Data Brief

Who Cares About Trust?

Clusters of Research on Trustworthy AI

Authors

Autumn S. Toney

Emelia S. Probasco

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

July 2023

Introduction

Given societal interests in the development and governance of artificial intelligence and machine learning systems, policymakers are keen to understand and track research and development on trustworthy AI topics. Tracking this topic is challenging, however, both because of the rapid developments in the field of AI/ML generally and also because trustworthy AI is a constantly developing, multifaceted concept that most scientific research papers address only in part or in a specific application.

In our previous Center for Security and Emerging Technology (CSET) report, *The Inigo Montoya Problem for Trustworthy AI*, we looked at how trustworthy AI keywords can (and cannot) be useful in identifying research papers related to trustworthy AI. In this follow-on analysis, we pair what we learned from our keyword approach with a different way of identifying areas of research: citations. Using CSET's research clusters derived from CSET's merged corpus of scholarly literature, we are able to contextualize the trustworthy AI keyword publications found in the prior report within the scientific research landscape. [1] This approach of identifying research clusters with a high percentage of trustworthy AI keyword papers opens the aperture for finding trustworthy AI research, surfacing papers that do not use a trustworthy-AI term and also overcoming the problems that come with using generic keywords in searches, like *safety* and *security*. Overall, we found:

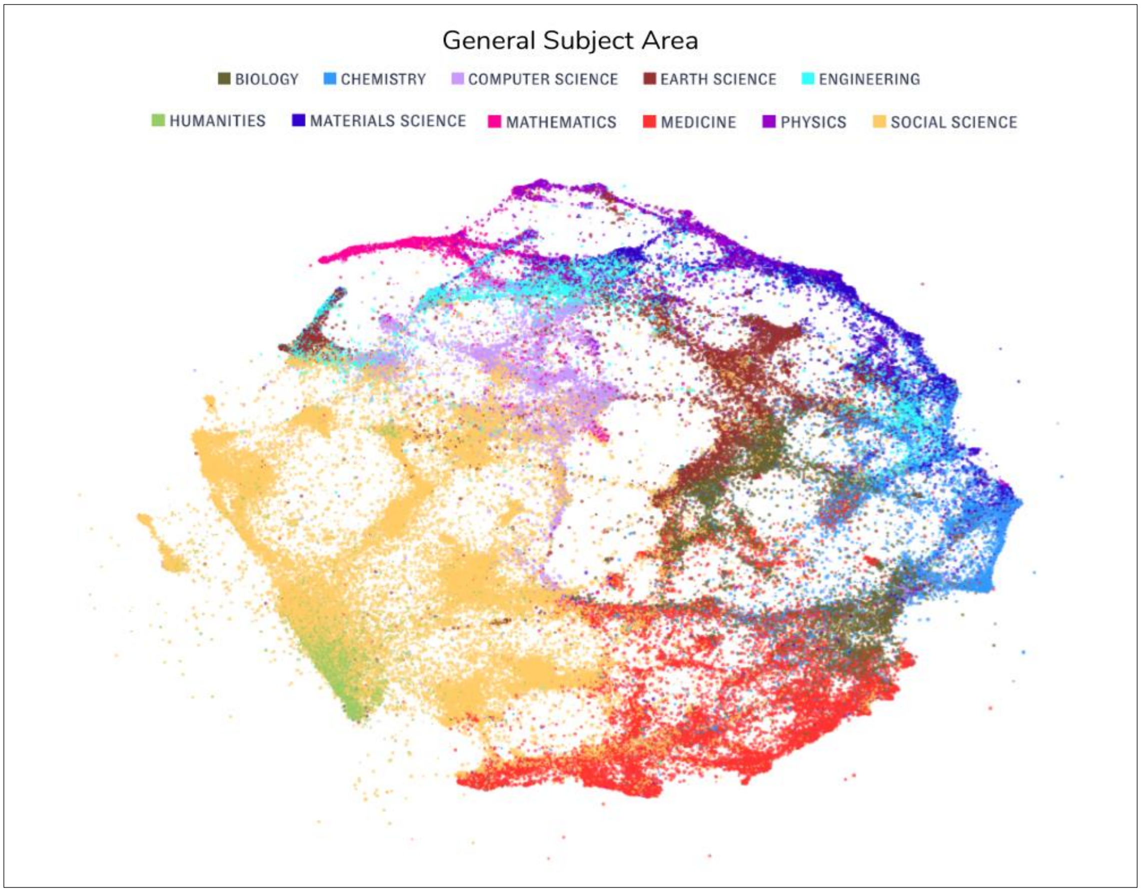
- The trustworthy AI terms *reliability*, *safety*, and *robustness* are widely dispersed across AI-research clusters, which further supports the idea from our previous research that these terms are widely studied and/or adopted metrics for AI systems.
- By contrast, we find the terms *interpretability, transparency, explainability, security,* and *privacy* are concentrated in fewer clusters, which may indicate that they are being studied as specific issues areas, rather than being broadly accepted and adopted characteristics across AI research.
- There are 18 clusters of AI-related research that are worth watching for those interested in the development of trustworthy AI. These 18 research clusters cover a broad spectrum of AI methods, techniques, and applications, including deep learning, adversarial attacks, word embeddings, image privacy, speech recognition, explainable AI, federated learning, algorithmic fairness, differential privacy, and robotics. For several of these clusters, the density of trustworthy AI keywords in the papers in the cluster is key to identifying the cluster as relevant to trustworthy AI research.

The Advantages of Research Clusters for Exploring Scholarly Literature

In our approach to finding trustworthy AI research, we used research clusters that appear in CSET's Emerging Technology Observatory's (ETO) Map of Science (displayed in Figure 1), which contains more than 120,000 research clusters that contain at least 50 total publications and five publications from the past five years.[2] CSET's research clusters represent groupings of scientific publications derived from citations. Specifically, a research publication must cite or be cited by at least one other publication to be assigned to a research cluster. This clustering approach is a reliable way of organizing millions of research publications around related research questions and approaches without being constrained by the limitations in topic modeling. Each dot in Figure 1 represents a research cluster and the colors reflect the general subject of the cluster.[3]

In terms of identifying related research papers, the clustering approach surpasses the performance of searching for keywords (one word can have multiple meanings and therefore pull unrelated papers) and the assignment of papers to topics (one paper may be appropriate to multiple topics and assignment can be arbitrary). Furthermore, CSET's research clusters contain aggregated metadata computed from member publications which can be used to analyze research at the cluster level, such as general subject area, percentage of AI-related publications, and key concepts.[4] Finally, a specific advantage for this analysis of trustworthy AI is that the citation cluster approach opens the aperture of analysis and surfaces publications that may not use a keyword but are nonetheless important or related to trustworthy AI.

Figure 1. ETO's Map of Science



Source: ETO Map of Science.

Finding Trustworthy AI Research Clusters

To select research clusters relevant to trustworthy AI out of the Map of Science, we began with a set of papers published between 2010-2021 that were classified by CSET as being AI-related and contained at least one of 13 trustworthy AI keywords drawn from the National Institute of Standards and Technology AI Risk Management Framework 1.0 (Box 1).[5]

Box 1. NIST AI RMF Terms Used for Keyword Search

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

* Officially, NIST uses the terms *bias-managed* and *privacy-enhanced*.

† *Robustness* is defined within NIST's discussion of *valid and reliable*.

This English-language-only keyword search resulted in a total of 322,209 publications for the terms in Box 1, or 14 percent of papers classified as AI-related in CSET's merged corpus. We refer to this set as trustworthy AI keyword publications.

Our previous research has established the limitations of a keyword search, especially the tendency for this approach to identify publications that use a trustworthy AI term but which are not actually relevant to trustworthy AI.[6] Despite this limitation, however, the admittedly flawed set of trustworthy AI keyword papers can be used to identify research clusters that do contain research relevant to trustworthy AI. To find these relevant clusters, we focused our attention on those that contained a high percentage of our trustworthy AI keyword papers as of 2023. Using prior analysis on the relevance of concentrations of AI-related papers to identify AI-related research clusters, we set a minimum threshold of 25 percent for trustworthy AI keyword publications. In other words, at least 25 percent of the papers in the cluster must include one of the keywords in the title or abstract.[7] This approach resulted in 18 research clusters of interest.[8]

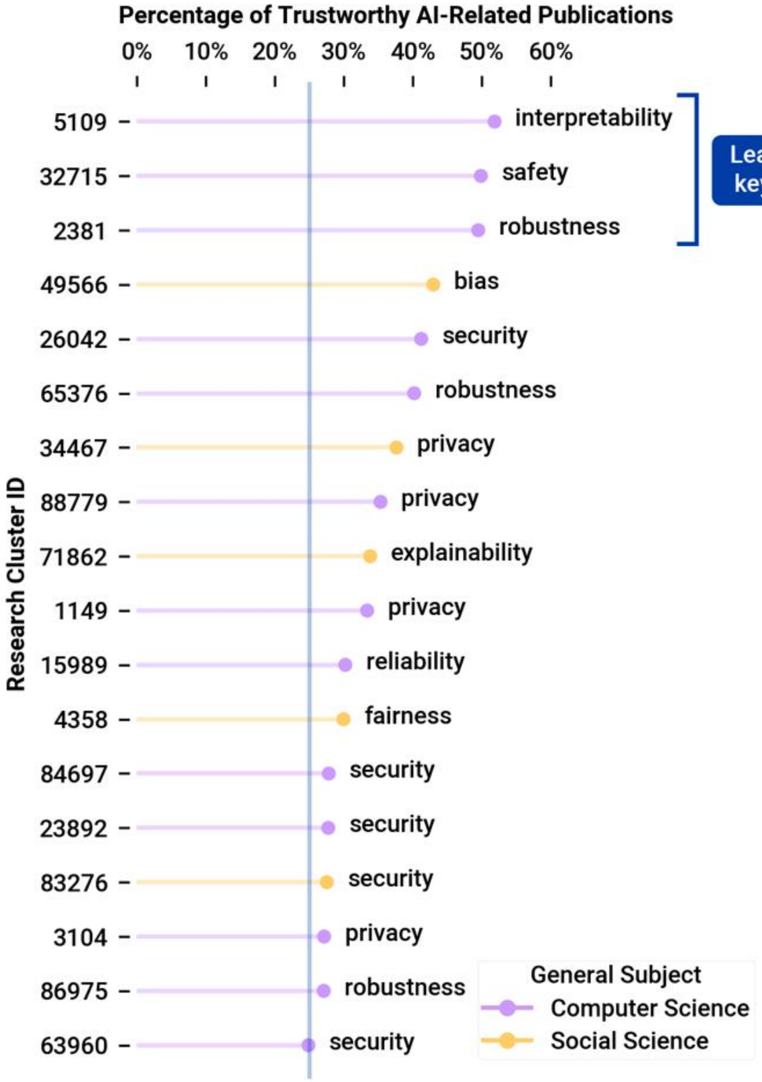
About the 18 Trustworthy AI Research Clusters

In Figure 2, we display the 18 trustworthy AI-related research clusters in descending order of their concentration of trustworthy AI keyword publications. The graph is annotated on the right with the trustworthy AI keyword that appears most frequently in the cluster (note that clusters contain publications that mention more than this one keyword). Among these 18 clusters, all have more than 25 percent trustworthy AI keyword publications and several have between 40 and 50 percent trustworthy AI keyword publications. For each cluster, we reviewed the titles of the “core” papers, (based on the article’s age, the total number of citations, and how often it cites articles within the cluster) and found that these core papers were all relevant to trustworthy AI issues, even in cases where the trustworthy AI term was not used in the article title.

Similar to the Map of Science, colors in Figure 2 represent the most common general subject category among articles in the cluster from the last five years.[9] We found that 72 percent of the clusters we identified fall under the computer science broad subject area and 28 percent of the clusters (five total) are labeled in the Map of Science as social science. For a complete listing of the clusters and their trustworthy AI term concentrations, as well as hyperlinks to more information, see Appendix B.

Finally, for each cluster, we identified the most frequently used trustworthy AI keyword and found that the terms *interpretability*, *safety*, *robustness*, *bias*, *security*, *privacy*, and *fairness* all featured as a lead term for one of our clusters. This is not to say that the other trustworthy AI terms were missing entirely, just that they were not the leading term in any of the 18 clusters.

Figure 2. Research Clusters with the Most Trustworthy AI Keyword Publications



Note: The trustworthy AI term that appears most frequently in each cluster is annotated on the right and does not represent the only trustworthy AI term present in the cluster.

Source: CSET research clusters.

Evidence of the Advantage of the Cluster Approach

In our previous research we found that AI papers with *safety* or *security* in the title or abstract in 2021 had a tendency to use those terms in a way unrelated to trustworthy AI policy concerns (for example, how to use AI to address non-AI safety issues like bike-helmet compliance or face mask detection). Given those prior observations we were concerned that of the 18 clusters with a high percentage of trustworthy AI terms, those with the lead term *safety* or *security* could in fact be unrelated to issues of trustworthy AI. However, after reviewing the most referenced papers in each of the clusters that had *safety* as a lead term, we found that the clusters were indeed relevant to trustworthy AI and they concerned deep learning safety and testing approaches. Of the five clusters with the lead word *security*, all were similarly relevant to trustworthy AI concerns because they included research on backdoor attacks on neural networks, facial recognition system attacks, and security for robotic systems. That the trustworthy AI-related clusters where *safety* and *security* were the leading terms were relevant to trustworthy AI concerns further demonstrates the value of a cluster-based analytical approach that leverages keywords to locate research of interest.

Trustworthy AI Cluster Development Over Time

Five clusters in particular consistently contained a higher percentage of trustworthy AI keywords over time (Table 1). Upon review, these clusters contained concepts important to trustworthy AI without necessarily using a trustworthy AI keyword, such as image data, machine learning models, testing, black-box adversarial attacks, and data poisoning.

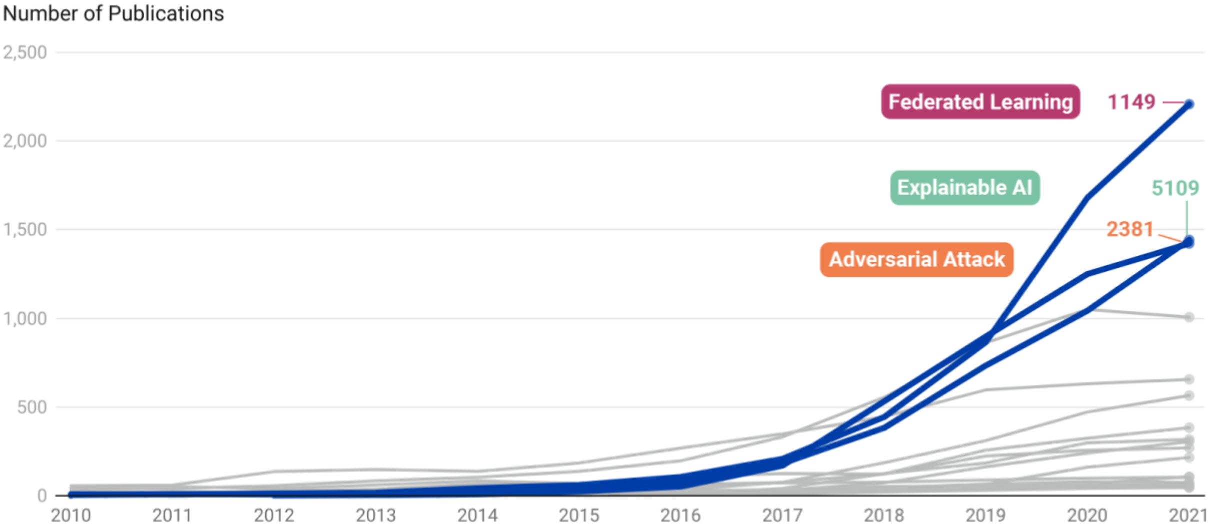
Table 1. Clusters with Consistently High Percentages of Trustworthy AI Keyword Publications Over Time

Research Cluster ID	Key Concepts	Top Trustworthy AI Keyword
34467	Image privacy protection, face images, generative adversarial network, privacy protection based, image data	Privacy
5109	Machine learning models, deep neural networks, deep learning, explainable machine learning, convolutional neural network	Interpretability
32715	Deep neural networks, neural network verification, deep learning, testing deep neural, deep learning applications	Safety
2381	Deep neural networks, adversarial perturbations, deep learning, black-box adversarial attacks, adversarial machine learning	Robustness
26042	Backdoor attacks, deep neural networks, data poisoning attacks, deep learning models, machine learning	Security

Source: CSET research clusters.

Furthermore, three clusters stood out for their rapid growth concurrent with the growth in the percentage of trustworthy AI keyword papers in the cluster. Cluster 2381 (related to adversarial attacks), cluster 1149 (related to federated learning), and cluster 5109 (related to explainable AI) grew significantly starting around 2017 (Figure 3). Additionally, all three showed a significant increase in the percentage of papers including a trustworthy AI keyword term over that same time period: from 25 to 57 percent in the case of adversarial attacks, from 24 percent to 45 percent in the case of federated learning, and from 42 percent to 63 percent in the case of explainable AI (for a full investigation of percentages, see Appendix C). The top trustworthy AI keyword for the federated learning cluster (privacy) echoed observations in our previous research about the dominance of federated learning as a key topic in highly cited AI papers using the word *privacy* in the title or abstract.[10]

Figure 3. Growth of Trustworthy AI Keyword Clusters Over Time, 2010-2021



Source: CSET research clusters.

Understanding Trustworthy AI Research by Keyword Dispersion and Lead Cluster Terms

After labeling each of our 18 clusters with the leading trustworthy AI term, we noticed that the three most frequently used terms in our paper titles and abstracts—*robustness*, *reliability*, and *safety*—did not appear as much as we would have expected. Only five of our 18 clusters featured any of these terms (three lead with *robustness*, one with *reliability,* and one with *safety*). This stood in notable contrast to the other keywords, like *interpretability*, *privacy*, *explainability*, and *fairness,* which appear less frequently in the papers but were nonetheless leading terms for one or more clusters. We hypothesized that the reason why some commonly used terms did not appear as a lead term for a cluster was that those terms were broadly dispersed across all AI clusters, whereas the other, lesser-used terms, were more concentrated in clusters as areas of research.

To test this hypothesis, we checked how dispersed each trustworthy AI keyword was across the 166,000 clusters by counting the number of clusters where you could find an AI paper that had each trustworthy keyword. Table 2 displays those counts, with the term on the left and the count of unique clusters that contained an AI paper using that keyword. This result supports the hypothesis from our earlier brief that certain terms (especially *reliability*, but also *safety* and *robustness*) are common to a wide swath of AI research, and not necessarily concentrated as a particular field or subject of research. The diffusion of these three keywords contrasts with other trustworthy AI terms like *interpretability*, *privacy*, *explainability*, and *fairness,* which appeared much less frequently in our set of keyword papers and appear in fewer clusters but were still a leading word for one or more of our 18 trustworthy AI clusters. This may be because these keywords represent more specific research areas and so represent higher concentrations in their member clusters.

Table 2. Number of Distinct Research Clusters Containing Trustworthy AI Papers, by Terms

Trustworthy AI Term	Number of Clusters
Reliability	15,103
Safety	9,120
Robustness	8,758
Bias	7,334
Security	6,882
Interpretability	4,070
Trust	3,823
Privacy	2,620
Transparency	1,867
Explainability	1,802
Resiliency	1,188
Fairness	958
Accountability	608

Source: CSET's research clusters.

Leveraging Cluster Metadata to Learn More about Clusters with Trustworthy AI Papers

As mentioned earlier, with research clusters we can use the aggregated metadata from the cluster's papers to better contextualize findings. Using the metadata, we found that the Map of Science key concepts allowed us to assign each cluster to one of 11 groups (which we call key concept groups). For a complete listing of clusters and our manually assigned short-title group based on Map of Science metadata, please see Appendix B.[11]

These key concept groups help us better understand and summarize the main subject of the clusters which have a high percentage of trustworthy AI papers (recall that we set the bar for "high" as 25 percent). The key concept groups we found included:

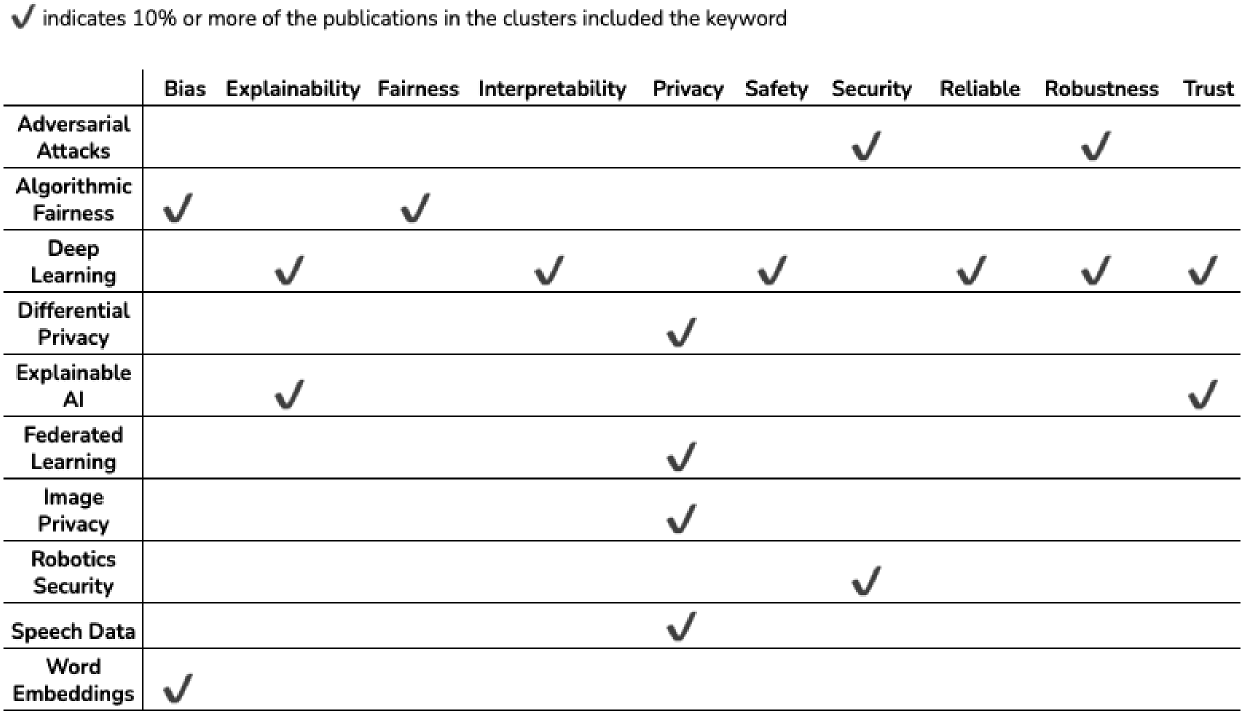
- Adversarial Attacks (6 unique clusters)
- Deep Learning (4)
- Word Embeddings (1)
- Image Privacy (1)
- Speech Data (1)
- Explainable AI (1)
- Federated Learning (1)
- Algorithmic Fairness (1)
- Robotics Security (1)
- Differential Privacy (1)

These groupings prompt several further observations. First, a number of the groups contain a trustworthy AI keyword but also give policymakers a deeper understanding of what is going on in the technical research using that keyword. For example, *privacy* is an important keyword, and the cluster groupings with the word *privacy* surface two particular aspects of *privacy* to research that are meant to enhance the privacy of AI systems: one on image privacy and one on a technique known as differential privacy. We can also see the phrase "Algorithmic Fairness," is the term of art, as opposed to AI fairness, for example. And that the term *security* is attached to the area of robotics (as opposed to, for example, cyber) is also interesting given the importance of cyber security to policymakers and hopes that AI may render cyber systems more—and not less—secure.

Second, there are several more general AI research areas surfaced in the key concept groups, such as deep learning, word embeddings, and speech data. Lastly, two of these groupings do not use a trustworthy AI keyword: adversarial attacks and federated learning.

To contextualize these key concept groups within our trustworthy AI keywords, we created a table showing the keywords that appeared in at least 10 percent of the publications in each of the key concept groups (Table 3 or Appendix D for a full listing with percentages).[12] From this table we better understand that adversarial attack clusters are concerned with the trustworthy AI keywords *security* and *robustness* and that federated learning has to do with *privacy*. We can also see that deep learning research clusters contain the largest variety of trustworthy AI keywords, but that there is a particular concern with the specific keyword, *privacy,* when it comes to the speech data cluster and the keyword, *bias,* when it comes to word embeddings.

Table 3. AI Research Cluster Key Concept Groups and Trustworthy AI Keyword Term Appearance



✓ indicates 10% or more of the publications in the clusters included the keyword

	Bias	Explainability	Fairness	Interpretability	Privacy	Safety	Security	Reliable	Robustness	Trust
Adversarial Attacks							✓		✓	
Algorithmic Fairness	✓		✓							
Deep Learning		✓		✓		✓		✓	✓	✓
Differential Privacy					✓					
Explainable AI		✓								✓
Federated Learning					✓					
Image Privacy					✓					
Robotics Security							✓			
Speech Data					✓					
Word Embeddings	✓									

Source: CSET research clusters.

The Exchange of Ideas Between Trustworthy AI Clusters

Finally, to analyze the citation-link relationships and how our clusters of interest may influence one another, we extracted the papers that were most often referenced by publications in any of the 18 trustworthy AI keyword clusters (Table 4). We consider these publications to be most influential to trustworthy AI-related research, as opposed to all of science, since we count citations from other trustworthy AI keyword publications specifically. For a visual representation of inter-cluster citations among trustworthy AI keyword publications, we provide a keyword cascade plot in Appendix E.[13]

Table 4. Most Referenced Publications within the Top 18 Trustworthy AI Keyword Clusters

Paper Title	Trustworthy AI Keyword Mentions	Research Cluster ID	Number of Cluster References
“Why Should I Trust You?”: Explaining the Predictions of Any Classifier	Interpretable, Trust	5109	4364
Towards Evaluating the Robustness of Neural Networks	Robustness	2381	4139
Intriguing Properties of Neural Networks	Interpretable	2381	3674
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization	Bias, Trust	5109	3070
Federated Machine Learning: Concept and Applications	Privacy, Security	1149	2852
Communication-Efficient Learning of Deep Networks from Decentralized Data	Privacy, Robust	1149	2797
A Unified Approach to Interpreting Model Predictions	Interpretability	5109	2761
DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks	Reliable, Robust	2381	2739
Towards Deep Learning Models Resistant to Adversarial Attacks.	Robustness, Security	2381	2669
Adversarial Examples in the Physical World	Security	2381	2288

Source: Authors’ analysis.

As could be expected, several of these highly-cited papers survey prior research of relevance to trustworthy AI or to a specific characteristic of trustworthy AI, or they have general applications of interest to a larger community of researchers as opposed to specific ones that would, in turn, be more likely to be cited in fewer papers.

Conclusion

AI-enabled systems are transforming society and driving an intense focus on what policy and technical communities can do to ensure that those systems are trustworthy and used responsibly. But to achieve the goal of trustworthy AI, policymakers and researchers must come together to establish the desired characteristics and adjudicate the proper technical approaches to establish or evaluate them. To that end, this analysis identifies 18 clusters of research within CSET’s Map of Science that are more relevant to policymakers and developers who are looking to understand trustworthy AI research. Furthermore, our analysis can help policymakers and developers better understand distinct areas of research concerned with trustworthy AI and investigate how the research clusters relate to each other.

These clusters are not the sum total of research relevant to trustworthy AI, but they are good places for interested parties to start exploring the topic further. AI is a still-developing area of scientific research, and the clusters we have identified now will add new papers, and new relevant clusters will emerge over time. All 18 clusters, as well as the entirety of the Map of Science, can be monitored at CSET’s Emerging Technology Observatory.

Authors

Autumn Toney is a data research analyst at CSET, where Emelia Probasco is a senior fellow.

Acknowledgments

For feedback and assistance, we would like to thank Catherine Aiken, Tessa Baker, Kevin Boyack, Shelton Fitch, Melissa Flagg, Richard Klavans, Margarita Konaev, Jason Ly, Igor Mikolic-Torreira, Dewey Murdick, Micah Musser, and Caleb Smith.

© 2023 by the Center for Security and Emerging Technology. This work is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License.

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

Document Identifier: doi: 10.51593/20230014b

Appendix A: Trustworthy AI Keyword Search Terms

We used a regular expression query that searched over publication titles and abstracts containing the following set of keywords. We drew our keywords from the National Institute of Standards and Technologies' AI Risk Management Framework (NIST AI RMF), which lists and defines a set of characteristics of trustworthy AI.[14] These included:

- Accountability, Accountable
- Bias
- Explainability, Explainable
- Fairness
- Interpretability, Interpretable
- Privacy
- Reliability, Reliable
- Robustness
- Safety
- Security
- Resilience
- Transparency
- Trust

For more information on why these terms were selected from NIST's list, please see Emelia Probasco and Autumn Toney, *The Inigo Montoya Problem for Trustworthy AI: The Use of Keywords in Policy and Research*, Center for Security and Emerging Technology, June 2023.

Appendix B: Trustworthy AI Research Cluster Numbers, Extracted Phrases, and Key Concept Groups

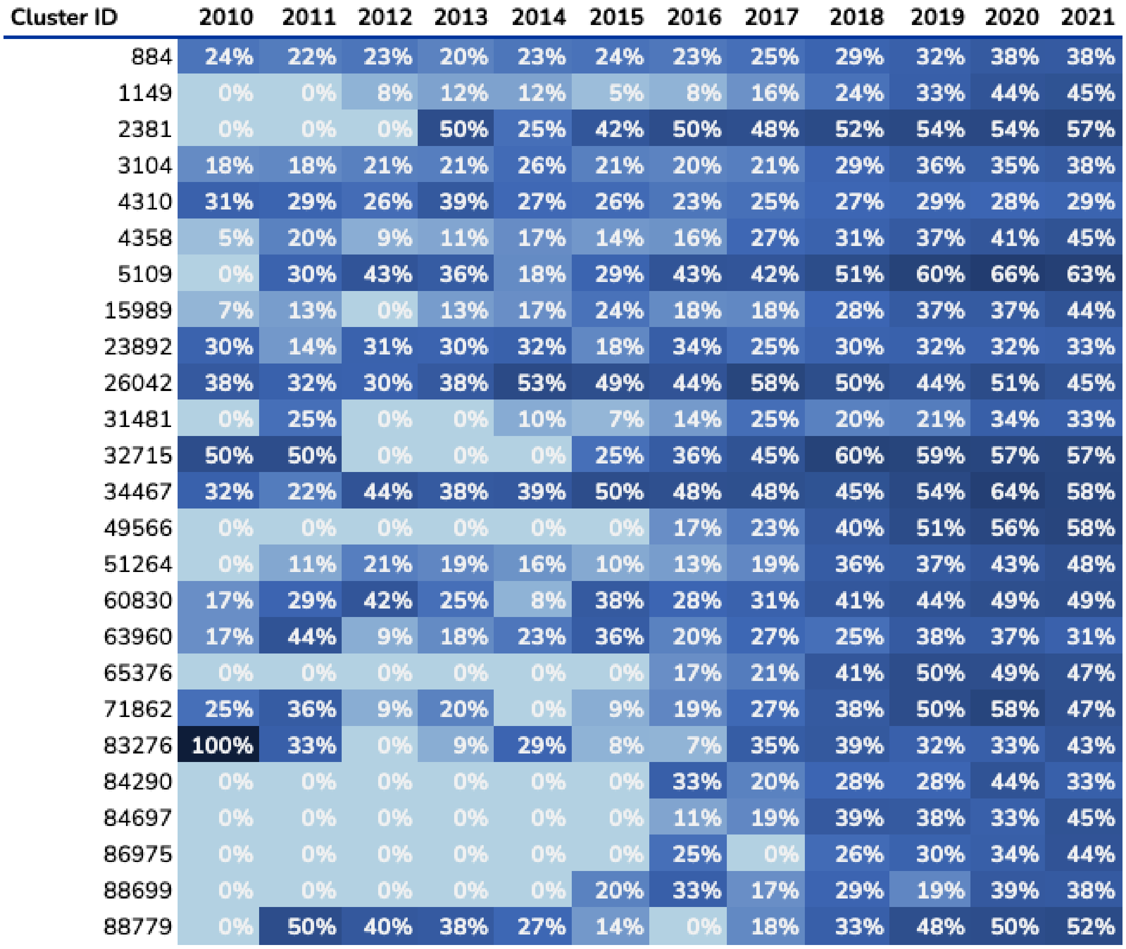
The table below lists each of the 18 clusters we identified as relevant to trustworthy AI, with the cluster number and hyperlink to that cluster in the Map of Science in the left-hand column. The "Trustworthy AI Keyword Percentage" is the percentage of papers in that cluster that contain any of our trustworthy AI keywords (listed in Appendix A). The "Key Concepts" are displayed in the Map of Science, and are found using the yake algorithm to identify the overall top twenty key phrases for a cluster based on their titles and abstracts.[15] We reviewed each of the clusters and their key concepts and used these to create key concept groups, which are listed in the right-hand column.

Cluster ID	Trustworthy AI Keyword Percentage	Key Concepts	Summary Group
5109	52%	Machine Learning models, deep neural networks, deep learning, explainable machine Learning, convolutional Neural Network	Deep Learning
32715	50%	Deep neural networks, neural network verification, deep learning, testing deep neural, deep learning applications	Deep Learning
2381	50%	deep neural networks, adversarial perturbations, Deep Learning, Black-box Adversarial Attacks, adversarial machine learning	Adversarial Attacks
49566	43%	Gender Bias, word embeddings, natural language processing, language models, mitigating gender bias	Word Embeddings
26042	41%	Backdoor attacks, deep neural networks, data poisoning attacks, deep learning models, machine learning	Adversarial Attacks
65376	40%	Adversarial attacks, NLP models, natural language processing, text classification models, generating adversarial texts	Adversarial Attacks
34467	38%	Image privacy protection, face Images, generative adversarial network, privacy protection based, image data	Image Privacy
88779	35%	Speech data, speech emotion recognition, automatic speech recognition, speech data publishing, privacy-preserving speech data	Speech Data

71862	34%	Explainable reinforcement learning, explainable artificial intelligence, reinforcement learning agents, eXplanation generation, human users	Explainable AI
1149	33%	Federated learning, machine learning models, distributed machine learning, training data, deep neural networks	Federated Learning
15989	30%	Bayesian deep learning, deep neural networks, Bayesian neural, deep learning models, uncertainty estimation	Deep Learning
4358	30%	Machine Learning, algorithmic fairness, fairness research algorithms, fairness constraints, data protection	Algorithmic Fairness
84697	28%	Face morphing attacks, morphed face images, morphing attack detection, face recognition systems, face image detection	Adversarial Attacks
23892	28%	Presentation attack detection, face presentation attack, face spoofing detection, face anti-spoofing, face recognition systems	Adversarial Attacks
83276	28%	Robot operating system, ROS system attacks, time location systems, robot security framework, industrial robots	Robotics Security
3104	27%	Differential privacy, local differential, private data, privacy protection, LDP	Differential Privacy
86975	27%	Deep neural networks, deep learning models, neural network watermarking, watermarking deep neural, DNN models	Deep Learning
63960	25%	Presentation attack detection, iris presentation attack, iris liveness detection, iris recognition, iris images	Adversarial Attacks

Source: CSET research clusters.

Appendix C: Percentage of Trustworthy AI Publications by Cluster, per Year, 2010-2021

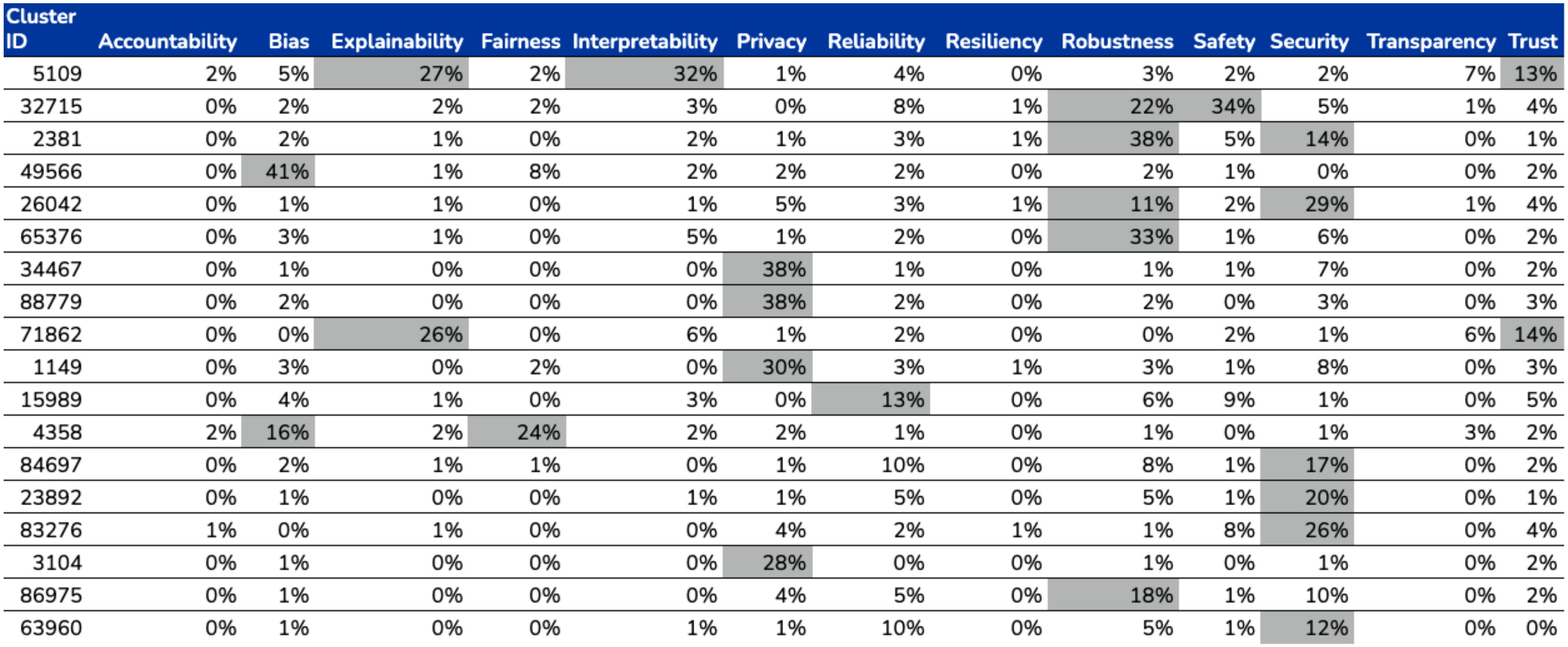


Cluster ID	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
884	24%	22%	23%	20%	23%	24%	23%	25%	29%	32%	38%	38%
1149	0%	0%	8%	12%	12%	5%	8%	16%	24%	33%	44%	45%
2381	0%	0%	0%	50%	25%	42%	50%	48%	52%	54%	54%	57%
3104	18%	18%	21%	21%	26%	21%	20%	21%	29%	36%	35%	38%
4310	31%	29%	26%	39%	27%	26%	23%	25%	27%	29%	28%	29%
4358	5%	20%	9%	11%	17%	14%	16%	27%	31%	37%	41%	45%
5109	0%	30%	43%	36%	18%	29%	43%	42%	51%	60%	66%	63%
15989	7%	13%	0%	13%	17%	24%	18%	18%	28%	37%	37%	44%
23892	30%	14%	31%	30%	32%	18%	34%	25%	30%	32%	32%	33%
26042	38%	32%	30%	38%	53%	49%	44%	58%	50%	44%	51%	45%
31481	0%	25%	0%	0%	10%	7%	14%	25%	20%	21%	34%	33%
32715	50%	50%	0%	0%	0%	25%	36%	45%	60%	59%	57%	57%
34467	32%	22%	44%	38%	39%	50%	48%	48%	45%	54%	64%	58%
49566	0%	0%	0%	0%	0%	0%	17%	23%	40%	51%	56%	58%
51264	0%	11%	21%	19%	16%	10%	13%	19%	36%	37%	43%	48%
60830	17%	29%	42%	25%	8%	38%	28%	31%	41%	44%	49%	49%
63960	17%	44%	9%	18%	23%	36%	20%	27%	25%	38%	37%	31%
65376	0%	0%	0%	0%	0%	0%	17%	21%	41%	50%	49%	47%
71862	25%	36%	9%	20%	0%	9%	19%	27%	38%	50%	58%	47%
83276	100%	33%	0%	9%	29%	8%	7%	35%	39%	32%	33%	43%
84290	0%	0%	0%	0%	0%	0%	33%	20%	28%	28%	44%	33%
84697	0%	0%	0%	0%	0%	0%	11%	19%	39%	38%	33%	45%
86975	0%	0%	0%	0%	0%	0%	25%	0%	26%	30%	34%	44%
88699	0%	0%	0%	0%	0%	20%	33%	17%	29%	19%	39%	38%
88779	0%	50%	40%	38%	27%	14%	0%	18%	33%	48%	50%	52%

Source: CSET research clusters.

Appendix D: Percentage of Trustworthy AI Keyword Publications within AI-related Research Clusters

Percentages of publications with trustworthy AI key term mentions in research clusters (percentages of 10 percent or higher are highlighted).



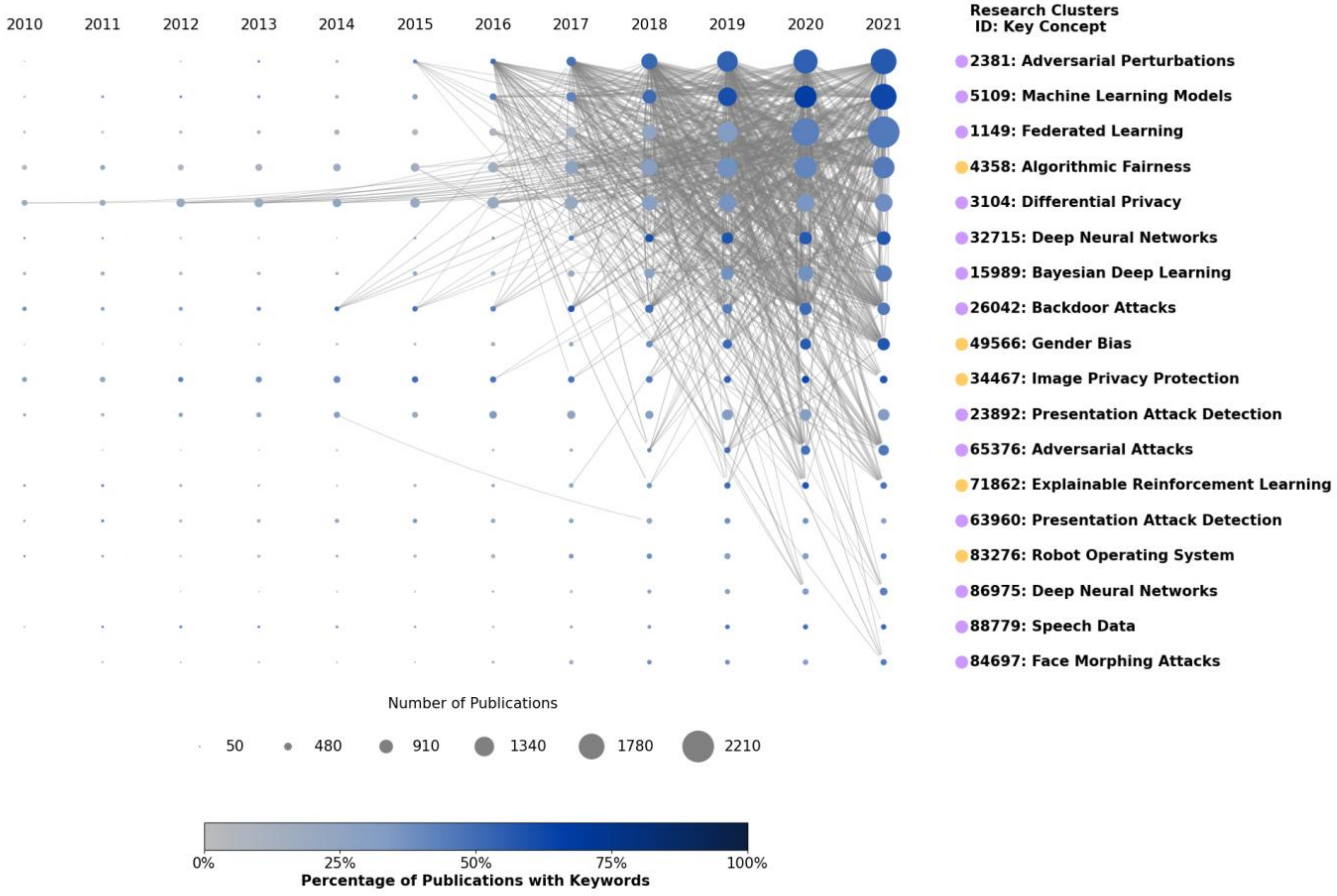
Cluster ID	Accountability	Bias	Explainability	Fairness	Interpretability	Privacy	Reliability	Resiliency	Robustness	Safety	Security	Transparency	Trust
5109	2%	5%	27%	2%	32%	1%	4%	0%	3%	2%	2%	7%	13%
32715	0%	2%	2%	2%	3%	0%	8%	1%	22%	34%	5%	1%	4%
2381	0%	2%	1%	0%	2%	1%	3%	1%	38%	5%	14%	0%	1%
49566	0%	41%	1%	8%	2%	2%	2%	0%	2%	1%	0%	0%	2%
26042	0%	1%	1%	0%	1%	5%	3%	1%	11%	2%	29%	1%	4%
65376	0%	3%	1%	0%	5%	1%	2%	0%	33%	1%	6%	0%	2%
34467	0%	1%	0%	0%	0%	38%	1%	0%	1%	1%	7%	0%	2%
88779	0%	2%	0%	0%	0%	38%	2%	0%	2%	0%	3%	0%	3%
71862	0%	0%	26%	0%	6%	1%	2%	0%	0%	2%	1%	6%	14%
1149	0%	3%	0%	2%	0%	30%	3%	1%	3%	1%	8%	0%	3%
15989	0%	4%	1%	0%	3%	0%	13%	0%	6%	9%	1%	0%	5%
4358	2%	16%	2%	24%	2%	2%	1%	0%	1%	0%	1%	3%	2%
84697	0%	2%	1%	1%	0%	1%	10%	0%	8%	1%	17%	0%	2%
23892	0%	1%	0%	0%	1%	1%	5%	0%	5%	1%	20%	0%	1%
83276	1%	0%	1%	0%	0%	4%	2%	1%	1%	8%	26%	0%	4%
3104	0%	1%	0%	0%	0%	28%	0%	0%	1%	0%	1%	0%	2%
86975	0%	1%	0%	0%	0%	4%	5%	0%	18%	1%	10%	0%	2%
63960	0%	1%	0%	0%	1%	1%	10%	0%	5%	1%	12%	0%	0%

Source: CSET research clusters.

Appendix E: Inter-cluster Citation Links of Trustworthy AI Keyword Publications

To analyze the citation-link relationships and how our clusters of interest may influence each other, we generated a keyword cascade plot for the 18 trustworthy AI keyword clusters. This plot displays each cluster (labeled on the right-hand side) and the citation links between the trustworthy AI keyword publications in those clusters over time. We only display connections using a gray line for clusters with more than four citations for the given year and we do not include intra-cluster citation links in our analysis.[16] For each year, the dots are colored by the percentage of trustworthy AI keyword publications (darker blue dots have a higher percentage of trustworthy AI papers) and the size corresponds to the total number of publications in the cluster.

Trustworthy AI Research Cluster Keyword Cascade Plot



Source: CSET research clusters.

This plot helps us contextualize publications within our trustworthy AI clusters by visualizing their development over time and their relationships with other trustworthy AI clusters. The density of the connections between the clusters as they develop helps us to further group the clusters as strongly connected, potentially as a consequence of the trustworthy AI theme.

Endnotes

[1] CSET's merged corpus of scholarly literature includes Digital Science's Dimensions, Clarivate's Web of Science, Microsoft Academic Graph, China National Knowledge Infrastructure, arXiv, and Papers With Code. 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, USA.

[2] Emerging Technology Observatory, "Documentation: Map of Science," accessed April 4, 2023, https://eto.tech/tool-docs/mos/.

[3] Research subject designation is explained in the Emerging Technology Observatory documentation: Emerging Technology Observatory, "Documentation: Merged Academic Corpus," accessed April 12, 2023, https://eto.tech/dataset-docs/mac/.

[4] Emerging Technology Observatory, "Documentation: Research Cluster Dataset," accessed April 4, 2023, https://eto.tech/dataset-docs/mac-clusters/#key-concepts-1.

[5] CSET's AI classifier was trained using the arXiv corpus of scientific preprints; 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. Also, 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; and Emelia Probasco and Autumn Toney, "The Inigo Montoya Problem for Trustworthy AI: The use of keywords in policy and research," Center for Security and Emerging Technology, June 2023.

[6] As one example, the word *security* is used in papers about techniques for securing AI and it is also used in papers about using AI for food security. We also know that a keyword approach is likely to miss research relevant to trustworthy AI that does not use one of the 13 terms in paper titles or abstracts. More information on the use of trustworthy AI keywords can be found in Emelia Probasco and Autumn Toney, "The Inigo Montoya Problem for Trustworthy AI: The Use of Keywords in Policy and Research," Center for Security and Emerging Technology, June 2023.

[7] Autumn Toney, "Locating AI Research in the Map of Science" Center for Security and Emerging Technology, July 14, 2021.

[8] We acknowledge that not every paper in a cluster may be considered to be about trustworthy AI explicitly. Rather, the papers in a cluster will be related in their focus on a common problem, technique, or approach. This contrasts research grouped by keyword, where papers may focus on substantially different problems or approaches (such as in the example of papers with the keyword security).

[9] Emerging Technology Observatory, "Documentation: Research Cluster Dataset," accessed April 4, 2023, https://eto.tech/dataset-docs/mac-clusters/#subjects.

[10] Emelia Probasco and Autumn Toney, “The Inigo Montoya Problem for Trustworthy AI: The use of keywords in policy and research,” Center for Security and Emerging Technology, June 2023.

[11] Emerging Technology Observatory, “Documentation: Research Cluster Dataset,” accessed April 4, 2023, https://eto.tech/dataset-docs/mac-clusters/#subjects.

[12] We chose the 10 percent threshold based on previous research indicating that a 10 percent threshold of papers within a cluster is often representative of cross-disciplinary research. For more on this topic, please see Autumn Toney, “Locating AI Research in the Map of Science,” July 14, 2021, https://cset.georgetown.edu/publication/locating-ai-research-in-the-map-of-science/.

[13] Autumn Toney and Melissa Flagg, “Keyword Cascade Plots,” Center for Security and Emerging Technology, February 1, 2023, https://cset.georgetown.edu/publication/keyword-cascade-plots/.

[14] National Institute of Standards and Technology, *AI Risk Management Framework: Second Draft*, August 18, 2022, https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf.

[15] Emerging Technology Observatory, “Documentation: Research Cluster Dataset,” accessed April 4, 2023, https://eto.tech/dataset-docs/mac-clusters/#key-concepts-1.

[16] Following the methodology from “Keyword Cascade Plots.”