI/ML) is rapidly expanding and evolving, its relevance to trustworthy AI is a quickly moving target. Second, trustworthy AI is a nascent, multifaceted concept that is not bound within a singular research area. And third, the key terms of trustworthy AI are not commonly understood across communities and are still evolving as well. Even where common terms are beginning to be visible across national and international policy documents, those same terms are not necessarily understood in the same way by the technical community. Hence policymakers and the technical community could find themselves in what this paper refers to as the "Inigo Montoya problem," referring to a character in the novel and film *The Princess Bride*—specifically a scene in which Inigo understands the word

inconceivable differently than the main character, Vizzini. Similarly, the policymaking and technical communities are ascribing different meanings to the same term, leading to obvious problems for ensuring the trustworthiness of AI/ML and its application.

To better understand the scope and potential impact of the Inigo Montoya problem on the research and policy communities, this study examines how policymakers use keywords which are considered characteristics of AI trustworthiness (henceforth referred to as "trustworthy AI keywords"), and how researchers use those same keywords in their publications. Keywords are drawn from the National Institute of Standards and Technologies' AI Risk Management Framework.[3] The NIST AI RMF was developed through a multi-year engagement with government, academia, and industry, both in the United States and abroad. This consensus-driven, living document lays out voluntary steps any organization can take to manage the risk of harm by AI systems, and it is complemented by a growing online resource center of tools and specific guidance.[4] As a part of the document, NIST lists a set of characteristics of trustworthy AI. In its discussion of those characteristics, NIST defines several more characteristics. Given NIST's broad engagement on the subject of trustworthy AI across the policy and technical communities, the characteristics it lists and defines serve as a foundation from which to draw policy-relevant trustworthy AI terms.

> "You keep using that word, I do not think it means what you think it means."
> – Inigo Montoya, *The Princess Bride*

Identifying AI Research with Trustworthy AI Terms

Characterizing and defining the sum total of research on trustworthy AI/ML is beyond the scope of this paper. Instead, the aim here is to examine the body of AI-related research that explicitly uses one of the terms extracted from the NIST AI RMF (Box 1). This scoping, while not exhaustive, touches on a broad array of issues that the technology and policy communities have identified as important.

Using this set of trustworthy AI terms extracted from the NIST AI RMF, researchers for this study conducted a keyword search of publications between 2010 and 2021 that CSET's AI classifier identified as AI-related.[5] This keyword search was performed over CSET's merged corpus of scholarly literature, including Digital Science Dimensions, Clarivate's Web of Science, Microsoft Academic Graph, China National Knowledge Infrastructure, arXiv, and Papers With Code.[6] Publications were selected for analysis if

they were classified as AI-relevant and their title or abstract contained at least one of the trustworthy AI keywords.[7] The term search in English yielded English-only publications from around the world, but many international authors publish their academic work in English.[8]

Two of the NIST trustworthy AI terms used in the initial search, *valid* and *accurate*, resulted in overly broad applicability in the field of AI research—which implicitly aims for valid results or accurate performance.[9] As a consequence, these two terms are eliminated from this analysis, though both concepts are important for the development of trustworthy AI and future research. The keyword *trust* was added to NIST's list of key characteristics as well, since NIST used it as an overarching term. With *trust* included and *valid* and *accurate* excluded, the search conducted for this study resulted in a set of 322,209 research publications, referred to here as trustworthy AI keyword publications.

Box 1. NIST AI RMF Terms Used for Keyword Search

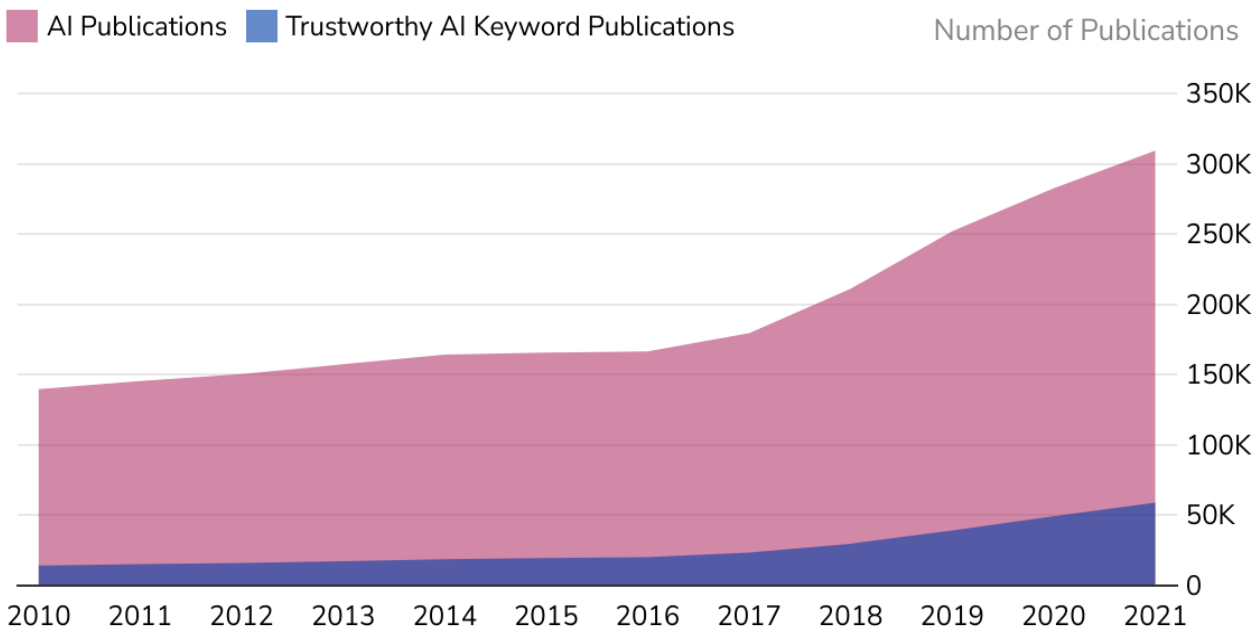
- Accountability, Accountable
- Explainability, Explainable
- Interpretability, Interpretable
- Reliability, Reliable
- Safe/Safety
- Resilience
- Trust
- Bias*
- Fairness
- Privacy*
- Robustness†
- Secure /Security
- Transparency

* Officially, NIST uses the terms *bias-managed* and *privacy-enhanced*.
† *Robustness* is defined within NIST's discussion of *valid and reliable*.

Publication Analysis

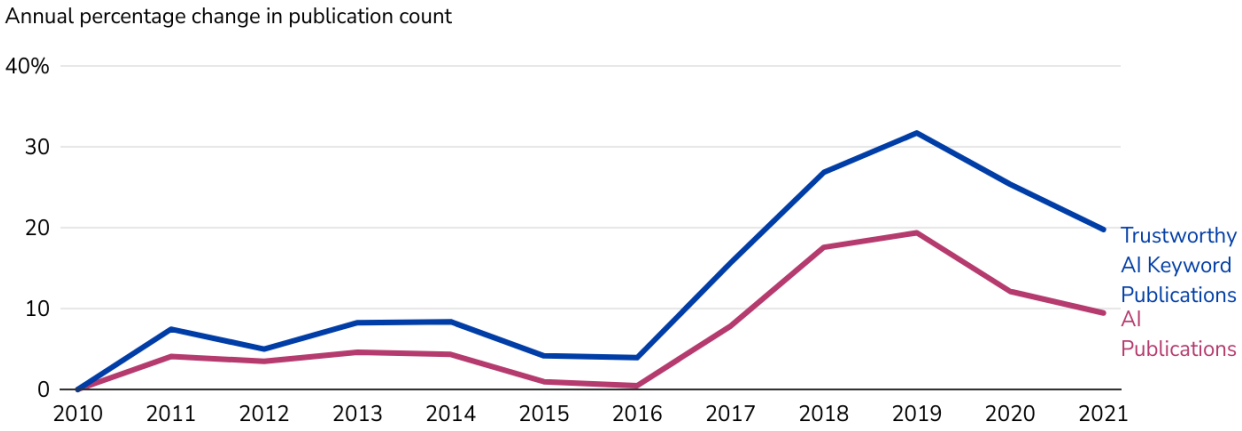
The trustworthy AI keyword publications identified for this study represent 14 percent of all the AI publications in CSET's merged corpus (322,209 papers out of 2,324,124 AI papers: see Figure 1). Additionally, the rate of growth of AI papers with trustworthy AI keywords exceeds the increase of AI papers overall, especially since 2016 (Figure 2). It is important to note when viewing these charts that they represent the appearance of trustworthy AI keywords in research papers, and not the count of papers focused on trustworthy AI topics. A forthcoming CSET paper will describe an approach to identifying research clusters with studies related to trustworthy AI topics.

Figure 1. Number of AI and Trustworthy AI Keyword Publications (2010–2021)



Source: CSET merged corpus of scholarly literature including Digital Science Dimensions, Clarivate's Web of Science, Microsoft Academic Graph, China National Knowledge Infrastructure, arXiv, and Papers With Code

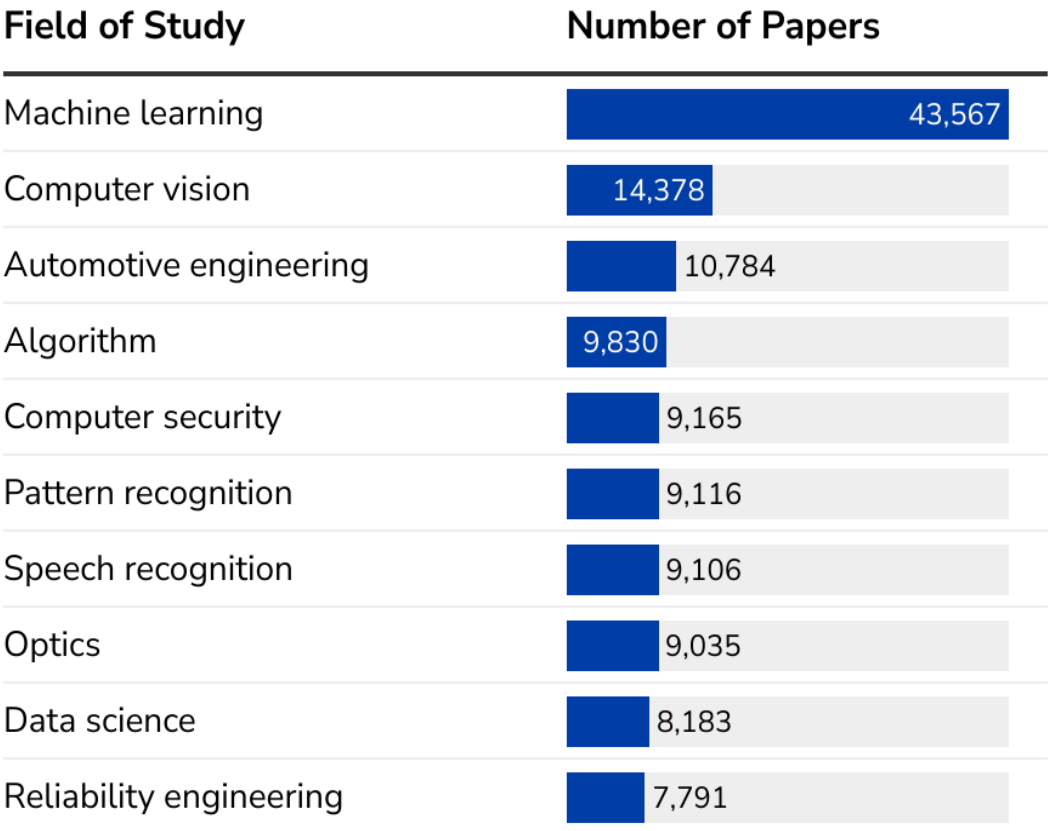
Figure 2. Annual Percentage Change in Number of Publications (2010–2021)



Source: CSET Merged Corpus

The CSET fields of study labels, derived from Microsoft Academic Graph (MAG), were also examined for the trustworthy AI keyword papers that were selected.[10] This hierarchical field of study taxonomy provides general, "level 0" labels (19 in total, including, for example, physics and art) and more granular "level 1" labels (292 in total). Of the trustworthy AI keyword publications in the corpus used for this study, 90 percent are categorized as "Computer Science" by MAG's most general field of study label (level 0), and the top 10 most common level 1 (more granular) fields of study are displayed in Figure 3. The 10 fields show that papers in the corpus used here span AI-related topics, including applications (such as computer vision and automotive engineering), fields (data science, for example), and cross-cutting topics (such as reliability engineering).

Figure 3. Trustworthy AI Keyword Publications by Top 10 Most Common Level 1 Fields of Study (2010–2021)



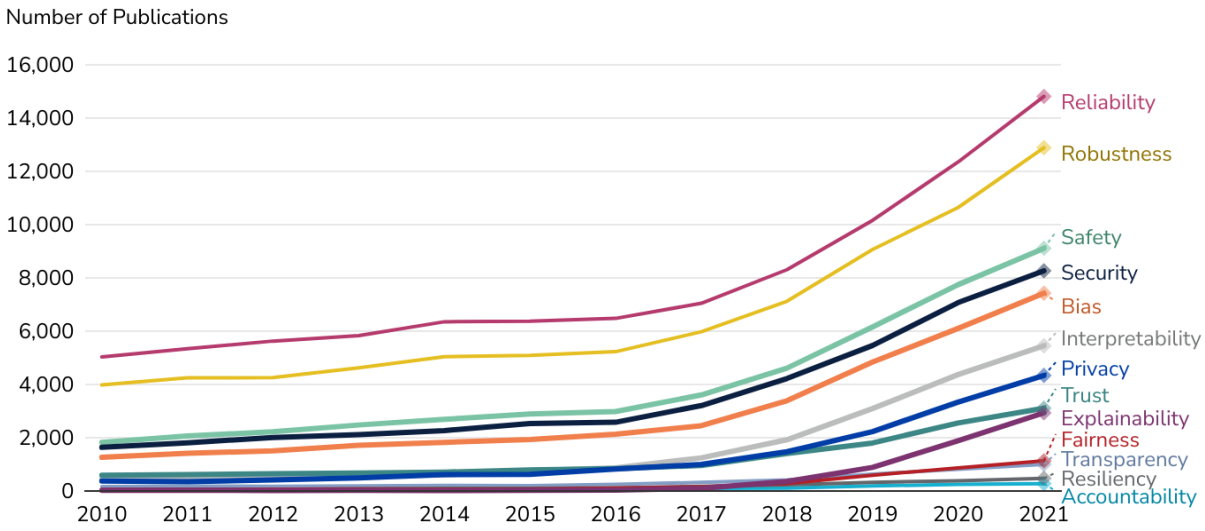
Source: CSET Merged Corpus

Trustworthy AI Term Use: Frequency and Variability

Overview of Publications and Conferences

The occurrence of trustworthy AI keywords were evaluated within the corpus by counting their appearance in AI-relevant titles and abstracts over time (Figure 4). Of these 332,209 trustworthy AI keyword publications, 12 percent included more than one of the keyword terms from the search used here. As just one example explored later in this paper, abstracts could include both the terms *privacy* and *security*. When analyzing term frequency, papers with more than one relevant keyword were counted for each one, and no attempt was made to disaggregate them into a single "most correct" category—therefore, some papers were counted in more than one category.

Figure 4. Trustworthy AI Keyword Use in Publications (2010–2021)



Source: CSET Merged Corpus

While observing the counts for each term begins to provide some sense of their appearance in research papers, it is important to recall that these counts do not communicate the context in which the word was used, and whether or not its use aligns with NIST's. In other words, without contextual understanding, a policymaker may encounter the Inigo Montoya problem and incorrectly believe that the terms are being used in ways that align with their own policy goals.

To understand the potential scope of this problem, the 50 most cited papers for each term in 2021 were manually reviewed (650 publications total). This study does not

claim that the distribution of topics and issues in these papers are entirely representative of the whole, but the review does provide insights into the range of potential uses, and helps to explain some of the limitations of tracking the use of trustworthy AI terms. Summaries of those explorations follow; they are critical to better understanding the term frequencies in Figure 4.

Reliability

Among the trustworthy AI terms that were counted in the research literature for this study, *reliability* and *robustness* consistently led as the most frequently mentioned from NIST's list of trustworthy AI terms between 2010 and 2021. Their frequent appearance may be attributed to the wide range of their usage. Of the 50 most cited 2021 papers in this study's corpus that used the word *reliability*, more than half used the term in a way that reflects NIST's definition: "the ability of an item to perform as required without failure, for a given time interval, under given conditions."[11] Papers that used the term in a way that aligned with NIST often did so to assert or document the reliability of a specific approach or application of AI/ML. For example, abstracts frequently included the phrase "our method produces *reliable* results," or "we confirmed the *reliability* of our models."

Approximately one-third of the articles, however, were concerned with the use of AI to improve the reliability of non-AI systems (for example, the reliability of COVID-19 detection or the reliability of a tunnel-boring machine), and thus used the term *reliability* but not with respect to AI. This revealed the focus, or perhaps enthusiasm, of researchers for the promise of AI to improve the reliability of current, non-AI systems, and contrasts with the policymakers' concern that AI systems must be made more reliable.

The remaining titles and abstracts using the term *reliable* included studies of methods for improving reliability in AI systems, as well as those mentioning the term in relation to an ancillary goal. To convey a sense of the range of publications that included the word *reliable* or *reliability* in the title or abstract, Box 2 provides a brief sample.

Box 2. Sample Publication Titles Using *Reliability* in the Title

- *Exploring the Effect of Image Enhancement Techniques on COVID-19 Detection: Using Chest X-Ray Images Sensor and Sensor Fusion Technology in Autonomous Vehicles: A Review*
- *The 10-M Crop Type Maps in Northeast China during 2017–2019: The Matthews Correlation Coefficient (MCC) Is More Reliable than Balanced Accuracy, Bookmaker Informedness, and Markedness in Two-Class Confusion Matrix Evaluation*

Robustness

Under the NIST and International Standards Organization (ISO) definitions, *robustness* is the "ability of a system to maintain its level of performance under a variety of circumstances."[12] Of the 50 most cited AI-related papers using the term in 2021, many did so in a way consistent with this definition and as an evaluation metric for a particular proposal (i.e., measuring the robustness of a federated learning approach). More than was observed for other keywords, assertions of robustness were frequently in relation to that of another method or approach, indicating that measures of robustness were perceived as relative rather than absolute. Other papers specifically focused on methods to improve robustness in algorithms, and yet others used the word as part of a statement about the importance of further developing robustness, either generally or specifically, for an application.

Safety

Following *reliability* and *robustness*, the keyword *safety* was the next most mentioned term. NIST quotes the ISO guidance in its AI RMF, specifying that the characteristic of safety requires that AI systems not, "under defined conditions, lead to a state in which human life, health, property, or the environment is endangered."[13] Roughly half of the 50 papers analyzed used the word in a way that aligned with the definition in the NIST AI RMF. The other half were about the application of AI to improve the safety of a current process or technology, for example, in construction, medicine, or manufacturing. Overall, about one-third of the 50 papers using the term *safety* were concerned with the use of AI to address that of autonomous vehicles (mostly cars but also seagoing

vessels). There were notable uses of the word in connection with the security or privacy of personal data, and several papers linked other NIST terms to *safety*, including *robustness*, *reliability*, and *explainability* (analysis of term co-occurrence follows later in this study).

Box 3. Sample Publication Titles using Safety in the Title

- *Second Opinion Needed: Communicating Uncertainty in Medical Machine Learning*
- *Physical Safety and Cyber Security Analysis of Multi-Agent Systems: A Survey of Recent Advances*
- *Safety Assurance Mechanisms of Collaborative Robotic Systems in Manufacturing*
- *Detection Algorithm of Safety Helmet Wearing Based on Deep Learning*

Security

NIST defines *security* as "AI systems that can maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorized access and use."[14] The use of the word *security* in research literature aligned with NIST's definition, but not as frequently as did other keywords. For example, similar to the term *safety*, roughly half of the top-cited 50 papers from 2021 that included the word *security* covered the application of AI to problems in security, usually cybersecurity or the security of the Internet of Things. Another eight of the 50 papers used the word *security* in a tangential reference, for example, in describing a paper about food security as a motivation for improving a crop monitoring algorithm. The remaining papers used *security* in a way that was more aligned with NIST's definition. For example, there were assessments of the relative security of a particular AI method and proposals to improve the security of an AI-enabled system.

Bias

NIST accounts for a broad definition of *bias* in its AI RMF, pointing out that "bias is not always a negative phenomenon," namely when discussed in a technical sense.[15] That said, NIST connects *bias* to *fairness* and is most concerned that bias in AI not create,

perpetuate, or amplify harm to individuals. But the appearance of the term *bias* in AI research is not necessarily focused on harmful discrimination. Rather, in research the term reflects NIST's broader statement that bias is not necessarily harmful and that it is in fact an essential part of building an AI algorithm. For example, nearly two-thirds of the 50 abstracts examined for this study referenced the role or need for "inductive bias," or the "weights and biases" necessary to develop an algorithm. The remaining third of the titles and abstracts using the word *bias* addressed racial or gender bias or methods for addressing bias through techniques connected to other keywords such as *explainability*.

Interpretability and Explainability

Interpretability and *explainability* are grouped here because, while distinct, they are intimately connected and can be challenging for non-specialists to separate in the research literature. NIST makes a clear distinction between the need to properly interpret the recommendation of an AI system (*interpretability*) and the related need to represent "the mechanisms underlying AI systems' operation" (*explainability*).[16] NIST's definitions of these terms and the distinction it makes are not universally accepted, however. For example, Amazon's definitions of the two terms are nearly reversed,[17] and some of the publications identified for this review make clear that *explainability* and *interpretability* are synonymous for some researchers.[18]

Notably, the research area of Explainable AI (XAI) focuses on designing AI systems that the end-user can trust, with both *interpretability* and *explainability* as critical components. The majority of the top 50 2021 publications reviewed here used the term *explainable* in the XAI phrase, and *explainability* to reference the ability to identify how the model makes decisions. Most publications that mention the term *interpretability* did so to describe the evaluation of or improvement to deep learning classification outputs, specifically in regards to an XAI framework. Several publications surveyed deep learning models that either lacked interpretability or asserted the state-of-the-art for interpretable outputs.

Privacy

The vast majority of the 50 top-cited AI papers that included the word *privacy* in 2021 were aligned with the NIST AI RMF and concerned techniques and approaches that would improve the privacy of user or device data, often for a very specific use. Moreover, among papers concerned with improving privacy, most discussed either the potential for federated learning to improve privacy or else ways to improve federated

learning approaches to minimize the loss of performance that has been observed as a tradeoff for the technique. Overall, one-third of the papers focused on privacy issues for the Internet of Things, and slightly less than a third focused on privacy in the medical field. This may be attributed at least in part to a high concern over the sharing of data in the midst of the COVID-19 pandemic (10 papers specifically mentioned COVID-19).

Trust

NIST's AI RMF is concerned on the whole with the creation of trustworthy AI, and so, in a way, defines trustworthy AI by the key characteristics examined in the rest of this paper. Given this overarching concern with the trustworthiness of AI systems, for this analysis the word *trust* is included in the examination of papers and abstracts. Unsurprisingly, however, the vast majority of appearances of the word *trust* in the titles and abstracts of the 50 most cited 2021 papers included at least one of the other NIST key characteristics, especially *privacy*, *security*, and *explainability* (40 of the 50 papers). Of those without the presence of another term, these papers were still concerned with the trustworthiness of an AI system as NIST would consider it. More on the term co-occurrence follows in the next section.

Fairness

The 50 most cited 2021 papers including the word *fairness* in the title or abstract were mostly well aligned with NIST's definition, but they also echoed NIST's assertion that fairness is a socio-technical issue that can vary across groups or cultures. Papers in this group included those that strongly linked fairness to bias and discrimination, others that explored fairness beyond bias or discrimination, some that examined definitions of fairness, and finally papers that focused on technical fairness (similar to the technical bias papers discussed previously). There were also papers explicitly examining or evaluating the fairness of a particular application of AI/ML.

Transparency

Transparency largely appeared in the 50 most cited AI papers examined here in ways that aligned with NIST's guidance that "information about an AI system and its outputs is available to individuals interacting with such a system."[19] More than two-thirds of these titles and abstracts also included another of NIST's key characteristics, especially *explainability* and/or *interpretability*. Several of the papers in this subgroup discussed the relative merits of post-hoc explainability for transparency as well as "transparent algorithms." Another grouping of the papers detailed the importance of transparency to

trust and/or the adoption of AI/ML. Several of the top papers included specific proposals to improve AI transparency for a given use or to improve the transparency of datasets.

Resiliency

Two-thirds of the 50 titles and abstracts using the word *resilience* did so in a way that aligned with NIST's ("withstand[ing] unexpected changes in their environment or use").[20] For example, papers mentioned the term as an evaluation metric, made proposals for general resiliency approaches such as digital twins, and voiced concerns about fault tolerance as a component of resilience. Similarly to the case for the terms *safety* and *security*, though to a lesser extent, approximately one-third of the top 50 papers were not about resilience as a characteristic of AI, but rather about the application of AI to improve resilience in a non-AI context. For example, paper topics in the group of 50 covered resilience in the face of climate change, the resilience of robots to environmental shifts, supply chain resilience, and even the application of AI to monitor pigs for indications of animal resiliency.

Accountability

The term *accountability* appeared in the top 50 titles and abstracts in ways that align with NIST's definition in the AI RMF. Roughly one-third of the titles and abstracts reviewed here also included another trustworthy AI term, especially *explainability*, *interpretability*, and *transparency*, or the umbrella concept of XAI, which echoes NIST's statement that "accountability presupposes transparency."[21] However, unlike some of the other trustworthy AI keywords, a majority of the top 50 most cited publications were studies of the concept or importance of accountability, and not specific proposals to improve or address accountability in a given case. This stood in contrast to the other keyword papers examined, where specific approaches or proposals were related to the key characteristic. An effort to understand what this might mean for efforts to realize *accountability* in AI systems could be useful research for policymakers.[22]

Term Usage in AI/ML Conferences

To further address the question of how the technical community's use of terms may align with NIST's AI RMF, key characteristic term use was examined in prestigious AI conferences' calls for papers between 2019 and 2023. Specifically, conference calls for papers and/or conference-assigned keywords for papers were sought, because these reflect specific research areas of interest for that year's conference proceedings. Table

1 shows the 11 trustworthy AI terms that were found across 11 top AI/ML annual or biannual conferences.[23] Table 2 tracks the progression of official conference keywords added to the Association for the Advancement of Artificial Intelligence conference from 2019 to 2023 (AAAI has one "main track" rather than several subject specific tracks, and author submissions must include one of the conference's official keywords). In each instance, the terms are used in ways consistent with NIST's AI RMF descriptions.

Notably, the use of key terms in calls for papers at major AI conferences does not mirror their use in the literature. For example, *reliability* and *resilience* do not appear in conference track titles, even though the count of publications with trustworthy AI terms used for this study shows that *reliability* is the most frequently used term in publication titles or abstracts. By contrast, *accountability*, *fairness*, *interpretability*, and *privacy* occur more frequently in the conference track titles or descriptions than they appear in publications. That major AI conferences are using these terms in ways that align with NIST's AI RMF may be good signs for the future of trustworthy AI research, given the influence that AI conferences and their calls for papers have on motivating research.

Table 1. Trustworthy AI Keyword Use in Major AI Conference Calls for Papers over Time, 2019–2023

	2019	2020	2021	2022	2023
Accountability	0	3	5	5	6
Bias	0	1	2	2	2
Explainability	1	1	3	5	3
Fairness	0	4	7	7	6
Interpretability	0	3	4	5	5
Privacy	5	6	6	7	8
Robustness	0	3	2	2	3
Safety	0	1	2	2	2
Security	3	4	1	2	3
Transparency	0	2	4	4	5
Trust	1	2	1	2	2

Source: Authors' calculations.

Table 2. AAAI Keywords for Main Track Research Papers, 2019–2023

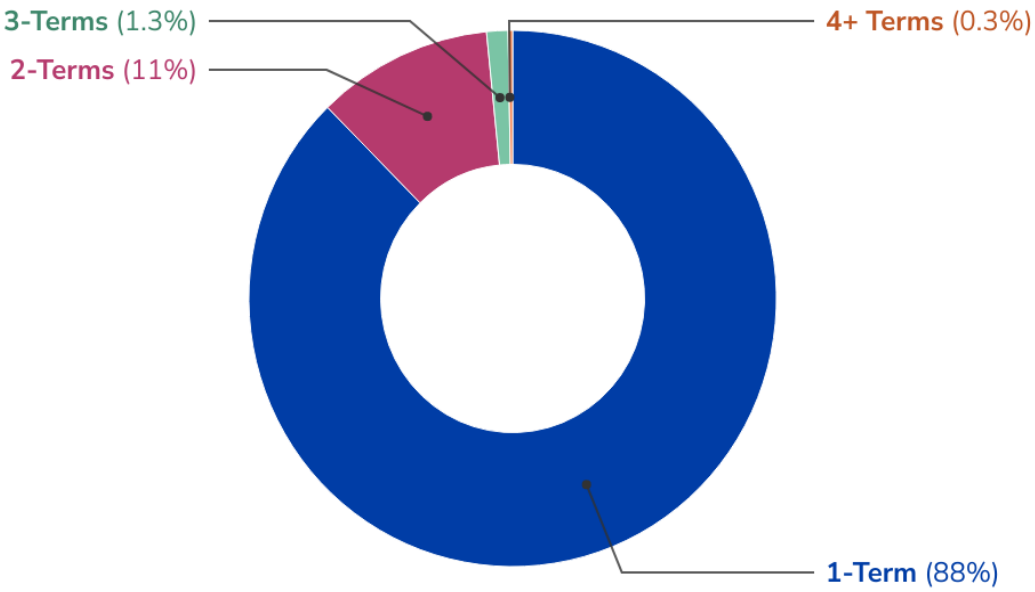
2019	2020	2021	2022	2023*
Explainability	Explainability	Explainability	Explainability	Explainability
Security	Security	Security	Security	Security
	Interpretability	Interpretability	Interpretability	Interpretability
	Privacy	Privacy	Privacy	Privacy
	Robustness	Robustness	Robustness	Robustness
		Accountability	Accountability	Accountability
		Bias	Bias	Bias
		Fairness	Fairness	Fairness
		Safety	Safety	Safety
		Transparency	Transparency	Transparency

*New track added on "Safe and Robust AI."
Source: Association for the Advancement of Artificial Intelligence.

Term Co-occurrence

Many of the key NIST characteristics of trustworthy AI are related conceptually if not technically, as evidenced in this study's review of titles and abstracts. For example, there is widespread concern about insecure systems that violate people's privacy and that biased AI algorithms result in unfair outcomes. To better understand the usage of the NIST key terms in the trustworthy AI keyword corpus, with its total of 322,209 publications between 2010 and 2021, the term co-occurrence was examined across four different grouping sizes (one term mentioned, two, three, and four or more terms mentioned). Figure 5 displays the results of the number of trustworthy AI keyword publications by their term usage grouping.

Figure 5. Term Co-occurrence in Trustworthy AI Keyword Publications, 2010–2021



Source: CSET Merged Corpus

A review of the trustworthy AI keyword publications between 2010 and 2021 revealed that the majority (88 percent) mentioned just one trustworthy AI term in the title and/or abstract. This results in 12 percent of trustworthy AI keyword publications containing more than one term mention. Table 3 shows the top five most frequently appearing groupings of 2-terms and 3-terms, where frequency represents the number of papers in which the listed terms co-occur. The 4-or-more-term grouping has drastically smaller usage percentages; thus, it is not included in Table 3.

Table 3. Top 5 2-Term and 3-Term Groupings by Number of Papers in the Trustworthy AI Keyword Corpus, 2010–2021

2-Term		3-Term	
Terms	**Frequency**	**Terms**	**Frequency**
robustness, reliability	4,500	reliability, security, safety	286
reliability, safety	4,195	robustness, reliability, safety	274
security, safety	2,775	trust, security, privacy	229
security, privacy	2,621	security, privacy, safety	177
reliability, security	2,223	robustness, reliability, security	150

Source: CSET Merged Corpus.

These co-appearances may offer a clearer understanding of where relationships may exist between terms, but the linkages in papers must be treated with caution. For example, the term co-appearance may connote two words that are technically distinct but closely linked conceptually. But the co-appearance could equally be a result of simply the frequency of use (as in the case of *reliability* and *robustness*, the two most frequently appearing terms overall), or it could be explained by the term confusion (i.e., conflating security and privacy), or even by the term variance (as detailed earlier, the top 50 AI-related papers from 2021 that used the word *security* did so in both the sense of AI security and using AI for security applications).

As another approach to understanding how terms may be linked technically or conceptually, calls for papers by major AI-related conferences were revisited, this time to see how those calls group keyword terms (Table 4). Six of these calls for papers in 2023 include *fairness* and *accountability* grouped together; of those six, four also include *transparency*. These co-appearances do not reflect the observations of this study's publications analysis, but they do reflect NIST's concerns and definitions as defined in the AI RMF, and they indicate important technical and conceptual relationships between *fairness*, *accountability*, and *transparency*.

Table 4. Trustworthy AI Terms and Groupings in Calls for Papers at Major AI-Related Conferences, 2023

Conference	Keyword Mentions in Calls For Papers/Research Tracks
Association for the Advancement of Artificial Intelligence Conference (AAAI)	Track on **Safe** and **Robust** AI
Conference on Computer Vision and Pattern Recognition (CVPR)	**Transparency**, **fairness**, **accountability**, **privacy**, and ethics in vision;
	Explainable computer vision
International Conference on Computer Vision (ICCV)	**Fairness**, **privacy**, ethics, social-good, **transparency**, **accountability** in vision
International Conference on Machine Learning (ICML)	**Trustworthy** Machine Learning (**accountability**, causality, **fairness**, **privacy**, **robustness**, etc.)
Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)	Foundations:...personalization**, security and privacy,** visualization**; fairness, interpretability,** ethics and **robustness**
Conference on Neural Information Processing Systems (NeurIPS)	Social and economic aspects of machine learning (e.g., **fairness**, **interpretability**, human-AI interaction, **privacy**, **safety**, strategic behavior)
Annual Meeting of the Association for Computational Linguistics (ACL)	**Interpretability** and Analysis of Models for NLP
Empirical Methods in Natural Language Processing (EMNLP)*	**Interpretability**, Interactivity and Analysis of Models for NLP
Special Interest Group on Information Retrieval (SIGIR)	**Fairness**, **Accountability**, **Transparency**, Ethics, and **Explainability** (FATE) in IR. Research on aspects of **fairness** and **bias** in search and recommender systems.
International World Wide Web Conference (WWW)	**Fairness**, **Accountability**, **Transparency**, and Ethics on the Web
	Security, **Privacy**, and **Trust**

Source: Association for the Advancement of Artificial Intelligence Conference, Conference on Computer Vision and Pattern Recognition, International Conference on Computer Vision, International Conference on Machine Learning, Special Interest Group on Knowledge Discovery and Data Mining, Conference on Neural Information Processing Systems, Annual Meeting of the Association for Computational Linguistics, Empirical Methods in Natural Language Processing, Special Interest Group on Information Retrieval, International World Wide Web Conference.

Conclusion

A broad array of agencies, industries, institutions, and individuals are contributing to the advancement of AI in ways that will alter and affect communities large and small—if not all of humanity. The widespread impact explains why institutions such as NIST are dedicating considerable time and effort to establish and define principles and characteristics that will avoid potential harm. Defining, much less achieving trustworthy AI characteristics, however, is a societal effort that necessitates clear communication through consensus on the meaning of field-specific terms. This is how the Inigo Montoya problem may be avoided. Through its publications, the research community is offering perspectives and approaches that will help society better avoid harm and achieve a positive impact. But to achieve trustworthy AI, these researchers must share a language—and an understanding—common to us all in our societal ambitions for trustworthy AI.

Examining the use of trustworthy AI terms in research publications can lead to better understanding the frequency of their appearance, how international research efforts may align with high-level policy goals, and where a focus on research in and development of trustworthy AI is evident. In the titles and abstracts reviewed for this study, each term has instances where its use aligns with NIST's AI RMF, but some terms have more variability than others. *Safety*, for example, refers to both the safety of an AI system and the opportunity for an AI system to improve the safety of a non-AI system. *Explainability* and *interpretability* appear frequently together, often as a part of research on XAI, though the two can easily be conflated. Finally, *accountability* appears the least frequently in the literature reviewed, and the titles and abstracts of the top papers from 2021 indicate that the concept may be less developed than other trustworthy AI characteristics, although large AI conferences may be driving more research in this area through multiple calls for papers.

Policymakers should remain aware of how the research community is using trustworthy AI terms so they can progress toward a more common understanding and track the development of commonly accepted techniques and frameworks. In publications around the world, researchers are using the terms analyzed here with some frequency, which in itself offers reason for hope that serious thought is being devoted to developing technology that will ensure trustworthy AI. But policymakers are essential for success. To communicate clearly and develop effective solutions, policymakers and researchers must share the same terminology and interpret it in the same way to avoid the Inigo Montoya problem.

Authors

Emelia Probasco is a senior fellow at the Center for Security and Emerging Technology, where Autumn Toney is a data research analyst and Kathleen Curlee is a research analyst.

Acknowledgments

For feedback and assistance, the authors thank Catherine Aiken, Tessa Baker, Shelton Fitch, Melissa Flagg, Heather Frase, Alex Friedland, Rebecca Gelles, Margarita Konaev, Cara LaPointe, Igor Mikolic-Torreira, Dewey Murdick, Micah Musser, and Lynne Weil.

The authors are solely responsible for all errors.

Document Identifier: doi: 10.51593/20230014a

Endnotes

[1] DoD Responsible AI Working Council, *U.S. Department of Defense Responsible Artificial Intelligence Strategy and Implementation Pathway* (Washington, D.C.: Department of Defense, 2022), https://www.ai.mil/docs/RAI_Strategy_and_Implementation_Pathway_6-21-22.pdf. Secretary of State for Digital, Culture, Media and Sport (UK), *National AI Strategy* (London: Office for Artificial Intelligence, HM Government, September 2021), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1020402/National_AI_Strategy_-_PDF_version.pdf. Graham Webster, Rogier Creemers, Paul Triolo, and Elsa Kania, trans., "Full Translation: China's 'New Generation Artificial Intelligence Development Plan,'" *New America Blog*, August 1, 2017, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/. Japan, The Conference toward AI Network Society, "AI Utilization Guidelines", August 9, 2019, https://www.soumu.go.jp/main_content/000658284.pdf.

[2] UNESCO, *UNESCO Member States Adopt the First Ever Global Agreement on the Ethics of Artificial Intelligence*, (Paris: United Nations Educational, Scientific and Cultural Organization, November 25, 2021), https://en.unesco.org/news/unesco-member-states-adopt-first-ever-global-agreement-ethics-artificial-intelligence. Carlo Casalone et al., "Human-Centric AI: From Principles to Actionable and Shared Policies" https://www.t20italy.org/2021/09/07/human-centric-ai-from-principles-to-actionable-and-shared-policies-2/. EU "The AI Act" https://artificialintelligenceact.eu/.

[3] NIST, *AI Risk Management Framework 1.0*, (Gaithersburg, MD: National Institute of Standards and Technology, January 2023) https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[4] "Trustworthy & Responsible AI Resource Center," National Institute of Standards and Technology, accessed April 13, 2023, https://airc.nist.gov/Home.

[5] The keyword search approach to the merged corpus differs from previous explorations of safety using a research clusters approach. For an example of the latter approach to examining research relevant to AI safety, see Helen Toner and Ashwin Acharya, "Exploring Clusters of Research in Three Areas of AI Safety" (Center for Security and Emerging Technology, February 2022), https://cset.georgetown.edu/publication/exploring-clusters-of-research-in-three-areas-of-ai-safety/. For more information about CSET's AI classifier, see James W. Dunham, Jennifer Melot, and Dewey A. Murdick, "Identifying the Development and Application of Artificial Intelligence in Scientific Text," ArXiv abs/2002.07143 (2020), https://arxiv.org/abs/2002.07143; and the forthcoming CSET report "Identifying AI Research."

[6] Data sourced from Dimensions, an inter-linked research information system provided by Digital Science (http://www.dimensions.ai). All China National Knowledge Infrastructure content is furnished for use in the United States by East View Information Services, Minneapolis, MN. However, 98.5 percent of CNKI publications are in Chinese, thus this English-only investigation leveraged only a very small percentage of CNKI papers.

[7] Full query here: https://github.com/georgetown-cset/trustworthy-ai-research.

[8] W. Liu, "The Changing Role of Non-English Papers in Scholarly Communication: Evidence from Web of Science's Three Journal Citation Indexes," *Journal of the Association of Learned and Professional Society Publishers* 30, no. 2 (2017): 115–23.

[9] For example, the keyword search for "accurate" returned results that included sentences such as: "Developing an accurate prediction model for housing prices is always needed" (S. B. Jha, R. F. Babiceanu, V. Pandey, and R. Jha, "Housing Market Prediction Problem Using Different Machine Learning Algorithms: A Case Study," ArXiv, abs/2006.10092 [2020]) and "Models that accurately capture fire propagation dynamics greatly help efforts for understanding, responding to and mitigating the damages caused by these fires" (J. Burge, M. Bonanni, M. M. Ihme, and R. L. Hu, "Convolutional LSTM Neural Networks for Modeling Wildland Fire Dynamics," ArXiv, abs/2012.06679 [2020]).

[10] Autumn Toney and James Dunham. "Multi-label Classification of Scientific Research Documents across Domains and Languages," *Proceedings of the Third Workshop on Scholarly Document Processing* (Gyeongju, Republic of Korea, Association for Computational Linguistics, 2022), 105–14.

[11] While this definition appears in NIST's AI RMF 1.0, it is the International Standards Organization's definition: ISO/IEC TS 5723:2022.

[12] ISO/IEC TS 5723:2022.

[13] ISO/IEC TS 5723:2022.

[14] NIST *AI RMF 1.0*.

[15] NIST *AI RMF 1.0*.

[16] NIST *AI RMF 1.0*.

[17] Joe King, Betty Zhang, Hanif Mahboobi, and Shantu Roy, *Model Explainability with AWS Artificial Intelligence and Machine Learning Solutions* (Amazon Web Services Whitepaper, September 10, 2021), https://docs.aws.amazon.com/whitepapers/latest/model-explainability-aws-ai-ml/interpretability-versus-explainability.html.

[18] Giulia Vilone and Luca Longo, "Notions of Explainability and Evaluation Approaches for Explainable Artificial Intelligence," *Information Fusion* 76 (2021): 89–106, https://doi.org/10.1016/j.inffus.2021.05.009.

[19] NIST *AI RMF 1.0*.

[20] NIST *AI RMF 1.0*.

[21] NIST *AI RMF 1.0*.

[22] For an interesting historical study of the conference publications concerning accountability, as well as fairness and transparency, at the Association for Computing Machinery's Conference on Fairness, Accountability, and Transparency (FAcct), see Benjamin Laufer, Sameer Jain, et al., "Four Years of FAccT: A Reflexive, Mixed-Methods Analysis of Research Contributions, Shortcomings, and Future Prospects," in *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea* (New York: ACM), https://doi.org/10.1145/3531146.3533107.

[23] We use the top 13 AI conference list, as identified by CSRankings, and remove two conferences (International Joint Conference on Artificial Intelligence and North American Chapter of the Association for Computational Linguistics) for not having calls for paper in 2023. Our list includes: The Association for the Advancement of Artificial Intelligence Conference (AAAI), Conference on Computer Vision and Pattern Recognition (CVPR), European Conference on Computer Vision (ECCV, held biannually), International Conference on Computer Vision (ICCV, held biannually), International Conference on Machine Learning (ICML), Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD), Conference on Neural Information Processing Systems (NeurIPS), Annual Meeting of the Association for Computational Linguistics (ACL), Empirical Methods in Natural Language Processing (EMNLP), Special Interest Group on Information Retrieval (SIGIR), and International World Wide Web Conference (WWW, now called "The Web Conference"). Of note, not all conferences had guidance on calls for papers each year, particularly in 2019.