nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-021-00339-6

Research community dynamics behind popular AI benchmarks

Fernando Martínez-Plumed[1,2] ✉, Pablo Barredo[3] ✉, Seán Ó hÉigeartaigh[4,5] ✉ and José Hernández-Orallo[2,4] ✉

The widespread use of experimental benchmarks in AI research has created competition and collaboration dynamics that are still poorly understood. Here we provide an innovative methodology to explore these dynamics and analyse the way different entrants in these challenges, from academia to tech giants, behave and react depending on their own or others' achievements. We perform an analysis of 25 popular benchmarks in AI from Papers With Code, with around 2,000 result entries overall, connected with their underlying research papers. We identify links between researchers and institutions (that is, communities) beyond the standard co-authorship relations, and we explore a series of hypotheses about their behaviour as well as some aggregated results in terms of activity, performance jumps and efficiency. We characterize the dynamics of research communities at different levels of abstraction, including organization, affiliation, trajectories, results and activity. We find that hybrid, multi-institution and persevering communities are more likely to improve state-of-the-art performance, which becomes a watershed for many community members. Although the results cannot be extrapolated beyond our selection of popular machine learning benchmarks, the methodology can be extended to other areas of artificial intelligence or robotics, and combined with bibliometric studies.

Experimental benchmarks are at the heart of the phenomenal progress that artificial intelligence (AI) has witnessed in recent years. In areas such as machine learning, the relevance of a scientific contribution is often linked to the level of performance achieved for a popular dataset or competition. Relatedly, technical contributions in AI go beyond single scientific papers in peer-reviewed journals or conferences to a more complex ecosystem of teams and community projects developing architectures or systems with evolving reports (usually on arXiv.org and other open repositories), source code, pre-trained models and results (usually on github.com). This activity is frequently driven by benchmarks. The importance of benchmarks in influencing AI research is poorly captured by traditional scientometric research, which mostly focuses on published papers and citations between them.

In this Article, we analyse how benchmarks may affect research dynamics in AI and the way different players—from academia to tech giants—behave. We perform an analysis of 25 popular benchmarks in AI, with 1,943 result entries overall. We extract the co-author communities from bibliographic repositories and plot the evolution of their performance results over time. For each benchmark, 'success' is related to their contribution to the SOTA fronts, a state-of-the-art curve defined by performance jumps on a bidimensional plot with time and performance as dimensions. We explore a series of hypotheses about the behaviour of communities that make repeated attempts on the benchmarks versus those making more isolated attempts, the composition of successful communities (single institution versus multiple institutions), their diversity (industry, academia or mixed) and the temporal dynamics in terms of the number of active members per community. Recent studies[1,2] have suggested that 'small teams disrupt whereas large teams develop', but this finding may be interpreted very differently in the context of AI benchmarks for a number of reasons: collaboration for AI challenges goes beyond a single paper, papers increasing the SOTA performance may be aligned to disruption or development, and finally, computer science is one of the few disciplines that does not follow the general disrupt–develop spectrum (Fig. 1a in ref. [2]). This creates a novel scenario where these phenomena can be investigated, but which also requires a new methodology, starting from the analysis of communities (clusters of authors usually working together) rather than teams based on the static co-authorship of a single paper or the volatile notion of affiliation. Similarly, benchmarks give us a new perspective about the concentration of efforts in a few major geographic areas, the increasing relevance of China and whether major academic institutions are giving way to industry. The findings can then be compared with recent bibliographical studies, such as figure 4 in ref. [3]. Finally, we also analyse results in an aggregated way, using indicators such as the level of activity, the number of jumps and, more importantly, the efficiency of each institution (measured as jumps on the SOTA front versus activity).

Our findings have to be considered carefully given the sample of benchmarks we explore in this paper (data-driven machine learning techniques where big industrial players may have an advantage). Nevertheless, this scientometric meta-review of AI benchmarks introduces a series of methodological innovations, raises new questions and makes a series of recommendations. For instance, data below the SOTA front capture an important and useful part of the activity. This suggests that a more fruitful use and sustainable appraisal of these benchmarks in the future should be based on consideration of other metrics beyond performance, such as technical innovations and resource use[4]. This should be encouraged by competitions and scientific venues in the future.

[1]European Commission, Joint Research Centre, Seville, Spain. [2]Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politécnica de Valéncia, Valencia, Spain. [3]Universidad de Oviedo, Oviedo, Spain. [4]Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, UK. [5]Centre for the Study of Existential Risk, University of Cambridge, Cambridge, UK. ✉e-mail: Fernando.MARTINEZ-PLUMED@ec.europa.eu; UO237136@uniovi.es; so348@cam.ac.uk; jorallo@upv.es

Table 1 | List of AI benchmarks used in the analysis by task

Task	Benchmark	Entries	Authors	Communities	Average size
Action recognition	HMDB-51[42]	31	103	23	4.96 ± 2.44
Action recognition	UCF101[43]	42	132	27	5.89 ± 3.57
Atari games	*Montezuma's Revenge*[44]	40	111	11	19.18 ± 23.2
Atari games	*Space Invaders*[44]	45	119	15	15.62 ± 25.36
Image classification	CIFAR-100[17]	110	327	69	6.91 ± 8.71
Image classification	ImageNet[8]	270	489	61	20.33 ± 41.12
Image generation	CIFAR-10[17]	239	605	104	9.38 ± 15.23
Image super-resolution	Set5[45]	64	194	36	8.18 ± 9.19
Language modelling	enwik8[46]	33	87	18	7.44 ± 5.66
Language modelling	Penn Treebank[47]	38	95	19	7.58 ± 5.98
Link prediction	WN18RR[48]	39	122	25	6.84 ± 4.56
Machine translation	WMT2014 English–French[49]	40	119	16	11.94 ± 12.42
Machine translation	WMT2014 English–German[49]	57	167	24	11.92 ± 17.29
Named entity recognition	CoNLL 2003[50]	52	178	32	8.19 ± 7.58
Named entity recognition	Ontonotes v5[51]	20	58	16	4.44 ± 2.76
Object detection	COCO Minival[52]	111	185	24	28.62 ± 46.18
Object detection	COCO test-dev[52]	198	395	41	24.16 ± 38.99
Pose estimation	MPII Human Pose[53]	32	107	21	6.57 ± 4.8
Question answering	SQuAD1.1[9]	196	157	21	15.95 ± 20.35
Question answering	WikiQA[54]	18	55	13	5.31 ± 4.13
Semantic segmentation	Cityscapes test[55]	91	259	39	11.23 ± 13.31
Semantic segmentation	PASCAL VOC 2012 test[56]	53	146	23	10.95 ± 11.23
Sentiment analysis	IMDb[57]	36	117	24	5.84 ± 6.12
Sentiment analysis	SST-2 Binary classification[58]	53	201	36	7.61 ± 8.99
Speech recognition	LibriSpeech[59]	35	139	13	17.69 ± 15.97

Number of entries, unique authors, communities and average size of the latter (that is, number of members) for each benchmark.

The following section describes our selection of benchmarks and explains the choice to perform the study at the level of communities. Next, we provide a detailed analysis of the results and explore a number of hypotheses and indicators. We close with a discussion that puts the paper into context, states the limitations of our study and makes some general recommendations. The Methods describes the new methodological pipeline, including data sources, algorithms, indicators and visualizations.

Benchmarks and communities

Benchmarks in AI come in many different forms and are arranged very differently depending on the area. They may be linked to a regular competition or a one-off challenge, released by an organization or maintained by a collective, and may be evaluated by their performance (for example, error rate for a machine learning classifier) or by efficiency (for example, computation time for a SAT solver). Because of this diversity, there is no fully comprehensive and unbiased source of results for all benchmarks in AI. However, Papers With Code (paperswithcode.com) has become the most comprehensive source so far, with hundreds of benchmarks and thousands of results from associated papers, with an emphasis on machine learning. Understanding this data source and its possible limitations, as well as the criteria we have used to make the selection of benchmarks, is key for the interpretation of our results.

Table 1 shows the list of benchmarks. The Methods discusses in detail the criteria we used for this selection, the inherent bias of this source and its implications.

Jointly with the use of benchmarks, the other cornerstone for our study is the use of communities. A community is formed of one or more authors that usually work together; authors can be from the same or different institutions. This is a novel feature of this methodology: unlike standard bibliometric analysis, where papers are used as units, communities can be used to track the continued efforts of a group of collaborating researchers, exploring longer-term endeavours and repeated attempts at a benchmark. Communities are of increasing relevance as the collaborations between different AI groups in different sectors (from universities and companies) have widened notably over the past few years. This is seen, for instance, in the quantitative analysis of industry–academia collaborations in different countries worldwide included in the 2019 AI Index (Fig. 1.5a in ref. [5]). We create communities using the Clauset–Newman–Moore hierarchical agglomeration algorithm[6], from the adjacency matrices between authors. All the details and an example of communities created from the graph of adjacencies is shown in Supplementary Fig. 1. Communities represent a better entity to identify trends spanning several organizations, sectors, areas or countries, than independent or small research teams with the same affiliation, or an occasional co-authorship[7]. Another important reason to consider communities is to avoid the overlap and repetition that would happen if we considered single authors, resulting in double counting of multi-author papers.

More information and links to alternative baselines and methods, and the relevance of benchmarks for scientometrics, can be found in the Supplementary Information.

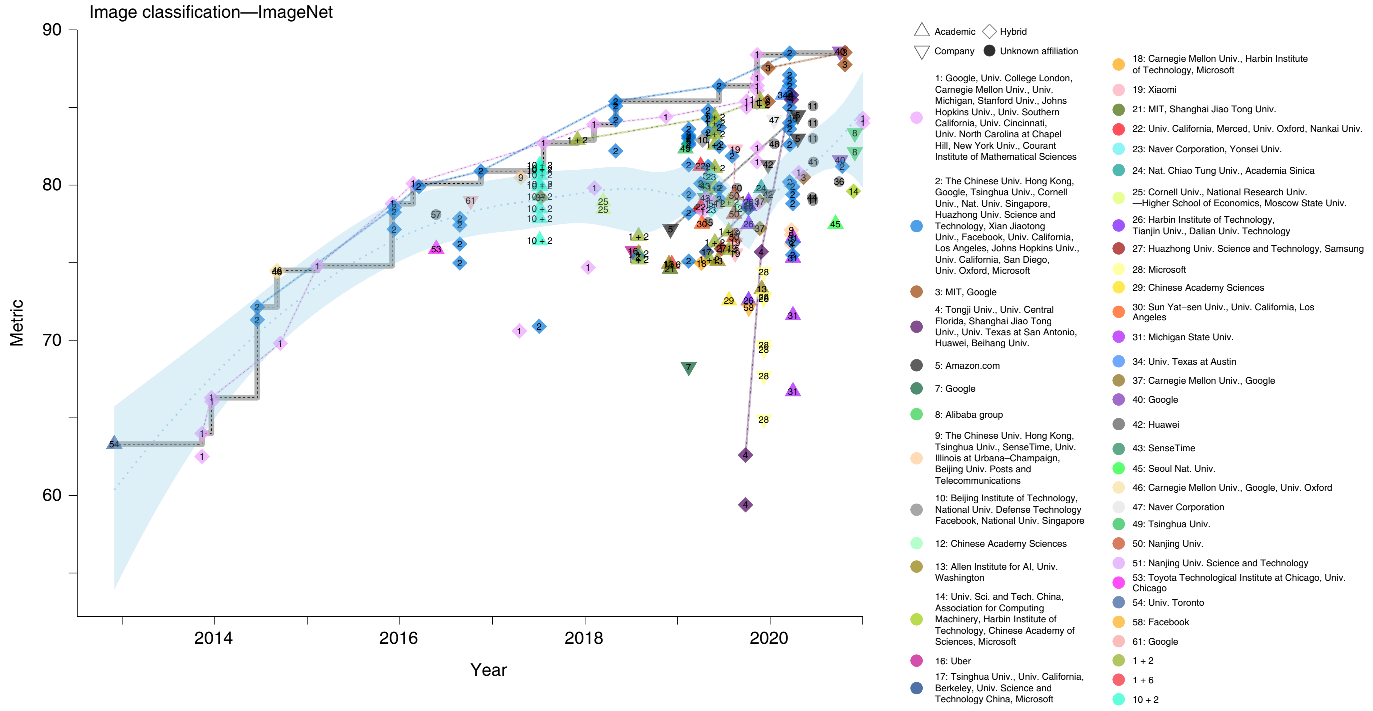

Fig. 1 | Progress in accuracy over time for ImageNet. Coloured shapes show the different communities (with one or more institutions in the legend). Dashed lines show the global SOTA front (in grey) for all the entries (results) and local SOTA front per community (in colour). The blue dotted line shows the smoothed means (all results) with 95% confidence level intervals. Different shapes indicate the types of institution (companies, universities or hybrid).

Findings

Once communities are identified for each benchmark, we can analyse specific trends, activity patterns or player dominance. For instance, Fig. 1 represents the results for the ImageNet dataset[8], which consists of 1.2 million images in 1,000 classes. The results of the different communities show that several long-term collaborative 'hybrid' groups, formed mostly by American universities (Johns Hopkins, University of California, Los Angeles, Cornell, Stanford, Toronto and so on) in collaboration with tech giants (Microsoft and Google) are those that have dominated the SOTA front from early on (communities numbered as 1 and 2). Although hybrid communities dominate the SOTA front, there are also some isolated company players, possibly representing different divisions, departments and research groups from companies such as Google, Xiaomi, Facebook and Microsoft. However, only a single non-hybrid community, Google, is able to achieve a score on the SOTA front.

The high number of points from 2016 to the present day is also notable. This is consistent with the general increased activity in AI and machine learning in the past five years. Another more intrinsic reason for the high activity may be that the benchmark is still not considered 'solved'. Human-level performance is reported to be around 95% of accuracy, whereas the SOTA front has gone from less than 65% in 2014 to 88% at the time of writing. No clear insights can be extracted from those entries that are far below the SOTA in the bottom-right part of the plot. While some have unknown affiliations, others (after detailed analysis of the papers) seem to represent improvements, modifications and adjustments over well-known algorithms that do not seem to have been evaluated by performance only but also by some other dimensions of what make a solution good[4].

While all benchmark plots can be found in the Supplementary Information (Supplementary Figs. 4–6), we include another example here, in Fig. 2. This is the Stanford Question Answering Dataset (SQuAD1.1)[9], a reading comprehension benchmark with more than 100,000 question–answer pairs from more than 500 articles. Questions derive from Wikipedia articles where the answer may be a segment of text from the corresponding reading passage, or may be unanswerable (for example, written adversarially to look similar to answerable ones). Like ImageNet, the SOTA front is dominated by hybrid long-term collaboration groups (communities numbered 2 and 3) formed by American universities (Stanford, Carnegie Mellon, Washington and so on) in collaboration with tech giants (Facebook and Google), but also by large hybrid communities formed by Asian universities (Beihang, Fudan, Peking and so on) jointly with Microsoft (community 1). We also observe that the participation of European universities initiatives is very low. Unlike ImageNet, most entries correspond to the period between 2016 and 2018, with a clear decline in activity from 2018 to 2020. This is probably due to the introduction of the new (and much more difficult) version of the benchmark (SQuAD2.0), with attention moving to the new challenge. However, SQuAD1.1 is still being addressed by communities 2 and 3, which have participated since 2016 and have led the SOTA from 2018 to 2020. Again, we see that long-term collaborative groups obtain better results than isolated communities.

Global hypotheses. We now provide a more general view by analysing the results of all 25 AI benchmarks together, to test a number of hypotheses. The hypotheses address questions about the research community dynamics, the key players behind the SOTA jumps and their characteristics. We postulate eight hypotheses (a detailed rubric can be found in the Supplementary Information):

1. $H_{Collaboration}$: the majority of points in the SOTA front belong to multi-institution communities.
2. $H_{Persistence}$: the ratio of points in the SOTA front belonging to multi-attempt communities is higher than the ratio below the SOTA.
3. H_{Hybrid}: the ratio of points in the SOTA front belonging to hybrid communities is higher than the ratio below the SOTA.

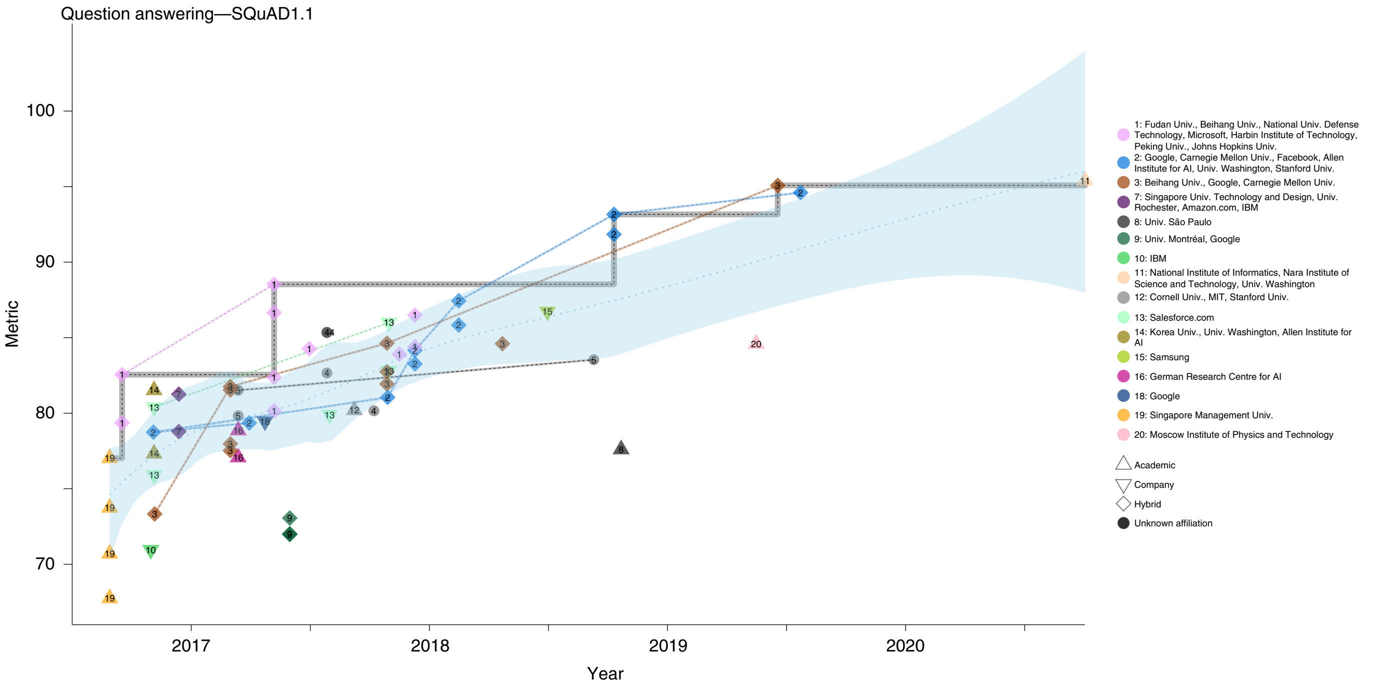

Fig. 2 | Progress in accuracy over time for SQuAD1.1. Shapes and lines are as in Fig. 1.

4. H_{Company}: points on the SOTA front are more likely to include companies than points below the SOTA.
5. H_{Increase}: jumps with a SOTA increase higher than the total increase of the SOTA in the previous year are associated with higher-than-average activity growth in the next year.
6. $H_{\text{Consecutive}}$: consecutive jumps (on the SOTA front) by the same community are followed by a lower-than-average activity growth in the following year.
7. H_{CommRise}: communities have more active members before a SOTA success than in other previous years.
8. H_{CommWane}: communities become relatively smaller after a SOTA success.

We test the above set of hypotheses to see whether they are met for each benchmark. Some hypotheses are expressed as implications; if we do not find the antecedent for a benchmark (for example, no jumps by the same community), then we consider this a non-case (a blank), and it is not used for hypothesis testing. Results are shown in Table 2. The *P* value is calculated for each hypothesis (two-sided binomial test at 95% confidence interval), but as we are considering eight hypotheses, we apply the Bonferroni correction for multiple comparisons[10].

From the above results, we reach a number of conclusions about the dynamics of communities engaging with AI benchmarks. We find that (1) SOTA jumps are mainly obtained by multi-institution communities, compared with the number of jumps obtained by single-institution communities; (2) multi-attempt communities are more likely to achieve SOTA jumps (compared with one-shot efforts); (3) jumps are mainly obtained by hybrid communities involving both universities and companies, meaning that heterogeneous communities achieve more success through collaborative efforts compared with 'pure' communities (only universities or companies); and, finally, (4) the presence of companies in a community, such as Google, Microsoft and Facebook, increases the odds of achieving a jump in an AI benchmark. All the above reinforces the usefulness of the increasing tendency of collaboration between universities and industry in AI research. Note, however, that association should not be interpreted as causation. It may well be that more success could increase the reputation for some institutions and attract further collaboration between different institutions. Also, reiterated attempts over a benchmark could provide more opportunities to work with other institutions.

It seems that the industry is engaging with academia not only to examine how best to produce and commercialize AI technologies but also to perform fundamental and applied research. Furthermore, over the past few years, we have witnessed an increasing number of high-profile university AI scholars being involved in research and leadership roles at tech giant companies where they have set up powerful research teams with both talented researchers and engineers, with a thin line between academia and industry[11].

The results for the fifth and sixth hypotheses are inconclusive. We cannot infer that new jumps are associated with an increase of activity (nor a decrease) immediately afterwards (5). Similarly, if the same community repeatedly achieves SOTA jumps, this does not seem to discourage the rest (6). These two hypotheses suggest that the motivations for working on a challenge go beyond what a single community or several communities are doing at the SOTA front.

The two last hypotheses bring more insight about the temporal dynamics behind community activity. We see communities usually grow before a SOTA jump (7), and their relative size diminishes after it (8). This seems to indicate that a SOTA jump is a watershed moment for the community members. Supplementary Fig. 2 shows consistent results too. The number of active researchers per activity is higher for SOTA communities (with mean 6.27) than for all communities (SOTA and non-SOTA, with mean 5.31). According to the disrupt–develop hypothesis (Fig. 1A in ref. [2]), larger communities are more incremental than disruptive. This seems to bring into question the idea that SOTA jumps are disruptive, and be consistent with the view that the real disruptive ideas are happening below the SOTA front. Nevertheless, we would require a deeper analysis of what contributions are considered disruptive in the discipline to answer this.

Global indicators. We have seen what kind of communities may be more successful and how they affect—or do not affect—the behaviour of other communities. Now we focus on characterizing their

Table 2 | Our eight hypotheses checked for each AI benchmark

Benchmark	$H_{Collaboration}$	$H_{Persistence}$	H_{Hybrid}	$H_{Company}$	$H_{Increase}$	$H_{Consecutive}$	$H_{CommRise}$	$H_{CommWane}$
HMDB-51	✓	✓	✗	✗	–		✓	✓
UCF101	✓	✓	✓	✓	✗		✓	–
Atari 2600 *Montezuma's Revenge*	✓				✗	✗	✓	–
Atari 2600 *Space Invaders*	✓		✓		✗	✓	✓	–
CIFAR-100	✗	✓	✓	✓	–	✗	✓	✓
ImageNet	✓	✓	✓	✓	–	✓	✓	✓
CIFAR-10	✓	✓	✗	✗			–	
Set5—4× upscaling	✓	✓	✓	✓	✗	✓		✓
enwik8	✓				–			
Penn Treebank (Word Level)	✓				✓			
WN18RR	✓	✓	✓	✓			–	
WMT2014 English–French	✓	✓	✓	✓	–	–	✓	✓
WMT2014 English–German	✓	✓	✓	✓	✗	✓	✓	✓
CoNLL 2003 (English)	✓		✓		–	–	✓	✓
Ontonotes v5 (English)	✓				✓			
COCO Minival	✓		✓			✓	✗	✓
COCO test-dev	✓	✓	✓	✓	✗	✓	–	✓
MPII Human Pose	✓	✗	✗	✗	–		✓	
SQuAD1.1	✓		✓			✓	✓	✗
WikiQA	✓	✓		–		✗	–	
Cityscapes test	✓	✓	✓	✓		✓	✗	✓
PASCAL VOC 2012 test	✓	✓	✓	✓	✗	✓	✓	✓
IMDb	✓	✓	✓	✓	✓		–	✓
SST-2 Binary classification	✓	✗	✓	✓	–			✓
LibriSpeech test-clean	✗	✓		✓	✓		✓	✓
Probability	0.920	0.882	0.842	0.794	0.421	0.714	0.775	0.861
P value	**0.000**b	**0.002**b	**0.004**b	**0.013**	0.648	0.180	**0.012**	**0.001**b

A tick indicates that the hypothesis is satisfied, the cross indicates that the hypothesis is violated, and '–' indicates that the hypothesis is neither satisfied or violated. We use a blank when the hypothesis is not applicable. Probabilities per each hypothesis are computed as the ratio between the number of benchmarks satisfying the hypothesis and the total number of benchmarks where the hypothesis is applicable (with '–' counted as half). Statistical significance (*P* values) is computed using the two-sided binomial test at 95% confidence level (with '–' counted as half), in bold when significant, and with 'b' when so with the Bonferroni correction (testing each individual hypothesis at 1 – 0.05/8 = 99.375% confidence level).

institutions. To do this, we use three main indicators: the level of activity (in terms of the number of papers or results published), the number of jumps (in terms of results placed on the SOTA) and the efficiency, the latter being the ratio between SOTA jumps and activity per institution.

Figure 3 shows the top institutions comparing activity and number of jumps obtained, as well as their efficiency. SOTA jumps on the *y* axis are dominated by a small group of players, with tech giants Google, Microsoft and Facebook, and a few universities, Carnegie Mellon, Oxford and Hong Kong, representing most jumps. Ignoring the institutions with zero SOTA jumps shown at the bottom, we see an important correlation between activity and SOTA jumps, but some universities such as Northwestern, New York and Oxford have the highest efficiency.

The observation that a few institutions dominate the activity and jumps suggests that benchmark activity is concentrated around a small number of key players. To better understand this, we can put this in the context of market scenarios when several players compete for a share. We can measure the concentration of institutions using the Herfindahl–Hirschman index (HHI)[12]. This measures the sum of the squares of the activity shares (s_i) for each institution i from a set of n institutions ($\mathrm{HHI} = \sum_{i=1}^{n} s_i^2$). The result can range from 0 to 1: $\mathrm{HHI} \leq 0.01$ indicates a well distributed scenario (highly competitive); $0.01 < \mathrm{HHI} \leq 0.15$ indicates an unconcentrated scenario, $0.15 < \mathrm{HHI} \leq 0.25$ indicates moderate concentration; and $\mathrm{HHI} > 0.25$ indicates high concentration (note that values close to 1.0 imply a single monopoly). In the case of the global set of benchmarks, computing the HHI of activity and SOTA jump share per institution results in scores of 0.043 and 0.07, respectively. This indicates a non-concentrated (highly competitive) research scenario in terms of the activity performed by the different institutions and the SOTA jumps achieved. Note that we do not compute the HHI for communities due to the fact that communities are very different for different benchmarks, implying little or no match between them (even if there are common institutions). A low number of common communities attempting several benchmarks is natural, since some benchmarks require specialized researchers that are not needed for others. Calculating the HHI for communities across benchmarks would simply give a low value of HHI.

We can also analyse the data geographically. Supplementary Fig. 3 shows a scatter plot of countries comparing activity and SOTA jumps. Efficiency, measured as jumps on the SOTA versus activity, distributes among actors very differently, especially geographically (United States versus China). This can be better seen in Table 3: the

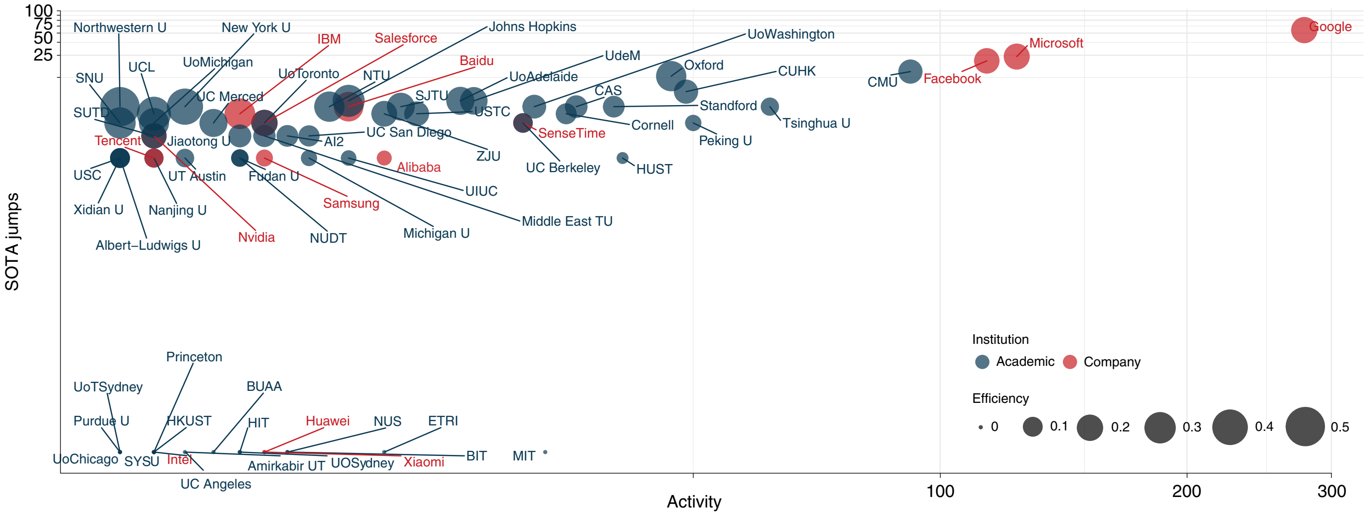

Fig. 3 | Most prolific institutions (at least ten entries) in terms of total SOTA jumps entries and activity. Both axes are logarithmic. Point size represents the efficiency (ratio between number of SOTA jumps and attempts). Note that 'Academic' represents both higher education and independent research institutions. We refer the reader to Supplementary Table 1 for further detail about the abbreviations of the institutions.

Table 3 | Aggregated results (number of institutions, activity, jumps and efficiency ratio) per continent

Continent	Institutions	Activity	Jumps	Efficiency (%)
Americas	118/68	1,327 (52.5%)/394 (43.4%)	223 (60.8%)/84 (58.7%)	16.8/21.3
Asia	106/76	858 (33.9%)/408 (45%)	82 (22.3%)/44 (30.8%)	9.6/10.8
Europe	83/29	261 (10.3%)/65 (7.2%)	53 (14.4%)/11 (7.7%)	20.3/16.9
Australia	13/10	74 (2.9%)/35 (3.9%)	9 (2.5%)/4 (2.8%)	12.2/11.4
Africa	2/1	8 (0.4%)/5 (0.5%)	0 (0%)/0 (0%)	0/0

Values are for all years/only for 2019.

aggregated participation per continent (considering all countries). While institutions from the United States represent about 56.7% of all jumps, China only represents about 18%. However, the gap becomes smaller if we only consider the recent years. For instance, Table 3 (orange) shows the same results for year 2019 only. Here the institutions from Asia come at the forefront in terms of activity compared with those from America. At the country level, activities from the United States and China are much more similar (41% versus 37%) and although the United States keeps leading the chart with respect to to the number of SOTA jumps compared with China (54% versus 26%), the difference has narrowed. This country-level concentration is also reflected when we compute the (country-wise) HHI to analyse concentration and competitiveness. In this case, the HHI is 0.33, showing a much higher concentration level per country compared with the analysis per institution.

These results are loosely consistent with analyses framing AI research progress as a 'race' being led by the United States and China (although we agree that the research community should avoid using this terminology due to its military connotations[13]). The data we analyse here represent only a small snapshot of all AI progress, but it still suggests that the United States has had a relevant lead if we look at the whole period, but the gap is being reduced by other countries such as China (Table 3). In the whole period, as Fig. 3 shows, six out of the top ten institutions are from the United States (the top three being tech giants).

The tech giants have three of the key ingredients of modern AI and the AI benchmark philosophy[14–16]: hardware (for example, computing infrastructure, especially through cloud services), software (for example, powerful libraries and platforms that are created and maintained by them, such as Keras, TensorFlow and so on) and data (for example, people's behaviour, satellite images, computer use and so on). The fourth ingredient is talent, and competitions and benchmarks play an important role in attracting young researchers.

Discussion

Benchmarks play an important role in shaping a field—deciding what is important and providing a shared challenge to focus on. They are increasingly prevalent in AI and machine learning. In machine vision, benchmarks such as ImageNet [8] and CIFAR-10[17] have supported and provided a focus for research activity. This success has led to an acceleration in the development of new and more varied benchmarks[18].

The role of benchmark performance as a metric of progress is reflected in some meta-analyses of the state of AI. For instance, the 'AI index'[19] has been summarizing the state of AI every year since 2017. In the 2019 edition of the AI index[5], chapter 3 ('Technical performance') collects the results of several benchmarks and their SOTA fronts over time. This complements more classical bibliometric analysis[20–26]. However, in this and other reports and repositories, benchmark results are not considered themselves as elements over which scientometric analysis can be done. Benchmark results are a genuine outcome of research, introducing innovations, techniques and even discoveries, and as such they warrant a level of consistent scientometric analysis that is currently lacking.

Benchmarks and competitions challenge AI to new levels, but they also create complex phenomena in which social behaviour and research policy play an important role. Jumps in some particular challenges have been identified as landmarks of the field[27–30]. However, the accompanying papers usually get many citations that simply refer to the achievement of the milestone rather than the actual use or extension of the techniques or the science underlying it. On the other hand, poor results on these benchmarks are associated with a low likelihood of getting the accompanying paper accepted in a good venue. Absence of progress on a benchmark more generally may be associated with stagnation in the field—or even AI winters. The more influential particular benchmarks become, the more careers and project funding may depend on good performance on them. Conversely, very quick progress on a benchmark may be a sign that the field is 'overfitting' on its performance, which may not translate ideally to the real world or to research challenges they are intended to be a benchmark for. This leads to a two-way dynamic: not only do benchmarks influence research, but research influences the new generation of benchmarks, in a 'challenge-solve-and-replace' evaluation dynamic[31,32].

The benchmarks may also focus overly on advances that are easily measured, in many cases reduced to a single metric of performance, to the exclusion of more important but less easily measured advances[4]. We have seen that the large number of attempts below the SOTA front may provide an important contribution beyond mere performance, such as new (disruptive) ideas that do not yet translate into SOTA jumps.

The behaviour we observe in this broad sample of AI benchmarks cannot be used as a single indicator of the way AI research evolves, its dynamics, competition and progress. This is especially the case given the original bias in the data source, Papers With Code, and our criteria for selection of benchmarks, as explained in Methods. However, the analysis of benchmarks in modern AI is sufficiently important to be a major source for the understanding of AI progress, in conjunction with some other indicators. In this Article, we have presented a new toolkit of methodological innovations, such as the use of communities, the associated plots and the aggregated indicators.

The methodology recognizes and identifies communities as teams that go beyond particular papers and perform one or more attempts on the same benchmark. In the sample of benchmarks we explore in this paper, focused on machine learning, the presence of industry is higher than in other areas of AI. This is also reflected by some of our hypotheses, and the finding that hybrid communities (industry–academia) are usually more successful than one-shot attempts. This is consistent with other scientometric studies analysing the role of hybrid teams[1,3]. However we also find some temporal dynamics relating to the size of the communities (whether new members are attracted after success), with the last two hypotheses (H_{CommRise} and H_{CommWane}) showing that the communities attract more members before a SOTA jump, but wane in relative size immediately after. Our observations of the size of the communities and the evolution of active members should be understood in the context of the develop–disrupt spectrum too[1,2]. This seems to reinforce the idea that some disruptions happen well below the SOTA front. We should then avoid a simplistic benchmark narrative that only encourages incremental developments (for example, using larger architectures and datasets).

The aggregated indicators in terms of institutions and countries are similar to those obtained with traditional bibliometric approaches[20,21,33]. However, there are important differences if we look at conferences, journals or citations. When focusing on major conferences only, four of the six top players (see www.marekrei.com/blog/ml-and-nlp-publications-in-2019/), Google, Microsoft, Carnegie Melon University and Facebook, also appear in the top four as per our Fig. 3. This overlap is expectable, so perhaps it is more meaningful to focus on particular cases where the difference is notable. For instance, Massachusetts Institute of Technology (MIT), the fourth-highest institution in terms of papers published at major AI conferences, does not appear in our 23 top institutions. Whether this is a coincidental exception or caused by different incentives at MIT with regard to AI competitions is difficult to tell. However, when looking at journals, for example, from Web of Science, the distribution is much more uniform in terms of players[20]. MIT, Toronto and Stanford top the ranking in the number of citations, and the Chinese Academy of Science, MIT and Hong Kong Polytech top the ranking in the number of papers. This discrepancy between the ranking given by the number of papers and the number of citations in bibliographic analyses can be compared with the discrepancy we found between the most active and most efficient institutions in our benchmark analysis. The research activity around AI benchmarks is an alternative source of insights into productivity and efficiency, highlighting different perspectives and incentives from the more traditional bibliometric approach.

Despite the advantages and insights, there are inherent limitations in our analysis that derive from the data source, Papers With Code. The alignment or divergence of observations may well originate from a biased sample of benchmarks. In the same way that bibliometric studies can be biased by the inclusion of journal papers only, or by using just a subset of conferences that reflect the recent dominance of machine learning in AI[3,34], the studies based on benchmarks may be affected by the selection criteria too. Similarly, there may be a bias in favour of recent papers and popular areas of research in the data of Papers With Code. This may affect the general increase of activity and the distribution of results below the SOTA fronts in later years, since old papers below the SOTA front may not have been included in Papers With Code. Furthermore, there are other factors affecting this in different directions, such as the chance of occurrence of extreme values as the scale for the performance metric approaches its limits, which in some cases may compensate for the exponential growth in submissions in recent years.

Many areas in AI cannot be reduced to benchmarks, and even where they can be, the benchmark data may not be representative or may not include the whole span of activity in an equally thorough way. If so, we encourage platforms such as Papers With Code, and any statistic or meta-review using them, to be more explicit in acknowledging these limitations, leading to an accurate reading of the results and a balanced comparison with other sources and methodologies.

Conclusions

We believe the novel way of exploring scientific activity we have introduced in this paper should be valuable in an era where measuring citations has been brought into question as a meaningful way of measuring innovation[35]. The scientometric analysis of benchmarks may present an excellent opportunity for distinguishing incremental research from more innovative undertakings. To do this, we believe that we should move beyond tracking performance alone when analysing the results of benchmarks. This does not mean that we should renounce metrics, plots and indicators. Rather, we should extend the analysis of AI results to include other dimensions of progress[4] that create more complex SOTA surfaces in the extended multidimensional space.

For instance, some results for the machine learning benchmarks are obtained with additional training data, beyond the data provided with the benchmark. The model is then adapted (through transfer) to the particular benchmark. Defining a coherent dimension for 'data economy' is challenging—but not impossible[36]. On other occasions, some teams achieve similar results with alternative algorithmic solutions that use much less computing resources[37]. As comparing computing resource use is not straightforward—but not impossible[38]— many competitions run the approaches on the same hardware, getting a normalized dimension for this (see, for example, ref. [39]). Finally, some performance results are obtained by specializing on some subpockets of problems, while failing to achieve results across all problems. There are already some proposals for defining a generality dimension[40]. We leave this more comprehensive analysis for future work; there is much to be explored in the activity that happens below the SOTA front. This activity may represent fruitful directions that may extend replicability into real reproducibility, support innovation through alternative approaches and contribute to the development of AI systems more suited to real-world applications.

Methods

We describe the methodology, including the data sources and their transformation into plots and aggregated indicators.

Data sources. Information on representative benchmarks in AI and the scientific papers associated with them have been obtained from Papers With Code (paperswithcode.com), the largest source so far hosting information on benchmarks in AI, focused on machine learning.

Papers With Code is not necessarily comprehensive. Not all areas in machine learning are equally represented, and some areas out of machine learning are covered, especially when machine learning is combined with other techniques (for example, video game agents may use deep learning for perception but also Monte Carlo tree search and planning for action). While areas such as theorem proving or combinatorial optimization are covered, no benchmarks are included yet in Papers With Code, despite the existence of important benchmarks in AI about SAT solvers, constraint satisfaction and other areas.

This coverage bias is common in some sources used by Papers With Code, such as NLP Progress (http://nlpprogress.com/), EFF AI metrics (https://www.eff.org/es/ai/metrics), SQuAD (https://rajpurkar.github.io/SQuAD-explorer/), RedditSota (https://github.com/RedditSota/state-of-the-art-result-for-machine-learning-problems) and so on, and new platforms such as stateoftheart.ai. The bias is also reinforced by a tendency of machine learning researchers to collect and process data for their own research, and may introduce a bias in favour of data-collecting giants in industry prioritizing AI techniques that excel when fed with huge amounts of data. The very nature of Papers With Code also favours a higher representation of recent papers over old papers, especially if these old papers were not on arXiv, did not reach the SOTA or are not very cited. In sum, the characteristics and limitations of Papers With Code must be recognized when interpreting our results.

Benchmarks are grouped into tasks such as image classification, action recognition or language modelling, where each task can contain several benchmarks. A benchmark is a specific problem that is measured by a single performance metric. Papers With Code collects information about the performance of different approaches during a given period, typically ranging from the introduction of the benchmark to the present day. Papers With Code links results to papers; several results can correspond to the same paper. A JSON interface can be used to export the name of the paper, the metric and the date that was published. Data comes originally from 'arXiv, Madewitml, Papers With Code (PwC), Connected Papers, ArXiv-sanity, GroundAI, Deep Learning Monitor, DistillPub, NLP Progress, and others', which may also introduce a bias in favour of the conferences they prioritize as a source (namely, the International Conference on Learning Representations, the Neural Information Processing Systems annual meeting, the International Conference on Machine Learning, the Conference on Computer Vision and Pattern Recognition, the International Conference on Computer Vision, the European Conference on Computer Vision, the International Conference on Robotics and Automation, the International Conference on Medical Image Computing and Computer Assisted Intervention, the International Conference on Intelligent Robots and Systems, the International Journal of Robotics Research, the Multidisciplinary Conference on Reinforcement Learning and Decision Making, and the Medical Imaging with Deep Learning conference). This selection reflects a shift in popularity of the AI venues, as shown in several studies[3].

Another cause of bias in this paper may come from our benchmark selection. Our criteria for inclusion are as follows. (1) Popularity: we only consider those benchmarks with at least 15 entries. (2) Diversity: we limit the maximum number of benchmarks per 'task' (the categories in Papers With Code) to two. (2a) When some problems are more challenging than others, we select the two most challenging ones. For instance, in the case of the more than 50 Atari games, deep reinforcement learning agents have struggled in a few challenging games: *Montezuma's Revenge*, *Space Invaders*, *Phoenix*, *Yars Revenge*, *Solaris*, *Ms. Pacman*, *Pitfall*, *Skiing* and so on. From these, only the first two have at least 15 entries. (2b) All things equal, we select the two most popular benchmarks for each particular 'task'. For instance, in machine translation we find the same benchmark (for example, WMT) for different pairs of languages (English–German, German–English, English–French, ...). We choose Eng–Fre and Eng–Ger. Finally, it should be noted that for a few benchmarks (for example, the MNIST database of handwritten digits and SQuAD2.0), and due to the extraction procedures performed by PwC (beyond our control), some data are systematically corrupted or missing (for example, date, performance values and so on) and, therefore cannot be properly analysed as this would introduce some bias in our results.

Papers With Code does not provide information about the authors' affiliation. We add this institution information by identifying authors' affiliations through searching Scinapse (scinapse.io), a free search engine that provides papers, authors and affiliations. Geographical information is extracted as follows. We gathered data to manually annotate the country of origin for the different authors' affiliations. We use the headquarters (for example, Google to the United States, but DeepMind to the United Kingdom), which benefits the United States even if many of their corporations have research centres in other parts of the world. See Supplementary Table 1 for more details about this annotation. A contribution counts for a country if at least one author of the paper is affiliated with an institution in that country. For robustness, we also tried doing this proportionally to the number of authors in the paper, and it produced very similar results.

Data processing. Further processing is needed to account for the following factors. Papers may have several authors; and authors may have multiple affiliations. One paper may detail several submissions to the same benchmark. Benchmarks may have positive (the higher the better, such as accuracy) or negative (the lower the better, such as error) metrics.

After this processing, how do we identify two papers as belonging to the same community? We first build adjacency matrices where each element indicates how many times a particular author has collaborated with other authors in one or more papers. From these networks, we apply the Clauset–Newman–Moore hierarchical agglomeration algorithm[6] for inferring the community structure. This algorithm gathers vertices into groups such that there is a higher density of edges within groups than between them. As an illustrative example, Supplementary Fig. 1 presents the communities discovered from the network of authors that have presented solutions for SQuAD1.1[9]. It is possible for two authors who worked on the same paper to be split into two different communities by this approach, although this is a rare occurrence (5% of authors on average). Only in this case can a paper be associated with more than one community. Overall, there is an average number of 30 ± 20.9 communities per benchmark, with 11.3 ± 6.5 members on average (note that active members in a year may be smaller, as shown in Supplementary Fig. 2).

Data availability

The data regarding all the papers analysed, their authors, community memberships, results for the different benchmarks and SOTA jumps can be found in the data folder on GitHub[41] ('data' folder).

Code availability

The code for reproducing results can be found on GitHub[41].

Received: 9 October 2020; Accepted: 7 April 2021;
Published online: 17 May 2021

References

1. Fortunato, S. et al. Science of science. *Science* **359**, eaao0185 (2018).
2. Wu, L., Wang, D. & Evans, J. A. Large teams develop and small teams disrupt science and technology. *Nature* **566**, 378–382 (2019).
3. Frank, M. R., Wang, D., Cebrian, M. & Rahwan, I. The evolution of citation graphs in artificial intelligence research. *Nat. Mach. Intell.* **1**, 79–85 (2019).
4. Martínez-Plumed, F. et al. Accounting for the neglected dimensions of AI progress. Preprint at https://arxiv.org/abs/1806.00610 (2018).
5. Perrault, R. et al. *The AI Index 2019 Annual Report* (AI Index Steering Committee, Human-Centered AI Institute, Stanford Univ. 2019); https://hai.stanford.edu/ai-index-2019
6. Clauset, A., Newman, M. E. J. & Moore, C. Finding community structure in very large networks. *Phys. Rev. E* **70**, 66–111 (2004).
7. Van Raan, A. The influence of international collaboration on the impact of research results: some simple mathematical considerations concerning the role of self-citations. *Scientometrics* **42**, 423–428 (1998).
8. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–55 (IEEE, 2009).
9. Rajpurkar, P., Zhang, J., Lopyrev, K. & Liang, P. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* 2383–2392 (Association for Computational Linguistics, 2016).
10. Bonferroni, C. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* **8**, 3–62 (1936).
11. Kwok, R. Junior AI researchers are in demand by universities and industry. *Nature* **568**, 581–584 (2019).
12. Rhoades, S. A. The Herfindahl–Hirschman index. *Fed. Res. Bull.* **79**, 188–189 (1993).
13. Cave, S. & Ó hÉigeartaigh, S. S. An AI race for strategic advantage: rhetoric and risks. In *Proc. 2018 AAAI/ACM Conference on AI, Ethics, and Society* 36–40 (Association for Computing Machinery, 2018).
14. Lee, K.-F. *AI Superpowers: China, Silicon Valley, and the New World Order* (Houghton Mifflin Harcourt, 2018).

15. Horowitz, M. C., Allen, G. C., Kania, E. B. & Scharre, P. *Strategic Competition in an Era of Artificial Intelligence* 8 (Center for New American Security, 2018).
16. Li, W. C., Nirei, M. & Yamana, K. *Value of Data: There's No Such Thing as a Free Lunch in the Digital Economy* Working Paper (US Bureau of Economic Analysis, 2019).
17. Krizhevsky, A. *Learning Multiple Layers of Features from Tiny Images.* https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf (2009).
18. Hernández-Orallo, J. et al. A new AI evaluation cosmos: Ready to play the game? *AI Magazine* **38**, 66–69 (2017).
19. Shoham, Y. Towards the AI index. *AI Magazine* **38**, 71–77 (2017).
20. Niu, J., Tang, W., Xu, F., Zhou, X. & Song, Y. Global research on AI from 1990–2014: spatially-explicit bibliometric analysis. *ISPRS Int. J. Geoinf.* **5**, 66 (2016).
21. Juan Mateos-Garcia, K. S., Klinger, J. & Winch, R. *A Semantic Analysis of the Recent Evolution of AI Research.* https://www.nesta.org.uk/report/semantic-analysis-recent-evolution-ai-research/ (NESTA, 2019).
22. Gao, F. et al. Bibliometric analysis on tendency and topics of artificial intelligence over last decade. *Microsyst. Technol.* 1–13 (2019).
23. Tran, B. X. et al. Global evolution of research in artificial intelligence in health and medicine: a bibliometric study. *J. Clin. Med.* **8**, 360 (2019).
24. Tang, X., Li, X., Ding, Y., Song, M. & Bu, Y. The pace of artificial intelligence innovations: speed, talent, and trial-and-error. *J. Inf.* **14**, 101094 (2020).
25. Qian, Y., Liu, Y. & Sheng, Q. Z. Understanding hierarchical structural evolution in a scientific discipline: a case study of artificial intelligence. *J. Inf.* **14**, 101047 (2020).
26. Serenko, A. The development of an AI journal ranking based on the revealed preference approach. *J. Inf.* **4**, 447–459 (2010).
27. Campbell, M., Hoane Jr, A. J. & Hsu, F.-h Deep Blue. *Artif. Intell.* **134**, 57–83 (2002).
28. Ferrucci, D. A. Introduction to 'This is Watson'. *IBM J. Res. Dev.* **56**, 235–249 (2012).
29. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
30. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
31. Schlangen, D. Language tasks and language games: on methodology in current natural language processing research. Preprint at https://arxiv.org/abs/1908.10747 (2019).
32. Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A. & Choi, Y. Hellaswag: can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* 4791–4800 (Association for Computational Linguistics, 2019).
33. Lei, Y. & Liu, Z. The development of artificial intelligence: a bibliometric analysis, 2007–2016. *J. Physi.* **1168**, 022027 (2019).
34. Martínez-Plumed, F. et al. The facets of artificial intelligence: a framework to track the evolution of AI. In *Proc. Twenty-Seventh International Joint Conference on Artificial Intelligence* 5180–5187 (International Joint Conferences on Artificial Intelligence Organization, 2018).
35. Bhattacharya, J. & Packalen, M. *Stagnation and Scientific Incentives* Technical Report (National Bureau of Economic Research, 2020).
36. Houghton, B. et al. Guaranteeing reproducibility in deep learning competitions. Preprint at https://arxiv.org/abs/2005.06041 (2020).
37. Lucic, M., Kurach, K., Michalski, M., Gelly, S. & Bousquet, O. Are gans created equal? A large-scale study. *Adv. Neural Inf. Process. Syst.* 700–709 (2018).
38. Hernandez, D. & Brown, T. B. Measuring the algorithmic efficiency of neural networks. Preprint at https://arxiv.org/abs/2005.04305 (2020).
39. Mattson, P. et al. MLPerf training benchmark. Preprint https://arxiv.org/abs/1910.01500 (2019).
40. Martínez-Plumed, F. & Hernández-Orallo, J. Dual indicators to analyse AI benchmarks: difficulty, discrimination, ability, and generality. *IEEE Trans. Games* **12**, 121–131 (2020).
41. Martínez-Plumed, F., Barredo, P., hÉigeartaigh, S. Ó. & Hernández-Orallo, J. AI research dynamics. *GitHub* https://github.com/nandomp/AI_Research_Dynamics (2021).
42. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T. & Serre, T. HMDB: a large video database for human motion recognition. In *2011 International Conference on Computer Vision* 2556–2563 (IEEE, 2011).
43. Soomro, K., Zamir, A. R. & Shah, M. UCF101: a dataset of 101 human actions classes from videos in the wild. Preprint at https://arxiv.org/abs/1212.0402 (2012).
44. Bellemare, M. G., Naddaf, Y., Veness, J. & Bowling, M. The arcade learning environment: an evaluation platform for general agents. *J. Artif. Intell. Res.* **47**, 253–279 (2013).
45. Timofte, R., De Smet, V. & Van Gool, L. Anchored neighborhood regression for fast example-based super-resolution. In *Proc. IEEE International Conference on Computer Vision* 1920–1927 (IEEE, 2013).
46. Hutter, M. Human knowledge compression contest. *Hutter Prize* http://prize.hutter1.net/ (2006).
47. Mikolov, T., Deoras, A., Kombrink, S., Burget, L. & Černocky, J. Empirical evaluation and combination of advanced language modeling techniques. In *Twelfth Annual Conference of the International Speech Communication Association* 605–608 (2011).
48. Dettmers, T., Minervini, P., Stenetorp, P. & Riedel, S. Convolutional 2D knowledge graph embeddings. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 32 (2018).
49. Bojar, O. et al. Findings of the 2014 workshop on statistical machine translation. In *Proc. Ninth Workshop on Statistical Machine Translation* 12–58 (Association for Computational Linguistics, 2014); http://www.aclweb.org/anthology/W/W14/W14-3302
50. Sang, E. F. & De Meulder, F. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003* 142–147 (2003).
51. Weischedel, R. et al. *Ontonotes Release 5.0 ldc2013t19 23* (Linguistic Data Consortium, 2013).
52. Lin, T.-Y. et al. Microsoft COCO: common objects in context. In *European Conference on Computer Vision* 740–755 (Springer, 2014).
53. Andriluka, M., Pishchulin, L., Gehler, P. & Schiele, B. 2D human pose estimation: new benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 3686–3693 (IEEE, 2014).
54. Yang, Y., Yih, W.-t. & Meek, C. Wikiqa: a challenge dataset for open-domain question answering. In *Proc. 2015 Conference on Empirical Methods in Natural Language Processing* 2013–2018 (Association for Computational Linguistics, 2015).
55. Cordts, M. et al. The cityscapes dataset for semantic urban scene understanding. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 3213–3223 (IEEE, 2016).
56. Everingham, M. et al. The Pascal visual object classes challenge: a retrospective. *Int. J. Comput. Vis.* **111**, 98–136 (2015).
57. Maas, A. L. et al. Learning word vectors for sentiment analysis. In *Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* Vol. 1, 142–150 (Association for Computational Linguistics, 2011).
58. Socher, R. et al. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. 2013 Conference on Empirical Methods in Natural Language Processing* 1631–1642 (Association for Computational Linguistics, 2013).
59. Panayotov, V., Chen, G., Povey, D. & Khudanpur, S. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 5206–5210 (IEEE, 2015).

Acknowledgements

F.M.-P. acknowledges funding from the AI-Watch project by DG CONNECT and DG JRC of the European Commission. J.H.-O. and S.Ó.h. were funded by the Future of Life Institute, FLI, under grant RFP2-152. J.H.-O. was supported by the EU (FEDER) and Spanish MINECO under RTI2018-094403-B-C32, Generalitat Valenciana under PROMETEO/2019/098 and European Union's Horizon 2020 grant no. 952215 (TAILOR).

Author contributions

The four authors, F.M.-P., P.B., S.Ó.h. and J.H.-O., participated in the definition and refinement of the goals of this study and the hypotheses. F.M.-P., J.H.-O. and P.B. conceived the technical methodology. P.B. and F.M.-P. implemented the code that collects and processes the data, and creates the communities. F.M.-P. generated the plots. All authors discussed the results and contributed to the writing of the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-021-00339-6.

Correspondence and requests for materials should be addressed to F.M.-P., P.B., S.Ó.h. or J.H.-O.

Peer review information *Nature Machine Intelligence* thanks Nima Dehmamy, Lars Kotthoff and Dashun Wang for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.