1st International Workshop on Evaluating Progress in Artificial Intelligence - EPAI 2020
In conjunction with the 24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Tracking the Impact and Evolution of AI: The *AIcollaboratory*

Fernando Martínez-Plumed[1,2] and **Jose Hernández-Orallo**[2] and **Emilia Gómez**[2,3]

Abstract. Artificial Intelligence (AI) is omnipresent today, but, what is AI capable of doing and where are its boundaries? Benchmarks, competitions and challenges are behind much of the recent progress in AI, but the dynamics of rushing breakthroughs at the expense of massive data and compute has led to a more complex AI landscape, in terms of what can be achieved and how. As a result, policy makers and other stakeholders have no way of assessing what AI systems can do today and in the future. In this paper, we identify a series of problems to track and understand what AI is capable of. Then we present the *AIcollaboratory*, a data-driven framework to collect and explore data about AI results, progress and ultimately capabilities, which is being developed in the context of AI Watch, the European Commission knowledge service to monitor the development, uptake and impact of AI in Europe.

1 INTRODUCTION

Artificial Intelligence (AI) has become an area of strategic importance with potential to be a key driver of economic development, with a wide range of potential social implications. In order to assess present and future impact, there is a need to analyse what AI can achieve both currently and in the future. But, what is AI capable of? This question is as crucial as elusive, and the answer becomes more difficult as AI is progressing in ways that are open-ended about the techniques and resources AI can operate with. The truth is that whenever a (computational) task is solved, researchers find it increasingly challenging to extrapolate whether the task can be reproduced, even when only a few things change: the available data, the domain knowledge, the level of noise or uncertainty, the (hyper)parameters, the techniques, the research team, the libraries, the compute, etc. In the end, we would like to infer whether a good result (or even a breakthrough) in task A transfers to a similar good result in task B. This extrapolation across tasks is precisely what the notion of *capability*, borrowed from psychology, tries to answer. However, we lack the tools, and the data, to do similarly in AI.

In this sense, several national AI strategies are accompanied by catalogues about what their research centres and AI companies claim they are able to do. The study of the impact of AI on the workplace is usually done in terms of some assumptions of what AI or ML is able to do too [11]. Similarly, the analysis of AI risks, both short-term and long-term, depend on the assessment of these capabilities [29, 30].

The need for assessing and understanding what AI is capable of is becoming more relevant as academia and industry in AI are rushing to achieve breakthroughs for particular benchmarks, specific problems or narrow tasks, usually at the expense of massive data, computation power, embedded heuristics, strong bias, system specialisation, etc. [35] Benchmarks, competitions and other kinds of challenges are behind much of the recent progress in AI, especially in machine learning (ML) [13, 27], but the dynamics of rushing breakthroughs at the expense of massive data, compute, specialisation, etc., has led to a more complex AI landscape, in terms of what can be achieved and how. As a result, policy makers and other stakeholders have no way of assessing what AI systems can do today and in the future, and how the field may get there. Similarly, there is no common framework for determining which capabilities to be included in AI systems are desirable, necessary or, on the contrary, dangerous. This does not mean that we must disregard or understate the valuable information that is provided by a plethora of benchmarks. On the contrary, the analysis of the progress of AI must be based on data-grounded evidence, relying on finding and testing hypotheses (e.g., distinguishing incremental improvement vs. real breakthroughs) through the computational analysis of big amounts of shared data [22], using open data science tools [33]. We need to assess whether new AI systems and techniques are simply an incremental improvement for a narrow collection of applications or a real breakthrough representing a more general cognitive ability, which can be established in comparison with the same abilities in humans and other animals. But this analysis must be abstracted from tasks to capabilities, for the purposes of integration[4] and evaluation [28].

In this paper, we identify a series of problems to track and understand what AI is capable of. Then we survey some previous initiatives addressing some of these problems, and we finally present the *AIcollaboratory*, a data-driven framework to collect and explore hetereogenous data sources about AI results, progress and ultimately capabilities, which is being developed in the context of *AI Watch*, the European Commission (EC) knowledge service to monitor the development, uptake and impact of AI in Europe[5]. We close the paper with some challenges for the community emerging around the collaboratory.

2 OPEN QUESTIONS

In other areas of science and technology, such as medicine or engineering, several catalogues exist, usually accompanied by methodologies and meta-analyses, where the results of several interventions (e.g., treatments in medicine or building procedures in engineering) are compared, also clarifying the operating point of each technique (when it works and what the costs and risks are).

[1] JRC, European Commission, email: {fernando.martinez-plumed, emilia.gomez-gutierrez}@ec.europa.eu@ec.europa.eu
[2] Universitat Politècnica de València, email: {fmartinez,jorallo}@dsic.upv.es
[3] Universitat Pompeu Fabra, email: emilia.gomez@upf.edu

[4] CCC AI roadmap: https://cra.org/ccc/ai-roadmap-integrated-intelligence/
[5] EC AI Watch: https://ec.europa.eu/knowledge4policy/ai-watch

Why is it so difficult to determine the capabilities that an increasingly wider range of AI systems and techniques display? The following list shows a series of questions that explain this difficulty as well as some other related problems:

- Limited understanding of how progress in some areas of AI makes new services and possibilities available (Technology Readiness Levels) [42].
- Lack of criteria to determine how specific or general AI systems are [28].
- Lack of transparency about the employed resources (e.g., data, compute, software, hardware, human oversight, etc.) and design factors (e.g., scalability, robustness, flexibility, etc.) [35, 2, 25].
- Evaluation happens beyond the traditional train and test split: train-test overlap in reinforcement learning, machine teaching, curriculum learning, self-play, generative models, etc., [6, 7, 21].
- Poor account of diversity in AI research. Are dominant paradigms, such as deep learning, reducing the technological and scientific diversity in the field? [40].
- Insufficient data and ability mapping on the AI side to determine whether AI progress is aligned with labour needs [20, 38, 46].
- Lack of comparative meta-analyses studying whether AI is converging or diverging with human capabilities and some other kinds of natural cognition [26, 31].
- Inadequate ways of determining how much overfitting to the test is taking place in AI research, lack of generalisation from current evaluation tasks to real-world scenarios, or the impact of Goodhart's law [47, 45].
- Confusion between repeatability, reproducibility and replicability [17, 24, 10, 34] and ways to certify and ensure them (e.g., tools that help to reproduce research such as CodaLab[6] or BEAT [7], or reproducibility challenges/workshops in ICLR [43] and ICML [1]).
- Need for benchmark taxonomies, their mapping to capabilities, subdisciplines and techniques, beyond a few websites and bibliographic studies [9].
- Insufficient number of AI meta-reviews based on evidence, as in other disciplines [44, 3, 31]

Most of the previous questions are intertwined and sufficiently relevant overall to justify the introduction of initiatives, settings and platforms to address them. In the following section see what has been attempted in this direction.

3 PREVIOUS INITIATIVES

In this section we describe several proposals which have been introduced in recent years to partially address one or more of the previous issues. The following list is illustrative, not comprehensive:

- **EFF AI metrics**[8]: A repository from the Electronic Frontier Foundation that collects problems and metrics to track progress from a subset of representative tasks from AI and machine learning, and aims at tracing progress on them [18].
- **Papers with Code**[9]: The largest, up to date, open repository of machine learning papers and their experimental results [4].
- **NLP-Progress**[10]: A hand-annotated repository to track the progress in Natural Language Processing (NLP), including the datasets and the current state of the art for the most common NLP tasks.
- **RedditSota**[11]: This repository provides state-of-the-art results for a variety of tasks across machine learning problems.
- **OpenML**[12]: This is an online platform for sharing and organising data, machine learning algorithms, experiments and results from predictive tasks.
- **Are we there yet ?**[13]: a crowd sourced list of result from machine learning approaches addressing some of the major visual classification, detection, and pose estimation datasets.
- **wer_are_we**[14]: an attempt at tracking recent results on speech recognition.
- **AI index Annual Report**[15]: an annual report to analyse and visualise data related to AI, it is aimed at policy makers, researchers, executives, journalists, and the general public, with the aim of developing their own criteria on the complex field of artificial intelligence.
- **Algorithmic Progress in Six Domains** [23]: it summarises data on algorithmic progress in six domains (e.g., SAT solvers, Chess and Go programs, machine learning models, integer programmings, etc.).
- **Robust Reading Competition**[16]: An on-line framework to facilitate the hosting, management and evaluation of robust reading-related competitions.
- **Animal-AI Olympics**[17]: A benchmark and competition to compare *capabilities* of RL agents using tasks/results from animal cognition [8].

However, all the previous approaches are limited in different ways: only cover parts of AI, are not fully integrated, are discontinued or not supported by stable institutions, or aim at improving AI research but without systematic analysis procedures, etc. While some of these initiatives can be used as sources for data, we think that a more solid, general and principled approach is needed for addressing the questions highlighted in section 2.

4 THE AI COLLABORATORY

As part of its Digital Single Market Strategy, the European Commission put forward in April 2018 a European strategy on AI in its Communication "*Artificial Intelligence for Europe*"[18]. The aims of the European AI strategy announced in the communication are:

- To boost the EU's technological and industrial capacity and AI uptake across the economy, both by the private and public sectors
- To prepare for socio-economic changes brought about by AI
- To ensure an appropriate ethical and legal framework.

Subsequently, in December 2018, the European Commission and the Member States published a "*Coordinated Plan on Artificial Intelligence*"[19], on the development of AI in the EU. The Coordinated Plan introduces the role of *AI Watch* to monitor its implementation:. Therefore, AI Watch aims at tracking European Union's in-

[6] https://codalab.org/
[7] https://www.beat-eu.org/platform/
[8] https://www.eff.org/ai/metrics
[9] https://www.paperswithcode.com/
[10] https://github.com/sebastianruder/NLP-progress
[11] https://github.com/RedditSota
[12] https://www.openml.org/
[13] https://rodrigob.github.io/are_we_there_yet/build/
[14] https://github.com/syhw/wer_are_we
[15] http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf
[16] https://rrc.cvc.uab.es/
[17] http://animalaiolympics.com/
[18] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN
[19] https://ec.europa.eu/knowledge4policy/publication/coordinated-plan-artificial-intelligence-com2018-795-final

dustrial, technological and research capacity in AI; AI-related policy initiatives in the Member States; uptake and technical developments of AI; and AI impact with a European focus but within the global landscape. AI Watch is being developed by the Joint Research Centre[20] (JRC) of the European Commission in collaboration with the Directorate-General for Communications Networks, Content and Technology (DG CONNECT).

Under the umbrella of the AI Watch, the *AIcollaboratory* is being materialised to monitor the impact and evolution of AI. The *AIcollaboratory* aims to develop a synergetic initiative for the analysis, evaluation, comparison and classification of AI systems. It is based on a thorough analysis of the requirements of the community [9], and an understanding of the difficulties but possibilities of using an ability-based view rather than a task-based AI evaluation approach [28, 26]. Regarding the latter, note that while specialised AI systems require a task-oriented evaluation, more general-purpose AI systems require an ability-oriented evaluation where a system is characterised by its skills and competence rather than by the tasks it is able to solve. The use of cognitive abilities when evaluating AI systems gives more explanatory and predictive power. While a task-oriented evaluation can hardly extrapolate results from task A to task B, an ability-oriented evaluation can do this if tasks A and B are similar or fall under the same category in a hierarchy. Regarding which abilities should be measured, we may use the hierarchical theories of intelligence from psychology, animal cognition and AI (e.g., CHC model [12]).

This is yet another indication that we need to characterise the different problems in a more informative way than using performance, which is insufficient to get a proper insight of what the systems are really able to do (and how they achieve it) and what the problems are evaluating. One key observation is the duality between tasks and systems, with capabilities being the latent variables that connect them. This idea is common in cognitive psychology and psychometrics, where different approaches and tools are useful to describe systems (humans or animals) by a collection of factors and constructs derived from observations and experiments. This is especially the case of Item response Theory (IRT) [19], a family of mathematical models which not only assigns these constructs to agents/systems (abilities), but it also derives task/problems indicators (difficulty and discrimination). Recently, IRT [19] has been adopted (and extended) to analyse ML/AI systems and problems in a more comprehensive way [36, 39, 32, 14, 37]. By considering AI problems as items and AI systems and algorithms as respondents, IRT is applicable to any area in AI. Then, we can reunderstand the proficiency (or ability) of a method as the difficulty level whose problems the method is able to solve, as well as difficulty and discrimination as key indicators for AI problems.

Another important insight is about the way systems and tasks may be organised and/or classified in the *AIcollaboratory*: there is no single true hierarchy; we can build different hierarchies and arrangements in both directions: (1) tasks gathered in the *AIcollaboratory* may be aggregated into clusters depending on their inherent characteristics (e.g., difficulty, discrimination, etc.) or goals (e.g., image classification, QA, language modelling, sentiment analysis, etc.), and ultimately into abilities (i.e., visual perception, communication, navigation, interaction, etc.); and (2) systems may be aggregated into species, families or technologies (e.g., connectionist, evolutionary, Bayesian, etc.) which can be obtained from the literature.

With this main insights in mind (collaborative platform, comprehensive evaluation and hierarchical organisation), in the following we further introduce the *AIcollaboratory* in terms of the infrastructure developed, the multidimensional design followed and the data-sources used. we also briefly describe a few research initiatives we are currently carrying out using this framework.

4.1 Infrastructure

The *AIcollaboratory* aims at integrating open data and knowledge in three different domains: an inventory of intelligent systems, a behavioural test catalogue and the measurement experiments obtained when systems take tests. Namely:

- *Inventory of intelligent systems*, which incorporates information about current, past and future intelligent systems. As previously indicated, they may be aggregated from individuals to species, groups or organisations, with populations and distributions over them. Examples of this kind of inventories are common in other areas, such as the animal taxonomies (Wikispecies[21], the Encyclopedia of Life[22] or the Catalogue of Life[23]), language taxonomies (Rosetta project[24]) and medical taxonomies [16].
- *Behavioural Test Catalogue*, which integrates a series of behavioural tests, including, if available, several other characteristics such as the dimensions they measure and for which kinds of systems, the possible interfaces and testing apparatus, etc. We also find similar examples in different areas, from AI (e.g.: OpenAI Gym[25] or the Deepmind PsychLab[26]) to psycology or animal cognition (e.g.: "Mental Measurement Yearbooks"[27])
- *Repository of experimentation*, which records the results (measurements) of a wide range of intelligent systems for several tests and benchmarks, as the main data source of the *AIcollaboratory*. Data is contributed from scientific papers, experiments, code repositories, AI/robotic competitions, etc. [28, 35, 39].

In a context of open science [15], these three pieces —roughly corresponding to subjects, instruments and results— have an open and collaborative character, encouraging researchers to be users but also contributors, and thus facilitating cross-comparison and reproducibility. In this regard, researchers and practitioners can share their new data (systems, tasks, results and metrics) for a new entry in the *AIcollaboratory* by sending a pull request to our GitHub repository[28] by using the simple template provided.

Finally, on top of the previous there is a series of exploitation and visualisation tools (see Figure **??**) where actual data science takes place, including interfaces for powerful (dis)aggregation and cross-comparisons, reuse of predefined multidimensional filters, trend analysis along time, interactive interfaces to perform projections, depicting trajectories, visual categorisations, etc.

4.2 Design

We follow a multidimensional perspective (over a relational database[29]) to model the information system behind the *AIcollaboratory* (see Figure 2). This type of perspective is designed to address

[20] https://ec.europa.eu/jrc/en

[21] https://en.wikipedia.org/wiki/Wikispecies
[22] http://www.eol.org/
[23] http://catalogueoflife.org/
[24] https://en.wikipedia.org/wiki/Rosetta_Project
[25] https://gym.openai.com/
[26] https://deepmind.com/blog/open-sourcing-psychlab/
[27] http://buros.org/mental-measurements-yearbook
[28] https://github.com/nandomp/AICollaboratory
[29] See the ER diagram developed here: https://github.com/nandomp/AICollaboratory/blob/master/MySQL/Atlas_ERR_v1.png

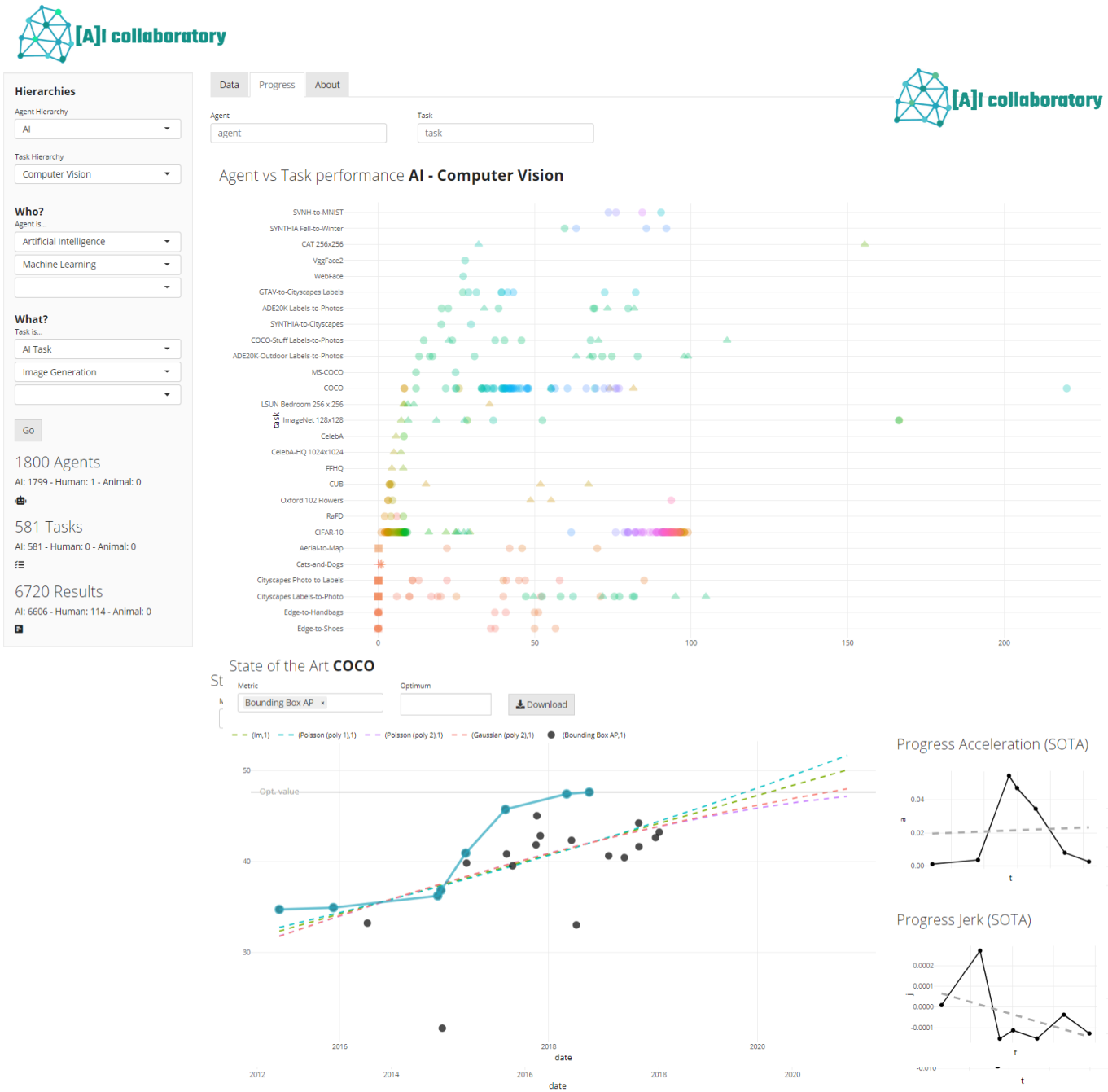

Figure 1: Screenshot of the AI Collaboratory. (Top) results from several DeepNN-based approaches addressing a set of computer vision benchmarks for image generation. (Bottom) progress over time for those AI systems addressing a particular benchmark for this task (including its pareto frontier).

the unique needs of very large databases designed for the analytical purpose. The main idea is that each piece of information in the *AIcollaboratory* (e.g., results, measures, etc.) is characterised by a number of variables or dimensions. Each dimension has a particular structure to capture (part of) the information, ontologies, hierarchies, constructs about intelligence, tests, etc., in the literature.

An example of (textual) information in the *AIcollaboratory* is:

> A DQN system, using the parameters of Mnih et al. 2015, is evaluated over Moctezuma game (from ALE v1.0) using 100,000 episodes, with a measured score of 23.3.

The key elements of the multidimensional model of the *AIcollaboratory* define the "WHO", "WHAT" and "HOW" dimensions for a specific result stored in the database. In further detail:

- WHO dimension: information regarding (artificial) agents (e.g., systems, architectures, algorithms, etc.). It resembles the Cognitive System Inventory. An example of information stored in this dimension:

> "RAINBOW" is a "Deep Learning architecture" according to "Hessel et al., 2013"
>
> "weka.PART(1)_65" is "Rpart" technique according to "OpenML"

- WHAT dimension: information regarding the taks to solve (e.g., instances, datasets, task, tests, etc.). It resembles the Behavioural Test Catalogue. An example of information stored in this dimensipn:

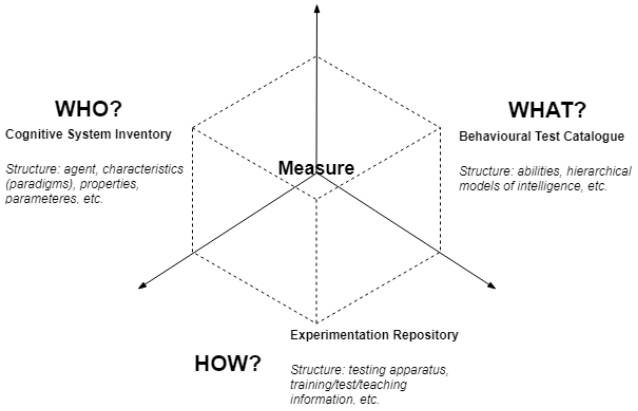

Figure 2: Multidimensional perspective of the *AIcollaboratory*

"Moctezuma" is an "Atari-Game" according to "Bellemare et al., 2013"
"Winograd schemas" requires "Commonsense" according to "Levesque 2014"

- HOW dimension: information about the testing apparatus (e.g., CV, hs, noop, splitting). It resembles the Experimentation Repository. An example of information stored in this dimension:

"Cross-Validation-Anneal" has "5" number of folds and "2" repetitions according to "OpenML"
"PriorDuel-noop" has "no-op actions" as procedure, "57" games in testing phase and "200M" training frames according to "Wang et al, 2015"

We have also defined different many-to-many relationships so each agent/task (1) *is* of a particular type; (2) *has* different attributes, which are shared by others; (3) and *belongs to* a (set of) specific hierarchy(ies) which allow us to define different grouping (and thus (dis)aggregations), all the above always according to some specific research source/reference[30].

4.3 Data

Initially, we consider a comprehensive set of AI benchmarks for our framework based on our own previous analysis and annotation of AI papers [28, 35, 39] as well as open resources which draw on data from multiple (verified) sources, including academic literature, review articles and code platforms focused on machine learning and AI. This includes some of those more active platforms such as *EFF AI metrics*, *Papers with Code* or *OpenML*, already introduced in section 3.

We rely on interfaces, APIs (when available) or web-scraping techniques in order to gather the data from the above repositories. In this regard, Papers with code provides their data (e.g., papers with abstracts, links to code repositories and evaluation tables) under a Creative Commons (CC-BY-SA-4) licence, and therefore enabled us to copy, redistribute, remix, transform, and build upon it. To access the data, we rely on the raw files (json) the platforms provide, which are updated once a week. For its part, EFF AI metrics provides a Jupyter/IPython notebook hosted in Github where the community can both contribute and gather the already collected data. Finally, OpenML offers interfaces in multiple programming languages which allow scientists to interact with the server using language-specific functions. Note that, once retrieved, all the data from the selected repositories should be then cleaned, transformed and structured (following an automatic ad-hoc ETL procedure) to fit *AIcollaboratory* multidimensional data model

Therefore, from the aforementioned sources we can track the reported evaluation results (when available or sufficient data is provided) on different metrics of AI performance across separate AI and machine learning benchmarks (e.g., datasets, competitions, awards, etc.). This is possible from a number of AI domains, including (among others) computer vision, speech recognition, music analysis, machine translation, text summarisation, information retrieval, robotic navigation and interaction, automated vehicles, game playing, prediction, estimation, planning, automated deduction, etc. This ensures a broad coverage of AI tasks. Currently, the *AIcollaboratory* stores data and knowledge from around 600 different tasks (involving around 350 benchmarks) and 1800 agents (AI systems, algorithms, approaches, etc.), having a total of 7000 results. Figure 3 shows some statistics about the type of data included in the *AIcollaboratory*. On the left we see the clear dominance of results (and benchmarks) related to visual perception, games and NLP tasks. On the right we show the astonishing number of different performance metrics used for evaluating the benchmarks.

4.4 Applications

As already introduced, the *AIcollaboratory* is a data-oriented instrument that collects and organises results at the technical level (e.g., benchmarks, competitions, resources, etc.), and linked to the research meta-level (e.g., research production, activity) and the societal level (e.g., jobs, skills). Among other applications, the *AIcollaboratory* can be used to perform in-depth (meta-)analyses and reviews using all the gathered data to structure and better understand (1) the potential impacts and benefits of AI and (2) the characteristics and evolution of the AI landscape. In the following we briefly describe some of the research independent initiatives and applications we are currently performing using the data and knowledge gathered in the *AIcollaboratory*.

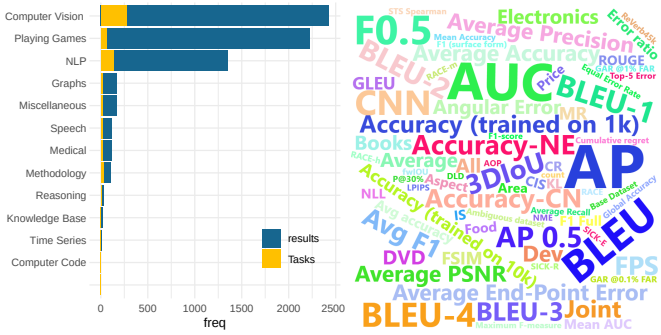

Figure 3: (Left) Number of results (blue) and benchmarks in the *AIcollaboratory* per type of task addressed. (Right) Tag cloud of the different metrics used. Figures dated December 2019, *AIcollaboratory* is constantly updated and will be enriched over time with further relevant data, knowledge and results.

Regarding the analysis of the potential impacts and benefits of AI, the data from the *AIcollaboratory* has recently been applied for analysing the influence of AI on occupations. In the framework developed we map the results of around 350 AI benchmarks with the results of labour-force surveys from three different sources: the Eu-

[30] See http://www.aicollaboratory.org/ for further information.

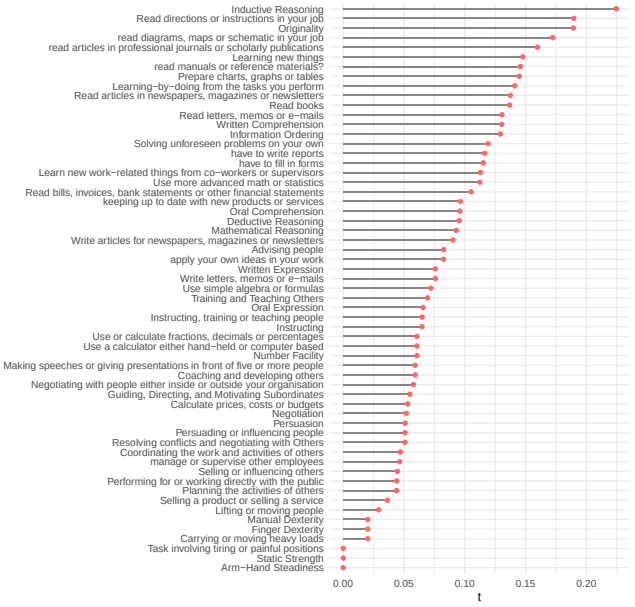

Figure 4: Labour-related tasks ranked in descending order regarding the level of impact of AI (figure from [38]). Those with the highest values consist almost entirely of information gathering and processing tasks as well as performing tasks without using explicit instructions, relying on patterns and inference instead. On the other hand, the lowest-scoring tasks are largely non-cognitive tasks that require a high degree of physical effort and dexterity (e.g., steadiness, manual/finger dexterity, etc.). At the same time, there are also plenty of non-routine interpersonal tasks that include a human component.

ropean Working Conditions Survey (EWCS)[31]; the OECD Survey of Adult Skills (PIAAC)[32]; and the database from the Occupational Information Network (O*NET)[33] (see [38, 46] for further information). This setting combines occupations and tasks from the labour market with AI research benchmarks through an intermediate layer of cognitive abilities which allows for examining the relation between the distribution of research intensity (or activity) in AI research (e.g., where the AI research community is putting the focus) and the relevance for a range of work tasks (and occupations) in current and simulated scenarios. The identification of the specific cognitive abilities that can be performed by AI gives a broader understanding on the impact of AI, as the inner layer is more independent of particular occupations, tasks or AI benchmarks. The final goal has been to analyse (i) what impact current AI research activity has on labour-related tasks (see Figure 4), and (ii) what areas of AI research activity would be responsible for a desired or undesired effect on specific labour occupations (see Figure 5).

As an application for analysing the characteristics and evolution of the AI landscape, data gathered in the *AIcollaboratory* (as well as from other sources such as *Scinapse*[34] providing papers, authors and affiliations) is also being used to perform (meta)-analyses about the breakthroughs obtained in different AI benchmarks and challenges, linking their underlying research papers, extracting the co-author communities from bibliometric sources and studying the Pareto fron-

[31] https://www.eurofound.europa.eu/surveys/european-working-conditions-surveys
[32] https://www.oecd.org/skills/piaac/
[33] https://www.onetonline.org/
[34] https://scinapse.io/

Figure 5: AI benchmarks ranked in descending order based on the interest they should attract from AI research in order to have an impact in the selected occupation(s) (figure adapted from [38]). Benchmarks are coloured according to the area they belong to (e.g., computer vision, dialogue and interaction, commonsense inference, etc.). For instance, in order for AI developments to have an effect on general office clerks, AI research should focus on those benchmarks related to inspection and data extraction as well as on the development of narratives, question answering and social interaction.

tiers over time, as defined by breakthroughs for each benchmark. The main idea is to derive research communities and analysing the results beyond specific performance metrics. This sort of analyses performed may show, among other things, whether activity patterns may be caused by breakthroughs and vice versa, whether the different benchmarks and challenges are dominated by a small group of players, with tech giants (e.g., Google, Microsoft or Facebook) and a few universities (e.g., Carnegie Mellon, Washington and Cornell) representing most breakthroughs, or whether efficiency, measured as breakthroughs vs activity, distributes among actors very differently, especially geographically (e.g., USA vs China). In Figure 6 we illustrate the different communities behind the participation and results in a specific benchmark for image classification (see [5] for further information).

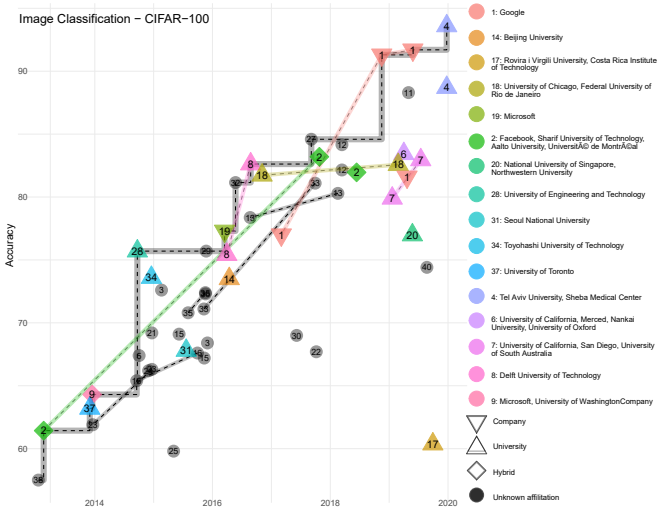

Figure 6: Activity patterns behind AI breakthroughs in the CIFAR-100 benchmark for image classification. The plot shows the key community/actors leading this task over the years.

5 CONCLUSIONS

In this report we have presented the *AIcollaboratory*, a data-driven framework developed in the context of the AI Watch initiative to collect and explore data about AI results, progress and ultimately capabilities. As stated, the main goal of the *AIcollaboratory* is to extract quantitative information related to the state of the art, challenges and trends of AI research and development in order to facilitate further qualitative analysis.

Several issues arise in this endeavour. One of the challenges of mapping systems with tasks, apart from the difficulty of extracting and integrating heterogeneous data from multiple sources, is that there are many possible hierarchies of abilities to map them. We have to realise that these hierarchies will always evolve and be refined as our understanding of AI, and intelligence in general, progresses.

A second challenge of the *AIcollaboratory* is maximising engagement by the AI community. Many initiatives do not get enough inertia, funding or popularity and are soon discontinued. We plan to address this in two ways. First, we take data and plan to co-operate with some other initiatives. For instance, we do not aim to replace current innitiatives, but to co-operate with them and address different goals and scopes, with the *AIcollaboratory* covering the whole of AI, and focused on analysing progress, impact, etc., and not application in meta-learning, auto-ml, etc. Second, the *AIcollaboratory* is an integral part of the EC's AI Watch initiative[4] of the European Commission. This ensures the future stability and continuity for the years to come, and an important number of regular users to perform meta-analysis, publications, etc.

Other challenges of AI also translate to the *AIcollaboratory*. For instance, we need to tackle the notion of generality in AI, better understand how theories of intelligence move between cognition and AI, clearly distinguish the results and the resources used in AI breakthroughs, and many others. Precisely because these are challenges to the *AIcollaboratory*, we are going to make all these questions more visible in the agenda of AI and involve more people in solving them.

Ultimately, the *AIcollaboratory* aims to be a networked ecosystem [41] allowing people all over the world to collaborate and build directly on each other's latest data, knowledge and results, also providing provide important benefits for the scientific community and policymakers, as well as produce innovative basic research at the core of the science of intelligence. This will contribute to a richer understanding of intelligence, and a better steering of AI progress and its effects on natural intelligence.

ACKNOWLEDGEMENTS

This material is based upon work supported by the EU (FEDER), and the Spanish MINECO under grant RTI2018-094403-B-C3, the Generalitat Valenciana PROMETEO/2019/098. F. Martínez-Plumed acknowledges funding of the AI-Watch project by DG CONNECT and DG JRC of the European Commission. J. Hernández-Orallo and S. Ó hÉigeartaigh are also funded by an FLI grant RFP2-152.

REFERENCES

[1] Reproducibility Workshop Series. https://sites.google.com/view/icml-reproducibility-workshop/icml2018, 2019.
[2] D. Amodei and D. Hernandez, 'Ai and compute', *Heruntergeladen von https://blog. openai. com/aiand-compute*, (2018).
[3] E. Aromataris, R. Fernandez, C. M. Godfrey, C. Holly, H. Khalil, and P. Tungpunkom, 'Summarizing systematic reviews: methodological development, conduct and reporting of an umbrella review approach', *International journal of evidence-based healthcare*, **13**(3), 132–140, (2015).
[4] AtlasML. Papers with code. https://www.paperswithcode.com/, 2019.
[5] P. Barredo, J. Hernández-Orallo, F. Martínez-Plumed, and S. Ó. hÉigeartaigh, 'The scientometrics of ai benchmarks: Unveiling the underlying mechanics of ai research', in *Evaluating Progress in Artificial Intelligence (EPAI 2020)*. ECAI, (2020).
[6] A. Benavoli, G. Corani, J. Demšar, and M. Zaffalon, 'Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis', *The Journal of Machine Learning Research*, **18**(1), 2653–2688, (2017).
[7] D. Berrar and W. Dubitzky, 'Should significance testing be abandoned in machine learning?', *International Journal of Data Science and Analytics*, **7**(4), 247–257, (2019).
[8] B. Beyret, J. Hernández-Orallo, L. Cheke, M. Halina, M. Shanahan, and M. Crosby, 'The animal-ai environment: Training and testing animal-like artificial cognition', *arXiv preprint arXiv:1909.07483*, (2019).
[9] S. Bhatnagar, A. Alexandrova, S. Avin, S. Cave, L. Cheke, M. Crosby, J. Feyereisl, M. Halina, B. S. Loe, F. Martínez-Plumed, et al., 'Mapping intelligence: Requirements and possibilities', in *"Philosophy and Theory of Artificial Intelligence*, pp. 117–135. Springer, (2017).
[10] X. Bouthillier, C. Laurent, and P. Vincent, 'Unreproducible research is reproducible', in *Proceedings of the 36th International Conference on Machine Learning*, pp. 725–734, (2019).
[11] E. Brynjolfsson and T. Mitchell, 'What can machine learning do? workforce implications', *Science*, **358**(6370), 1530–1534, (2017).
[12] J. B. Carroll et al., *Human cognitive abilities: A survey of factor-analytic studies*, Cambridge University Press, 1993.
[13] D. Castelvecchi, 'Tech giants open virtual worlds to bevy of ai programs', *Nature*, **540**(7633), (2016).
[14] Y. Chen, T. de Menezes e Silva Filho, R. B. C. Prudêncio, T. Diethe, and P. A. Flach, '$\beta^3-irt: A new item response model and its applications', in *AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pp. 1013–1021, (2019).
[15] O. S. Collaboration et al., 'Estimating the reproducibility of psychological science', *Science*, **349**(6251), aac4716, (2015).
[16] N. R. Council et al., *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease*, National Academies Press, 2011.
[17] C. Drummond, 'Replicability is not reproducibility: nor is it good science', *Evaluation Methods for ML (ICML)*, (2009).
[18] P. Eckersley, Y. Nasser, et al. Eff ai progress measurement project. https://www.eff.org/es/ai/metrics, 2017.
[19] S. E. Embretson et al., *Item response theory*, Psychology Press, 2013.
[20] E. Fernández-Macías, E. Gómez, J. Hernández-Orallo, B. S. Loe, B. Martens, F. Martínez-Plumed, and S. Tolan, 'A multidisciplinary task-based perspective for evaluating the impact of ai autonomy and generality on the future of work', in *AEGAP, July 2018, Stockholm, Sweden*. IJCAI, (2018).
[21] P. Flach, 'Performance evaluation in machine learning: The good, the bad, the ugly and the way forward', in *33rd AAAI*, (2019).
[22] V. Gewin, 'Data sharing: An open mind on open data', *Nature*, **529**(7584), 117–119, (2016).
[23] K. Grace, 'Algorithmic progress in six domains', Technical report, Machine Intelligence Research Institute, (2013).
[24] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, 'Deep reinforcement learning that matters', *CoRR abs/1709.06560*, (2017).
[25] D. Hernandez and T. B. Brown, 'Measuring the algorithmic efficiency of neural networks', *arXiv preprint arXiv:2005.04305*, (2020).
[26] J. Hernández-Orallo, *The measure of all minds: evaluating natural and artificial intelligence*, Cambridge University Press, 2017.
[27] J. Hernández-Orallo et al., 'A new AI evaluation cosmos: Ready to play the game?', *AI Magazine*, **38**(3), 66–69, (2017).
[28] J. Hernández-Orallo, 'Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement', *Artificial Intelligence Review*, **48**(3), 397–447, (Oct 2017).
[29] J. Hernández-Orallo, F. Martínez-Plumed, S. Avin, and S. Ó. hÉigeartaigh, 'Surveying safety-relevant AI characteristics', in *Workshop on Artificial Intelligence Safety 2019 co-located with the Thirty-Third AAAI Conference on Artificial Intelligence 2019 (AAAI-19), Honolulu, Hawaii, January 27, 2019*, (2019).
[30] J. Hernández-Orallo, F. Martínez-Plumed, S. Avin, and S. Ó. hÉigeartaigh, 'Ai paradigms and ai safety: Mapping artefacts and tech-

niques to safety issues', in *24th European Conference on Artificial Intelligence (ECAI 2020), Santiago de Compostela, Spain, September 1-5*, (2020).
[31] J. Ioannidis, 'Meta-research: Why research on research matters', *PLoS biology*, **16**(3), e2005468, (2018).
[32] J. P. Lalor, H. Wu, T. Munkhdalai, and H. Yu, 'Understanding deep learning performance through an examination of test set difficulty: A psychometric case study', in *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4711–4716, (2018).
[33] J. S. Lowndes et al., 'Our path to better science in less time using open data science tools', *Nature ecology & evolution*, **1**(6), 0160, (2017).
[34] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, 'Are gans created equal? a large-scale study', in *Advances in neural information processing systems*, pp. 700–709, (2018).
[35] F. Martínez-Plumed, S. Avin, M. Brundage, A. Dafoe, S. Ó. hÉigeartaigh, and J. Hernández-Orallo, 'Accounting for the neglected dimensions of ai progress', *arXiv preprint arXiv:1806.00610*, (2018).
[36] F. Martínez-Plumed, R. Prudêncio, A. Martínez-Usó, and J. Hernández-Orallo, 'Making sense of item response theory in machine learning', in *Proceedings of the Twenty-second European Conference on Artificial Intelligence*, pp. 1140–1148. IOS Press, (2016).
[37] F. Martínez-Plumed, R. B. C. Prudêncio, A. M. Usó, and J. Hernández-Orallo, 'Item response theory in AI: analysing machine learning classifiers at the instance level', *Artif. Intell.*, **271**, 18–42, (2019).
[38] F. Martínez-Plumed, S. Tolan, A. Pesole, J. Hernández-Orallo, E. Fernández-Macías, and E. Gómez, 'Does AI qualify for the job? A bidirectional model mapping labour and AI intensities', in *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2020, New York, USA, February 7-8, 2020*, (2020).
[39] F. Martínez-Plumed and J. Hernández-Orallo, 'Dual indicators to analyze ai benchmarks: Difficulty, discrimination, ability, and generality', *IEEE Transactions on Games*, **12**(2), 121–131, (2020).
[40] F. Martínez-Plumed, B. S. Loe, P. Flach, S. hÉigeartaigh, K. Vold, and J. Hernández-Orallo, 'The facets of artificial intelligence: A framework to track the evolution of ai', in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 5180–5187, (7 2018).
[41] M. Nielsen, *Reinventing discovery: the new era of networked science*, Princeton University Press, 2011.
[42] OECD, *Artificial Intelligence in Society*, OECD Publishing, 2019.
[43] J. Pineau, K. Sinha, G. Fried, R. N. Ke, and H. Larochelle, 'ICLR Reproducibility Challenge 2019', *ReScience C*, **5**(2), 5, (May 2019).
[44] V. Smith, D. Devane, C. M. Begley, and M. Clarke, 'Methodology in conducting a systematic review of systematic reviews of healthcare interventions', *BMC medical research methodology*, **11**(1), 15, (2011).
[45] R. Thomas. The problem with metrics is a big problem for AI. https://www.fast.ai/2019/09/24/metrics/, 2019.
[46] S. Tolan, A. Pesole, F. Martínez-Plumed, E. Fernández-Macías, J. Hernández-Orallo, and E. G. and, 'Measuring the occupational impact of AI beyond automation: tasks, cognitive abilities and AI benchmarks', *Submitted*, (2020).
[47] C. Wiggins. Ethical Principles, OKRs, and KPIs: what YouTube and Facebook could learn from Tukey. https://datascience.columbia.edu/ethical-principles-okrs-and-kpis-what-youtube-and-facebook-could-learn-tukey, 2018.