April 2021

Mapping Research Agendas in U.S. Corporate AI Laboratories

CSET Data Brief

CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

AUTHORS
Rebecca Gelles
Tim Hwang
Simon Rodriguez

Executive Summary

Within the broad research field of artificial intelligence (AI), it is worth understanding, specifically, what leading U.S. companies invest in. This data brief conducts an analysis of the research papers published by Amazon, Apple, Facebook, Google, IBM, and Microsoft over the past decade to better understand what work their labs are prioritizing, and the degree to which these companies have similar or different research agendas overall.

We find the following:

- **Major "AI companies" are often focused on very different subfields within AI.** While companies like Amazon, Apple, Facebook, Google, IBM, and Microsoft are often grouped together generically as leaders in AI, an analysis of their publications shows considerable differentiation in the areas of research they prioritize. While publications may not provide the full picture of these companies' research agendas, as companies may not choose to publish on work that will form the basis of valuable intellectual property, it still provides a window into the differences in research agendas between these companies. Policymakers should be careful to consider these differences in framing national assessments of technological competitiveness and in strategizing government investments in research.

- **The private sector may be failing to make research investments consistent with ensuring long-term national competitiveness.** None of the leading companies examined in this analysis appear to be prioritizing work on problem areas within machine learning that will offset the broader structural challenges the United States faces in deploying and benefitting from the technology when competing against authoritarian regimes.[1] This includes work in areas such as few-shot learning, federated learning, simulation learning, interpretability, and ML fairness.

Introduction

Private companies are investing heavily in advancing the cutting edge of AI and the subfield of ML. As a result, national policy around AI must take into account the state of play of corporate investment in the technology. To the extent that national interests and private sector agendas converge, the U.S. government may only need to encourage existing research activity. To the extent that these interests diverge, U.S. government strategy may need to intervene more extensively in order to ensure national competitiveness in underinvested areas.

Effectively positioning the U.S. government as a “gap filler” in basic research requires a nuanced understanding of the various subfields within ML and the level of research effort and investment they are receiving from various organizations, both public and private. ML is not a monolithic technology. The field of ML is better characterized as a broad family of related techniques, with many subfields focused on specific types of applications and technical challenges. By necessity, research organizations prioritize their limited resources within this universe of potential research opportunities, focusing on making progress on certain problems while leaving others by the wayside.

This brief analysis sheds light on the research priorities of six major U.S. technology companies in the field of ML through an analysis of their research publications over the past decade. While companies do not choose to publish on every subject they conduct research on, saving some of their work as proprietary intellectual property, an analysis of the publications data from companies should still provide a broad overview of the topics companies considered important enough to invest time, money, and effort into pursuing. Through topic modeling, our analysis breaks down ML into a series of subfields, and identifies the areas that have been the greatest focus for these companies over time. While the major U.S. companies identified with the modern breakthroughs in ML do share some overlap in research agendas, publication activity reveals some significant differences in their prioritization of problem areas within the field. This suggests important underlying divergences in the level of investment, effort, and interest of these companies in certain areas of ML.

Policymakers should be aware of these differing areas of focus, as well as the areas not receiving as much attention. This analysis makes the case for a framing of AI competitiveness that is more multifaceted in nature, and suggests a potential mismatch that exists between private research investments and national priorities.

Methodology

To analyze organizational research agendas, we relied on Digital Science's Dimensions dataset of scientific literature. Dimensions is a platform maintained by Digital Science that tracks over 128 million scholarly publications, grants, policies, data, and metrics across academic fields.[2] Using a method described in earlier work by CSET, we extract a set of 1,269,033 papers related to the topic of AI, along with academic citations to these papers from within the field.[3]

While this dataset does not span the entire universe of scientific literature available, it is quite broad, and carries the unique advantage of a much higher-quality form of author affiliate entity resolution than most comparable scientific literature datasets. This means that we are much better able to capture which authors within the dataset who are affiliated with an organization actually have published particular papers within the dataset, and produce a complete picture of the literature within.[4]

From the initial set of papers drawn from Dimensions, we narrowed our focus to enable us to examine the research agendas of major U.S. corporate labs focusing on ML. We focused on six companies: Amazon, Apple, Facebook, Google, IBM, and Microsoft.[5] This specific selection of companies is based on their representation on the board of the Partnership on AI, a nonprofit coalition committed to the responsible use of AI founded in 2016, that we take as a signal of public prominence in the industry and in the research field.[6] To provide a baseline to the research activity of these companies, we appended papers published by the top 100 universities as measured by publication volume in the field of ML since 2010. Publication counts for each company are provided for reference in Appendix A, and a list of these universities and their publication counts is provided in Appendix B.

To provide a structured analysis of the topics covered by the papers in our resulting dataset of 270,802 papers, we then applied topic modeling, a method that categorizes papers into 60 high-level areas.[7] Our use of topic modeling was predicated on our goal of identifying organizational research agendas based on the granular technical problems being addressed in their publications.[8] Topic modeling is well-established[9] as an approach to classifying documents by their relevance to various themes.[10]

Results

Topic modeling provides insight into the topic areas that various organizations have published in most frequently over the past decade. The top five topic areas for each of the six leading U.S. AI companies tracked in our analysis is provided below, along with our qualitative interpretation of the topic areas and counts for the number of papers classified under each topic. We also provide the top five topics by publication count across our aggregate dataset of papers from universities and companies. For reference, the specific terms associated with each topic and paper counts for each topic cluster are available on Github.[11]

Table 1: Top five topics by organization, as measured by publication volume.

Apple (72)	Amazon (138)
Topic 12 - robotics and grasping **(9)** **Topic 18** - model performance evaluation **(4)** **Topic 16** - image segmentation **(4)** **Topic 2** - model architecture **(3)** **Topic 8** - optimization **(2)**	**Topic 12** - robotics and grasping **(10)** **Topic 46** - fault / failure diagnosis **(6)** **Topic 41** - graph based learning **(6)** **Topic 8** - optimization **(5)** **Topic 5** - image (denoising / resolution) **(5)**

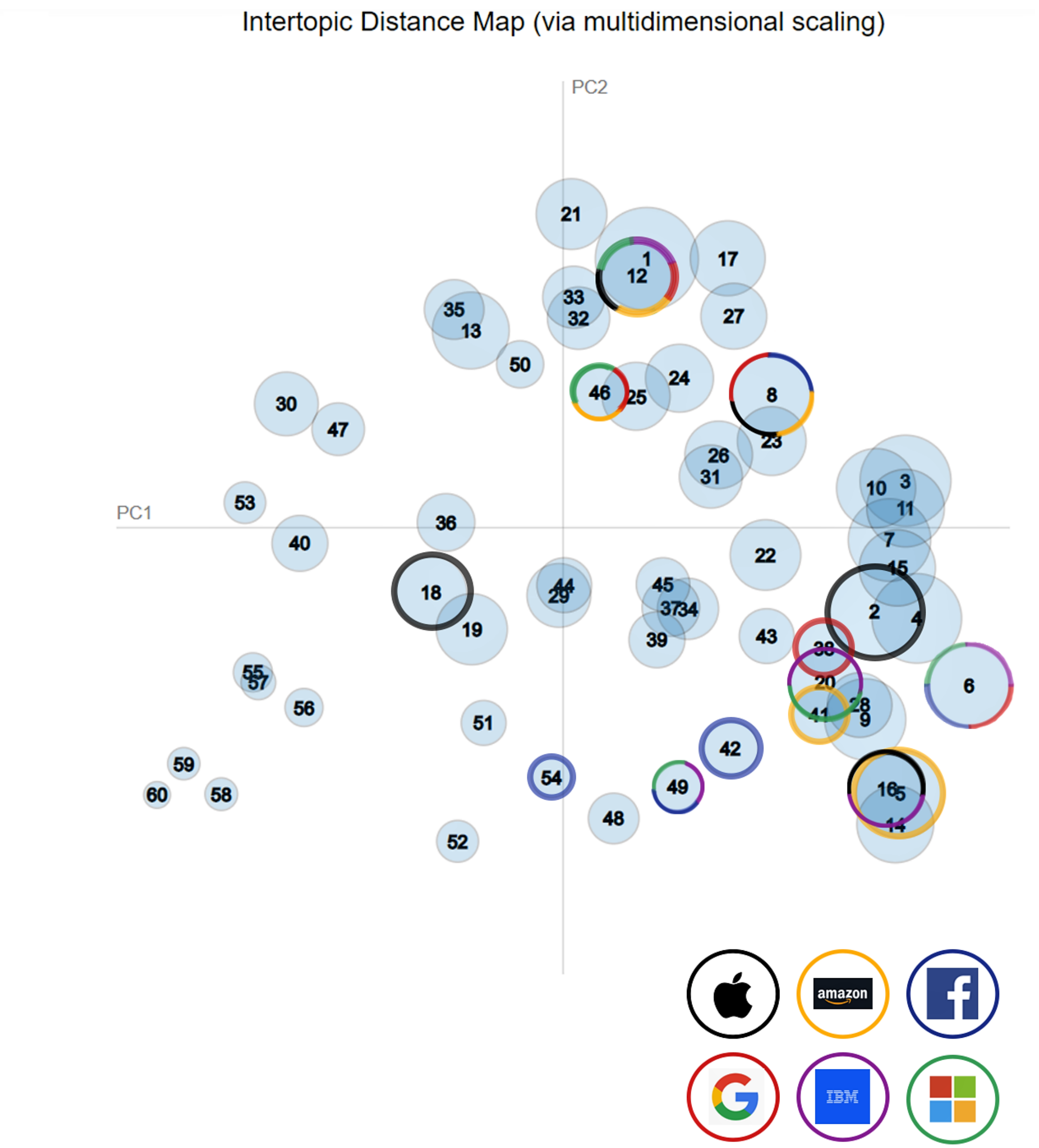
Facebook (173)	Google (1,033)
Topic 49 - images (color / light enhancement) **(12)** **Topic 54** - GANs and image generation **(6)** **Topic 42** - images (stereoscopy / 3D) **(6)** **Topic 8** - optimization **(5)** **Topic 6** - sparse matrices and representation **(5)**	**Topic 12** - robotics and grasping **(52)** **Topic 46** - fault / failure diagnosis **(45)** **Topic 38** - object tracking **(33)** **Topic 6** - sparse matrices and representation **(32)** **Topic 8** - optimization **(31)**
IBM (2,659)	Microsoft (3,346)
Topic 12 - robotics and grasping **(130)** **Topic 20** - modeling of 3D shapes **(105)** **Topic 49** - images (color / light enhancement) **(90)** **Topic 16** - image segmentation **(87)** **Topic 6** - sparse matrices and representation **(86)**	**Topic 12** - robotics and grasping **(175)** **Topic 20** - modeling of 3D shapes **(134)** **Topic 46** - fault / failure diagnosis **(122)** **Topic 6** - sparse matrices and representation **(119)** **Topic 49** - images (color / light enhancement) **(109)**
Aggregate (270,802)	
Topic 12 - robotics and grasping **(12,834)** **Topic 20** - modeling of 3D shapes **(10,853)** **Topic 49** - images (color / light enhancement) **(9,500)** **Topic 6** - sparse matrices and representation **(8,674)** **Topic 46** - fault / failure diagnosis **(8,633)**	

Source: Dimensions.

In order to gain a better sense of the relationship between the topics generated by our modeling, and the distribution of private company publication activity among the topics, we created an intertopic distance map (Figure 1).[12]

In this visualization, the size of the circles denote the relative prevalence of the topics in the dataset. Topics that appear closer to one another share terminology, suggesting that they may be more closely related to one another semantically. We then color coded the top topics for each of the companies tracked in our dataset in order to identify areas of overlap and divergence between them.

Figure 1: Intertopic distance map with leading company topics marked.



Source: Dimensions.

This visualization reveals a number of interesting aspects of corporate publishing activity in ML.

First, it is clear that the six companies tracked in our analysis do share considerable overlap in their research activity. Topic 8 (optimization), Topic 6 (sparse matrices and representation), and Topic 12 (robotics and grasping) all appear in listings of the most published topics in our dataset for four or more of the companies. On some level, the frequent appearance of these topics is not surprising. Techniques in optimization and sparse matrices, for instance, are fundamental to the performance of ML algorithms, so widespread publishing on the topic across major corporate labs is to be expected. Similarly, the potential business opportunities offered by expanding the applications of ML to real-world settings is consistent with the appearance of robotics and grasping as a top research focus across these labs.[13]

However, beyond this tightly linked cluster of topics, the research agenda of the companies diverge. There are numerous topics that appear to be clear priorities for a handful of companies but not the others. Facebook, for instance, appears to have made research into image generation and transformation a priority, focusing on Topic 54 (GANs and image generation) and Topic 42 (stereoscopic imagery) in a way distinct from the other companies in our analysis. Similarly, Apple and IBM appear to have focused efforts on image segmentation (Topic 16) in a way distinct from the other four companies in our analysis.

Perhaps the most important aspect of the research landscape revealed by Figure 1 is that publishing from corporate labs covers only a subset of the full range of active research areas within ML. Significant areas of research such as Topic 35 (gait analysis) and Topic 27 (robotic navigation), for example, do not appear in the lists of top five topics that are most frequently published by the six companies tracked in our analysis. This may mean these labs are not working on these research areas, or it may mean they believe the potential for profit in these areas is high enough such that they do not yet want to share what they have discovered with the wider research community. Either is notable; identifying which is the case would require a more in-depth analysis of the patent literature to evaluate whether any of these companies are producing intellectual property in areas they are not publishing, which could be a productive avenue for future work.

Conclusion

CSET has published previous work about the structural issues that may limit the ability for democracies to quickly adopt and benefit from advances in ML.[14] Commitments to privacy, for instance, may make it more challenging to acquire the data necessary to train high-performance ML systems, and to deploy these technologies in certain contexts. While one path is to compromise on these values to move faster, investments in certain technical areas with ML can allow democracies to benefit from the technology without these sacrifices. For instance, advances in "one-shot" or "few-shot" learning may enable the creation of ML systems that achieve high performance with significantly smaller training datasets.

This "terrain strategy" seeks to shape the field of ML to mitigate the limitations that democracies face under the current state of play in the technology, and to upset some of the advantages that authoritarian regimes may enjoy in the status quo. This previous work identifies a range of technical areas—from simulation and federated learning to ML interpretability and fairness—that might overcome some of the structural hurdles that democracies face in effectively developing and deploying ML. Interestingly, none of the high-impact research areas identified in this previous work are represented among the top areas of publication by the six leading companies examined in this analysis.

This suggests that there may be a place for the government and policymakers to play a role as a "gap filler" in offsetting the structural challenges that the United States may face in ML. The major private labs that have invested aggressively in ML in recent years may not be investing in the specific areas that are most beneficial to the overall U.S. position in the technology. Policy can work to influence the research agenda of the leading labs, as well as rally the wide range of other universities, companies, and funding agencies to direct their efforts on these topics. Prioritizing and advancing these subdomains of ML may be a critical part of ensuring U.S. competitiveness in the technology going forwards.

This analysis challenges simplistic notions of “leadership” in the field of ML. The six companies tracked in this analysis are often grouped together as leaders in the field. However, a closer look reveals that, while commonalities exist in their research agendas, they are far from being aligned in their priorities. No one company has established a dominant publishing record across the totality of topics that exist within ML.

This picture of the field suggests a notion of “leadership” in the technology that is multivalent. It is unhelpful to ask which company hosts the “top” research lab in ML, the answer to this question depends critically on identifying a specific problem area of interest. Facebook may credibly claim to lead in ML-generated imagery while Google leads in optimization. None of these companies may be leaders in applications of ML to problems like gait analysis—even if they produce proprietary work on these topics, by not sharing their work they are removing themselves from the leadership process of setting or moving forward the research agendas on these problem areas.

This point has broader implications beyond the six companies reviewed in this analysis. In the context of national competition in AI and ML, policymakers should reconsider whether “leadership” in the technology—in an absolute, categorical sense—is a practicable objective. As with the companies reviewed in this data brief, the issue should be explored with more nuance: what are the subfields of ML that the United States should be prioritizing in order to best advance the national interest? Seeking effective prioritization, rather than leadership broadly, is a more productive framing of the issue.

Appendix A

Major U.S. technology companies and counts of AI publications in Dimensions as of May 29, 2020.

Institution	AI Publications
Microsoft	3,346
IBM	2,659
Google	1,033
Facebook	173
Amazon	138
Apple	72

Source: Dimensions.

Appendix B

Top colleges and universities and counts of AI publications in Dimensions as of May 29, 2020.

College/University	Country	AI Publications
Tsinghua University	China	9,399
Shanghai Jiao Tong University	China	7,841
Harbin Institute of Technology	China	7,436
Beihang University	China	6,918
Zhejiang University	China	6,916
University of the Chinese Academy of Sciences	China	6,168
Nanyang Technological University	China	6,080
Wuhan University	China	5,351
Carnegie Mellon University	U.S.	5,055
Peking University	China	5,003
Beijing Institute of Technology	China	4,819
National University of Singapore	Singapore	4,665

Huazhong University of Science and Technology	China	4,662
Xidian University	China	4,655
University of Tokyo	Japan	4,532
University of Electronic Science and Technology of China	China	4,407
Xi'an Jiaotong University	China	4,406
National University of Defense Technology	China	4,381
Northeastern University	China	4,367
University of Science and Technology of China	China	4,286
Northwestern Polytechnical University	China	4,274
Chinese Academy of Sciences	China	4,204
Tianjin University	China	4,152
Anna University, Chennai	India	4,143
Institute of Automation, Chinese Academy of Sciences	China	4,085
Massachusetts Institute of Technology	U.S.	3,907
Beijing University of Posts and Telecommunications	China	3,700
Technical University of Munich	Germany	3,691
South China University of Technology	China	3,630
University College London	UK	3,625
Stanford University	U.S.	3,520
Southeast University	China	3,483
Dalian University of Technology	China	3,328
Imperial College London	UK	3,258
Georgia Institute of Technology	U.S.	3,223
Nanjing University	China	3,220
Sun Yat-sen University	China	3,196
University of Southern California	U.S.	3,160

Shandong University	China	3,159
Hong Kong Polytechnic University	China	3,087
Korea Advanced Institute of Science and Technology	South Korea	2,955
Nanjing University of Science and Technology	China	2,947
Chinese University of Hong Kong	China	2,945
Tongji University	China	2,937
University of Technology Sydney	Australia	2,929
University of Toronto	Canada	2,904
Seoul National University	South Korea	2,879
University of Illinois at Urbana Champaign	U.S.	2,869
National Taiwan University	Taiwan	2,867
Johns Hopkins University	U.S.	2,829
University of Michigan	U.S.	2,773
Shanghai University	China	2,765
ETH Zurich	Switzerland	2,761
Beijing Jiaotong University	China	2,754
University of São Paulo	Brazil	2,726
French National Centre for Scientific Research	France	2,706
University of California, Berkeley	U.S.	2,679
University of Oxford	UK	2,661
UNSW Sydney	Australia	2,577
Jilin University	China	2,503
KU Leuven	Belgium	2,493
University of Waterloo	Canada	2,475
Delft University of Technology	Netherlands	2,454
Xiamen University	China	2,447
Shenzhen University	China	2,440
University of Sydney	Australia	2,416

Sichuan University	China	2,403
Ministry of Education of the People's Republic of China	China	2,381
University of Pennsylvania	U.S.	2,376
Beijing University of Technology	China	2,372
Harvard University	U.S.	2,370
Chongqing University	China	2,370
City University of Hong Kong	China	2,342
University of California, Los Angeles	U.S.	2,326
Swiss Federal Institute of Technology in Lausanne	Switzerland	2,320
University of Cambridge	UK	2,288
University of Lisbon	Portugal	2,272
Nanjing University of Aeronautics and Astronautics	China	2,271
University of Tehran	Iran	2,270
Osaka University	Japan	2,237
Central South University	China	2,233
Harbin Engineering University	China	2,232
University of British Columbia	Canada	2,173
Arizona State University	U.S.	2,158
University of Alberta	Canada	2,150
Fudan University	China	2,111
Waseda University	Japan	2,102
University of California, San Diego	U.S.	2,074
Wuhan University of Technology	China	2,057
University of Granada	Spain	2,012
University of Washington	U.S.	2,005
University of Maryland, College Park	U.S.	1,987
Columbia University	U.S.	1,976
Yonsei University	South Korea	1,955

Purdue University West Lafayette	U.S.	1,955
Hefei University of Technology	China	1,948
University of Edinburgh	UK	1,918
Korea University	South Korea	1,900
Kyoto University	Japan	1,900
Rutgers, The State University of New Jersey	U.S.	1,885

Source: Dimensions.

Authors

Rebecca Gelles is a data scientist at CSET, where Tim Hwang is a research fellow, and Simon Rodriguez is a research assistant.

Acknowledgments

For feedback and assistance, we would like to thank Avonelle Davis, Shelton Fitch, Matt Mahoney, Igor Mikolic-Torreira, Nils Newman, and Dewey Murdick.

Document Identifier: doi: 10.51593/20200037

Endnotes

[1] Tim Hwang, “Shaping the Terrain of AI Competition” (Center for Security and Emerging Technology, June 2020), https://cset.georgetown.edu/research/shaping-the-terrain-of-ai-competition/.

[2] Dewey Murdick, James Dunham, and Jennifer Melot, “AI Definitions Affect Policymaking” (Center for Security and Emerging Technology, June 2020), https://cset.georgetown.edu/research/ai-definitions-affect-policymaking/.

[3] Murdick, Dunham, and Melot, “AI Definitions Affect Policymaking.”

[4] Dimensions relies on the GRID (https://grid.ac/) resolution system for companies and universities. Through CSET’s analysis, we have found that organizations that are resolved to GRID through this system are reliably captured in our datasets in a consistent way, with alternate names (like acronyms), typos, parents, children, mergers and acquisitions, generally well-handled. Dimensions is imperfect; in particular, not every organization is linked to a GRID. However, organizations without GRIDs tend to be smaller and have lower publication counts, so for an analysis of top organizations this is unlikely to affect our analysis. We also observe that Dimensions may undercount the total number of papers attributed to an organization. For instance, the dataset used in this brief counts 1,033 papers attributed to Google, while the company itself lists some 3,895 papers across the topics of machine intelligence, machine perception, machine translation, natural language processing, robotics, and speech processing on its site as of March 29, 2021. To the extent that an organization appears to not be publishing in a given topic, this may be an artifact of its exclusion from the Dimensions database. “Publications Database,” Google, accessed March 29, 2021, https://research.google/pubs/.

[5] To resolve these companies in Dimensions, we used not only their parent GRID identifier but the whole series of interconnected organizations resolved for each company. For example, for IBM, this included GRIDs for the following organizations: IBM (United States), IBM (United Kingdom), IBM (Germany), IBM (France), IBM (Netherlands), IBM (Canada), IBM (Brazil), IBM (Italy), IBM (India), IBM (Czechia), IBM (Egypt), IBM (Ireland), IBM Research – Austin, IBM Research – Almaden, IBM Research – Thomas J. Watson Research Center, IBM Research – Zurich, IBM Research – Australia, IBM Research – Brazil, IBM Research – Tokyo, IBM Research – India, IBM Research – China, IBM Research – Africa, IBM Research – Haifa, and IBM Research – Ireland. Each individual GRID resolves numerous names, as well: for example, just a few of the raw publication affiliations for IBM (United States) include the following: "IBM Corporation, Endicott, NY 13760,” “IBM, Berkeley, USA,” “IBM Microelectronics Div. (USA),” “International Business Machines Research Center, Ossining, New York,” and “Storage Techol. Div., IBM Corp., San Jose, CA, USA.”

[6] "Meet the Team," Partnership on AI, accessed August 12, 2020, https://www.partnershiponai.org/team/.

[7] This analysis applies a Latent Dirichlet Allocation (LDA) topic model based on Python's Gensim library to the abstracts of these papers, see "Topic Modeling for Humans," Gensim, accessed October 29, 2020, https://radimrehurek.com/gensim/.

[8] For an alternative technique using citation-based clusters, see Richard Klavans, Kevin Boyack, Dewey Murdick, "A Novel Approach to Predicting Exceptional Growth in Research," *arXiv [cs.DL]* (April 27, 2020), arXiv, https://arxiv.org/abs/2004.13159.

[9] Daniel Maier et al., "Applying LDA topic modeling in communication research: Toward a valid and reliable methodology," *Communication Methods and Measures* Issues 12, no. 2-3 (2018): 93-118, https://boris.unibe.ch/112835/7/Maier et al_2018_Applying LDA topic modeling in communication research.pdf.

[10] Our LDA model treated the combined title and abstract of each paper as a document, and extracted topics from these documents. To tune the model, we employed the standard metrics for model validation in the literature. This validation step involved tuning topic coherence, adjusting the model iteratively in order to optimize the c_v score and ensure that our topics were interpretable to our analysts. We also evaluated our model's perplexity, aiming to minimize it in order to best predict the true topics, although our focus was on maximizing coherence. Our final c_v coherence score was 0.55, and our final log-perplexity was -7.23. At the end of this tuning process, our topic model ultimately classified the papers in our dataset into a set of sixty, human-interpretable topics.

[11] "Topic Terms," Github, https://github.com/georgetown-cset/unicorn-topics/blob/master/data/results/topic_terms.txt; "Cluster Counts," Github, https://github.com/georgetown-cset/unicorn-topics/blob/master/data/results/cluster_counts.csv.

[12] The intertopic distance map was generated using pyLDAvis, with proximity between topics defined using a Jensen-Shannon divergence. To produce a two-dimensional map, pyLDAvis leverages the commonly used tool of Principal Components Analysis (PAC). For more information on this approach, see Carson Sievert and Kenneth E. Shirley, "LDAvis: A method for visualizing and interpreting topics", *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (2014): 63-70, https://www.aclweb.org/anthology/W14-3110/.

[13] Tom Simonite, "Alphabet's Dream of an 'Everyday Robot' is Just Out of Reach", *WIRED*, November 21, 2019, https://www.wired.com/story/alphabets-dream-everyday-robot-out-reach/.

[14] Tim Hwang, "Shaping the Terrain of AI Competition."