ECAI 2020
G.D. Giacomo et al. (Eds.)

doi:10.3233/FAIA200451

Tracking AI: The Capability Is (Not) Near

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

1 INTRODUCTION

AI is an area of strategic importance with potential to be a key driver of economic development and 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 (and will) achieve. But, what is AI capable of? This question is as crucial as elusive, 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 task is solved, researchers find increasingly challenging to extrapolate whether this task can be *reproduced*, even when only a few things change: the data, the domain knowledge, the level of uncertainty, the (hyper)parameters, the techniques, the team, the compute, etc. In the end, we would like to infer whether a good result (or a breakthrough) in task A *transfers* to a similar good result in task B. This extrapolation 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.

Benchmarks, competitions and challenges are behind much of the recent progress in AI, especially in machine learning (ML) [10], 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. 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 through the computational analysis of big amounts of shared data [6], using open data science tools [11]. But this analysis must be abstracted from tasks to capabilities, for the purposes of integration[3] and evaluation [8].

In this paper, we identify a series of problems to track and understand what AI is capable of, surveying some previous initiatives. We present the *AIcollaboratory*, a data-driven framework to collect and explore data about AI results, progress and ultimately capabilities, 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[4]. We close the paper with some challenges for the community emerging around the collaboratory.

2 OPEN QUESTIONS AND INITIATIVES

In other areas of science and technology, 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). Why is it so difficult to determine the capabilities that AI systems and techniques display? Some possible reasons are:

- Lack of criteria to determine how specific or general AI systems are [8], and no transparency about the employed resources [12].
- More complex evaluation: train/test overlap in RL, machine teaching, curriculum learning, self-play, generative models, etc. [5].
- Poor account of diversity in AI research. Are dominant paradigms (e.g., DL) reducing the scientific diversity in the field? [15].
- Insufficient data and ability mapping on the AI side to determine whether AI progress is aligned with labour needs [14].
- Lack of comparative meta-analyses studying whether AI is converging or diverging with natural intelligence [9].
- Confusion between repeatability, reproducibility and replicability [3] and ways to certify and ensure them (see e.g., [17]).
- Limited understanding of how progress in AI makes new services and possibilities available (Technology Readiness Levels) [16].
- Need for benchmark taxonomies, their mapping to capabilities, subdisciplines and techniques [1].

Most of the previous questions are intertwined and sufficiently relevant overall to justify initiatives and platforms to address them. Several proposals exist (see Tab. 1), but 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 rather than really understanding it. While some of these initiatives can be used as sources for data, a more solid, general and principled approach is needed for addressing the above questions.

Table 1. AI repositories, projects, research initiatives and reports

Repository	Description
EFF AI metrics	Problems and metrics to track progress from a subset of tasks from AI and ML (https://www.eff.org/ai/metrics)
Papers with Code	The largest, up to date, open repository of ML papers and their results (https://www.paperswithcode.com/)
NLP-Progress	A hand-annotated repository to track the progress in Natural Language Processing (NLP) (https://github.com/sebastianruder/NLP-progress)
RedditSota	State-of-the-art results for a variety of tasks across ML problems (https://github.com/RedditSota)
OpenML	Online ML platform for sharing and organising data, ML algorithms and experiments (https://www.openml.org/)
AI Index	Annual report analysing and visualising data related to AI, aimed at policy makers, researchers, executives, journalists, etc. (https://aiindex.org/)
Algorithmic Progress	Summary of data on algorithmic progress in six domains (e.g., SAT solvers, Chess/Go, ML models, integer programmings, etc.) [7]
Animal-AI olympics	A benchmark and competition to compare *capabilities* of RL agents using tasks/results from animal cognition (http://animalaiolympics.com/).

3 THE AI WATCH'S COLLABORATORY

The *AIcollaboratory* is being developed in the context of the AI WATCH initiative to monitor the European Commission's "*Coordinated Plan on Artificial Intelligence*"[5] on the development, uptake and impact of AI in the EU. From AI WATCH developments and in-depth analyses, the strengths and needs of the AI landscape will be

[1] Technical University of Valencia, email: {fmartinez,jorallo}@dsic.upv.es
[2] JRC, European Commission, email: {fernando.martinez-plumed, emilia.gomez-gutierrez}@ec.europa.eu@ec.europa.eu
[3] CCC AI roadmap: https://cra.org/ccc/ai-roadmap-integrated-intelligence/
[4] EC AI Watch: https://ec.europa.eu/knowledge4policy/ai-watch
[5] See https://ec.europa.eu/knowledge4policy/publication/coordinated-plan-artificial-intelligence-com2018-795-final

identified, providing an independent assessment of the impacts and benefits of AI on growth, jobs, education, and society[4].

For its part, the *AIcollaboratory* aims to develop a synergetic initiative for the analysis, evaluation, comparison and classification of AI systems. It is based on an understanding of the difficulties but possibilities of using an ability-based view rather than a task-based AI evaluation approach (where a system is characterised by its competence rather than by the tasks it is able to solve) [8, 9], and a thorough analysis of the requirements of the community [1]. One of the key observations is the duality between tasks and systems, with capabilities being the latent variables that connect them [13]. This idea is common in psychometrics, and especially in IRT [4], which not only assigns these constructs to agents (abilities), but it also derives task indicators (difficulty and discrimination). An important insight is that there is no single true hierarchy, we can build different hierarchies in both directions: (1) tasks are aggregated into clusters (e.g., according to their characteristics or goals) and ultimately into abilities, and (2) systems are aggregated into families or technologies.

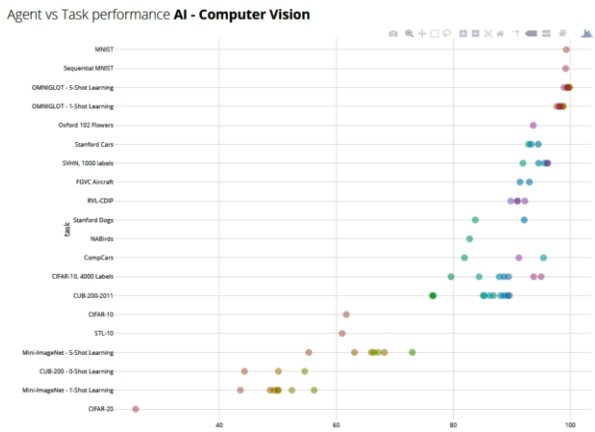

Figure 1. Screenshot of the AI Collaboratory[7]. Progress over time for those AI systems addressing a particular set of benchmarks for computer visión.

The *AIcollaboratory (a)* is conceived as a data-oriented instrument that incorporates information about current, past and future intelligent systems; *(b)* integrates a series of behavioural tests, the dimensions they measure and for which kinds of systems; and *(c)* records the results (measurements) of a wide range of intelligent systems for several tests and benchmarks. Furthermore, these three elements must rest on (and also affect) a cumulative corpus of knowledge in cognitive science and intelligence research, covering constructs, theories, ontologies, etc. The representations, aggregated information and data analysis would come by using exploitation tools over this platform (see Fig. 1). In a context of open science [2], this platform is populated in an open and collaborative fashion, facilitating cross-comparison and reproducibility.

We follow a multidimensional perspective to model the information system behind the *AIcollaboratory*. The main idea is that each piece of information is characterised by a number of dimensions defining the "WHO" (e.g., AI systems) and the "WHAT" (e.g., tasks), so covering the duality mentioned above. Finally, there is a third major dimension: "HOW" (e.g., testing apparatus) for a specific result (fact) stored in the collaboratory. 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(s) which allow us to define different grouping (and thus (dis)aggregations)[8].

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

4 CHALLENGES

One of the challenges of mapping systems with tasks 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 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, such as *OpenML* (with the *AIcollaboratory* covering the whole of AI and also natural intelligence, and focused on analysing progress, impact, etc.). Second, the *AIcollaboratory* is an integral part of the EC's AI WATCH initiative[4], which ensures future stability and continuity.

Other challenges of AI also translate to the collaboratory. 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 AI Collaboratory, 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 provide important benefits for the scientific community and policymakers, as well as produce innovative basic research at the core of the science of intelligence, contributing to a richer understanding of intelligence, and a better steering of AI progress and its effects on natural intelligence.

ACKNOWLEDGEMENTS

Work supported by EU (FEDER), Spanish MINECO (RTI2018-094403-B-C3), Generalitat Valenciana (PROMETEO/2019/098), INCIBE, EC, UPV (PAID-06-18) and FLI (RFP2-152).

REFERENCES

[1] S. Bhatnagar et al., 'Mapping intelligence: requirements and possibilities', in *Conf. Phil. and Theory of AI*, pp. 117–135. Springer, (2017).
[2] Open Science Collaboration et al., 'Estimating the reproducibility of psychological science', *Science*, **349**(6251), aac4716, (2015).
[3] C. Drummond, 'Replicability is not reproducibility: nor is it good science', *Evaluation Methods for ML (ICML)*, (2009).
[4] S. E Embretson et al., *Item response theory*, Psychology Press, 2013.
[5] P. Flach, 'Performance evaluation in machine learning: The good, the bad, the ugly and the way forward', in *33rd AAAI*, (2019).
[6] V. Gewin, 'Data sharing: An open mind on open data', *Nature*, **529**(7584), 117–119, (2016).
[7] K. Grace, 'Algorithmic progress in six domains', Technical report, Machine Intelligence Research Institute, (2013).
[8] J. Hernández-Orallo, 'Evaluation in AI: from task-oriented to ability-oriented measurement', *AI Review*, **48**(3), 397–447, (2017).
[9] J. Hernández-Orallo, *The measure of all minds: evaluating natural and artificial intelligence*, Cambridge University Press, 2017.
[10] J. Hernández-Orallo et al., 'A new AI evaluation cosmos: Ready to play the game?', *AI Magazine*, **38**(3), 66–69, (2017).
[11] 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).
[12] F. Martínez-Plumed et al., 'Accounting for the neglected dimensions of AI progress', *arXiv:1806.00610*, (2018).
[13] F. Martínez-Plumed et al., 'IRT in AI: Analysing machine learning classifiers at the instance level', *Artificial Intelligence*, **271**, 18–42, (2019).
[14] Fernando Martínez-Plumed et al., 'Does AI qualify for the job? A bidirectional model mapping labour and ai intensities', in *AIES*, (2020).
[15] F. Martínez-Plumed et al., 'The facets of artificial intelligence: A framework to track the evolution of ai', in *IJCAI-18*.
[16] OECD, *Artificial Intelligence in Society*, OECD Publishing, 2019.
[17] J. Pineau, K. Sinha, G. Fried, R. N. Ke, and H. Larochelle, 'ICLR Reproducibility Challenge 2019', *ReScience C*, **5**(2), 5, (May 2019).