
BETWEEN PROGRESS AND POTENTIAL IMPACT OF AI: THE NEGLECTED DIMENSIONS

A PREPRINT

✉ **Fernando Martínez-Plumed***
VRAIN, Universitat Politècnica de València
fmartinez@dsic.upv.es

Shahar Avin
CSER, University of Cambridge
sa478@cam.ac.uk

Miles Brundage
OpenAI
miles.brundage@asu.edu

Allan Dafoe
DeepMind
allandafoe@deepmind.com

Seán Ó hÉigeartaigh
CSER, University of Cambridge
so348@cam.ac.uk

José Hernández-Orallo
VRAIN, Universitat Politècnica de València
jorallo@dsic.upv.es

July 5, 2022

ABSTRACT

We reframe the analysis of progress in AI by incorporating into an overall framework both the task performance of a system, and the time and resource costs incurred in the development and deployment of the system. These costs include: data, expert knowledge, human oversight, software resources, computing cycles, hardware and network facilities, and (what kind of) time. These costs are distributed over the life cycle of the system, and may place differing demands on different developers and users. The multidimensional performance and cost space we present can be collapsed to a single utility metric that measures the value of the system for different stakeholders. Even without a single utility function, AI advances can be generically assessed by whether they expand the Pareto surface. We label these types of costs as neglected dimensions of AI progress, and explore them using four case studies: Alpha* (Go, Chess, and other board games), ALE (Atari games), ImageNet (Image classification) and Virtual Personal Assistants (Siri, Alexa, Cortana, and Google Assistant). This broader model of progress in AI will lead to novel ways of estimating the potential societal use and impact of an AI system, and the establishment of milestones for future progress.

Keywords Artificial Intelligence · Evaluation · AI Costs · AI resources · AI progress

1 Introduction

The impact a new technology has on society depends on many factors and interests, but it is undeniably and ultimately linked to how powerful and versatile the technology is. This is a two-way association, as measuring the actual progress of a technology depends on how transformative it is for society. AI is no different. In this paper we argue that we need a more thorough account of the elements that (1) play a major role on how efficient it is to deploy a new AI technology or (2) imply some unaccounted costs on the AI life cycle or on society as a whole.

The prevailing approach to assessing AI progress consists of measuring *performance*, such as the raw or normalized score in a game, ELO rating, error rate, accuracy, and so forth. All these measures are often plotted over time to evaluate temporal progress Eckersley and Yomna [2017], Shoham et al. [2017]. Performance, however, does not exactly correspond with social value or scientific progress in AI. Misalignment between what is measured and what is desired can lead to misallocation of energy and resources. Specifically, excessive effort is likely to go towards achieving novel performance milestones Martínez-Plumed et al. [2021a], and insufficient effort towards progress on other dimensions relevant to social value, economic value, and scientific progress, such as compute efficiency, data efficiency, novelty, replicability, autonomy, and generality.

*Corresponding author

This does not mean that quantitative assessment through benchmarking should be abandoned Martínez-Plumed and Hernández-Orallo [2016], Martínez-Plumed et al. [2016], Martínez-Plumed and Hernández-Orallo [2018], Martínez-Plumed et al. [2019], Hernández-Orallo et al. [2022], Martínez-Plumed et al. [2022]. On the contrary, we need more and better measurement Hernández-Orallo [2017a], Hernández-Orallo et al. [2021]: measurement which is more comprehensive, general, and focused on the cost function of the ultimate beneficiaries. Ultimately we would like to weight all the resources that users (or receivers) of a technology require to achieve their goals. For instance, to what extent does progress on a particular metric of performance in machine translation map on to user’s satisfaction? Does the progress also correspond to a reduction in cost per translation, or in time for execution? If a paper develops a new technique, how easily can this be brought from the laboratory to a generally impactful application?

In general, users seek the benefits of high performance (at a set of tasks), while they seek to minimize the costs of developing and deploying their system. Sensitivity to costs is true for individual consumers, firms and developers, as well as other scientists. Some kinds of hidden costs can appear during development, when an application is produced, when reproduced at a large scale, or when adapted to other domains. Some future costs will be borne by future developers or scientists, sometimes referred to as “technical debt” or “research debt”. Other costs may be spread more broadly, and are thus harder to account for. As in other sectors, there are externalities from AI development and deployment which are important to be aware of; among the negative externalities are environmental footprints, user privacy, skill atrophy (e.g., the Google effect), opacity in decision making, etc. Attention to, and ideally measurement of the possible impact of these side effects is beneficial, as it is a first step towards internalizing them.

In this paper we consider this wide range of costs. We will identify how costs are distributed depending on the stage in which they are incurred, the number of times they are replicated, and the actor covering each cost. These dimensions should be integral to the measurement of AI progress, even if their measurement is not always straightforward. As an illustration of these difficulties and how they can be overcome, we will analyze several case studies where we evaluate performance alongside these other dimensions. As a result, we overhaul the notion of progress in these domains.

Our paper makes several contributions. First, we offer the most detailed and formal analysis to date of the dimensions of AI progress. While previous work has attempted to quantify progress in the performance of a specific system, we more fully account for the resources required and the generality of solutions. Second, in so doing we surface neglected dimensions of AI progress that may be optimized more directly. Third, we offer a novel framing under Pareto optimality for assessing performance and costs of an AI system, which suggests a more principled approach to forecasting the impact of future developments in AI, with myriad applications for policy, ethical, and economic analysis, and better research portfolio optimization within the field of AI itself.

2 Background

Many benchmarks and competitions are used in AI, but they vary in how representative they are of the fundamental problems in their respective subfields Hernández-Orallo [2017b], Hernández-Orallo et al. [2017]. As a reaction, challenges in AI are realigned to see if they can better capture the potential impact on automation Frey and Osborne [2017], Brynjolfsson and Mitchell [2017], Aghion et al. [2017], Korinek and Stiglitz [2017], or the aspiration of more human-like AI Lake et al. [2017], Marcus [2018]. A deeper concern is that most benchmarks are not really fostering the basic scientific advances needed to move the field forward, be they theoretical advances, explanatory insights, or tools to facilitate other work. This issue of *non-representativeness* is partly addressed through the review process, and requirements such as controlling the percentage of papers in different areas Shah et al. [2017].

A second issue, *specialization*, is related to representativeness. When a benchmark or competition becomes the target, researchers will have incentives to overly specialize their systems to performance on that benchmark, at the cost of other features of their system, such as generalizability. If we had a satisfactory metric of generality then we could use that as a benchmark measure, but it remains an open question how best to operationalize generality Hernández-Orallo [2017a], balancing between putting all the distribution mass possibly falling on a few tasks Legg and Hutter [2007]—and not really being general—or distributing it in a block-uniform way—facing the no free lunch theorems Wolpert [2012].

A third issue is *reproducibility*, and the wider notion of replicability. In AI this was usually understood as requiring the sharing of data and code, but the concept is becoming richer Drummond [2009], Bonsignorio and Del Pobil [2015], Henderson et al. [2017]. Indeed, we must distinguish between specifically reproducing the results, and replicating the findings with some variations Zwaan et al. [2017]. Several initiatives have been proposed to facilitate (or even require) a wider replicability. For instance, with the “open leaderboards” Spohrer [2017], participants have to upload their code so that other participants can make modifications and submit another proposal.

Finally, users are generally sensitive to the effort of developing and deploying an AI system, which performance benchmarks rarely take into account. Much AI progress is attributed to advances in computational power Reagen

et al. [2017], Hwang [2018]. However, it is not straightforward to quantify what exactly can be attributed to software progress, hardware progress or several other resources Brundage [2016], Grace [2017]. Accordingly, perhaps it is more effective to just measure the so-called “end-to-end performance”, including computational time and quality of the models, such as the recent *DAWN Bench* Coleman et al. [2017] for deep learning, or MLPerf MLPerf [2018] for a variety of AI models and (hardware) chips. Other resources, such as data, are at least as important, especially in machine learning². But it seems subjective to determine what input is seen positively or negatively, or even considered as cheating: too much data (versus better algorithms), too much knowledge (constraints, rules or bias), enriched input Bougie and Ichise [2017], etc. The question depends mostly on the cost of the resource. Human resources (“human computation”) are also common in AI to increase performance or generality (but at the cost of autonomy).

Overall, there are many resources involved but, at the moment, there is no integrated framework taking into account all of them. Related approaches involve the ideas of utility functions, Pareto-optimal analysis and, most especially, cost-sensitive learning Elkan [2001]. Turney (2002) identifies costs related to inputs and outputs in classification (errors, instability, attributes, labeling, actioning) data (cases), computation and human preprocessing. In this paper, we offer a general statement of this idea, applied to AI progress.

In the end, when assessing AI progress in a comprehensive way, one should consider the whole life cycle of research, innovation, production, and reproduction. Notions such as technical or research debt are becoming more recognized, as they incorporate some costs that are not perceived at early stages of the process but have an impact later on, when the technology or product is put into practice Sculley et al. [2015], Henderson et al. [2017], Olah and Carter [2017], Desislavov et al. [2021].

3 Components and integration

We now flesh out a comprehensive list of dimensions that are required for an “AI system” to work. We use the term “system” in a flexible way, including an agent, an algorithm, a product, etc., proposed in a research paper or by a company.

Given the fuzzy contours of AI Martínez-Plumed et al. [2018], one relevant way of assessing the impact of AI technology is through the potential for “automation” Frey and Osborne [2017], Brynjolfsson and Mitchell [2017], Aghion et al. [2017], Korinek and Stiglitz [2017], Martínez-Plumed et al. [2020a], Tolan et al. [2021]. However, some of these studies are usually assuming conditions such as “at a reasonable cost”, “to a high degree of automation”, etc., versus “full automation at whatever cost”. The estimated probability of automation for a given task might change completely depending on these conditions. In the end, automation is important, but it is the efficiency of the whole system what matters to assess its potential impact, including any “human computation” involved. This view of efficiency links us directly to the resources involved in an AI system.

Table 1 shows the resources we identified as frequently involved in developing and deploying AI systems. These resources have fuzzy boundaries and are often fungible with each other. For instance, the distinction between data and knowledge is not always clear, and hardware and software may be highly intertwined. Human resources are typically considered under “manipulation”, but can appear in other resources (e.g., labeled data and teaching a robot might be assigned to r_d and r_m respectively). Similarly, r_t represents calendar time, which cannot be accelerated by putting more human resources, as we have to wait for some events to happen—unless we use simulations or historical data from other domains. The existence of these fuzzy boundaries is not a problem, as long as all the resources are identified.

For some dimensions, we can find methods to evaluate their cost. For instance, software effort can be evaluated using analogy-based, WBS-based or size-based estimation models Putnam [1978], Sommerville [2015]. In the hardware category, some models consider both the equipment used for the development and deployment of the system to more complex hardware cost estimation methods, models and tools Ragan et al. [2002]. Similarly, compute can be estimated using mathematical (simulation) models such as in MathWorks [2018]. In some other cases, these ingredients can be grounded to economic terms Veryard [2014], or linked to the concept of “value proposition”, what a company or product actually delivers to its customers or society Anderson et al. [2006].

The ultimate criterion for identifying the resources is that they must incur costs during the development or deployment of an AI system. There are other dimensions that are not necessarily seen as increasing the overall costs, such as *fairness*, *privacy* and *transparency*. Because they have more to do with trust in AI or several ethical issues, they are not included in Table 1. Also, they are less neglected nowadays than they were a few years ago Friedler and Wilson [2018], fat [2019], Fernando et al. [2021]. In any case, separately or jointly with those in the table, fairness, privacy and transparency could be considered as well when analyzing some particular technologies (especially machine learning), as they can have an impact on their applicability, if some constraints or regulations are not met, or must be traded off

²See <https://sites.google.com/site/dataefficientml/bibliography> for a bibliography on data-efficient ML.

	Description	Example
r_d	Data: All kinds of data (unsupervised, supervised, queries, measurements).	A self-driving car needs online traffic information.
r_k	Knowledge: Rules, constraints, bias, utility functions, etc., that are required.	A spam filter requires the cost matrix from the user.
r_s	Software: Main algorithm, associated libraries, operating system, etc.	A planner uses a SAT solver over a complex ecosystem of libraries.
r_h	Hardware: Computer hardware, sensors, actuators, motors, batteries, etc.	A drone may need a 3D radar for operation, instead of a camera.
r_m	Manipulation: Manual (human-operated) intervention through assistance	A robot needs to be manually re-calibrated or overseen real-time.
r_c	Computation: Computational resources (CPU, GPU usage) of all the components	A (vanilla) nearest neighbor classifier computes all distances.
r_n	Network: Communication resources (Internet, swarm synchronisation, distribution).	A delivery system needs online connectivity for all drones.
r_t	Time: Calendar (physical) time needed: waiting/night times, iteration cycles.	A digital assistant requires cyclical data (weeks) to find patterns.
r_l	Load: Volume, size or dimension of the solution (length of the parameter vector in a DNN, model size in units of bytes, memory usage, etc.)	A specific DNN (GoogLeNet) trained on CIFAR10 using 8 layers requires 7M parameters and 40MB of storage.
r_e	Energy: Power consumption per unit of time required to build or operate.	A personal assistant (PA) has a peak power consumption of 2.20W when keyword spotting and 0.4W when idle.

Table 1: Resources that are frequently needed by AI systems.

with performance or the other dimensions in Table 1. In general, as we will see in the following sections, for a particular new innovation or technology, only a subset of dimensions may be relevant.

It is appealing to collapse several of these dimensions for an AI system to a single metric. For any given user with rational (transitive and complete) preferences, their preferences can be represented using a utility function. A firm’s utility function, for example, might correspond to risk-adjusted expected profit. A user’s utility function might be harder to quantify, but is generically increasing in the performance of the system and decreasing in the costs of the system. Denote a performance vector, ψ , for a given problem, which is often a unidimensional quantitative score (such as the error), but could also have several components. A utility function maps performance and all associated resources to a single dimension:

$$U(\psi, \bar{r}) = U(\psi, r_d, r_k, r_s, r_h, r_m, r_c, r_n, r_t, r_l, r_e) \rightarrow u \quad (1)$$

In some cases this is an additively separable function, such that $U(\psi, \bar{r}) = B(\psi) - \sum_x C_x(r_x)$, with the first term accounting for the benefit according to the performance of the system minus the costs produced by the use of resources (note that the cost functions C_x are different for each resource). For economic applications, we might be able to separate the utility function into performance generating revenue (in dollars), and resources imposing costs (in dollars).

In many cases, we are not able to collapse performance and costs into a single metric, perhaps because the utility function is not known or varies across a population of users. Still, we can productively examine the relative performance and costs of different systems. For any number of dimensions, we can assess the Pareto-optimal surface. For example, Fig. 1 shows algorithms and architectures according to their MNIST prediction error and power consumption, revealing that most solutions are not on the Pareto surface on these dimensions, with notable exceptions, such as some ASIC architectures, which focus on efficiency in terms of chip space, speed and “energy footprint” Chen et al. [2014].

4 The full range of accounting

The benefits and costs of developing and deploying an AI system are not incurred only once, but throughout the many uses, reuses, and follow-on contributions. Some costs are borne exclusively during the initial conception and development, while others recur with each adaptation to a new application, or even each application to a particular user. In general, the total resource burden should be accounted for according to the whole cycle of the AI system.

Fig. 2 shows how the dimensions we identified can become relevant at different stages of the life cycle of an AI system. Consider we want to assess the potential impact of a new algorithm for voice recognition. Apart from all the human

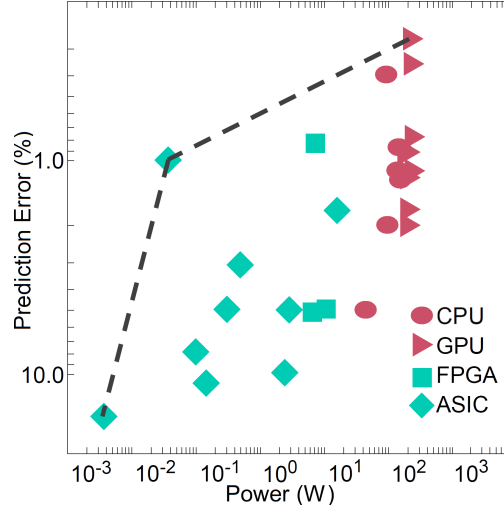


Figure 1: Performance for MNIST LeCun et al. [1998], for 22 papers, compared to power consumption (data from Reagen et al. [2017]). The Pareto frontier is also shown (we will later discuss whether the points can actually be joined by straight segments).

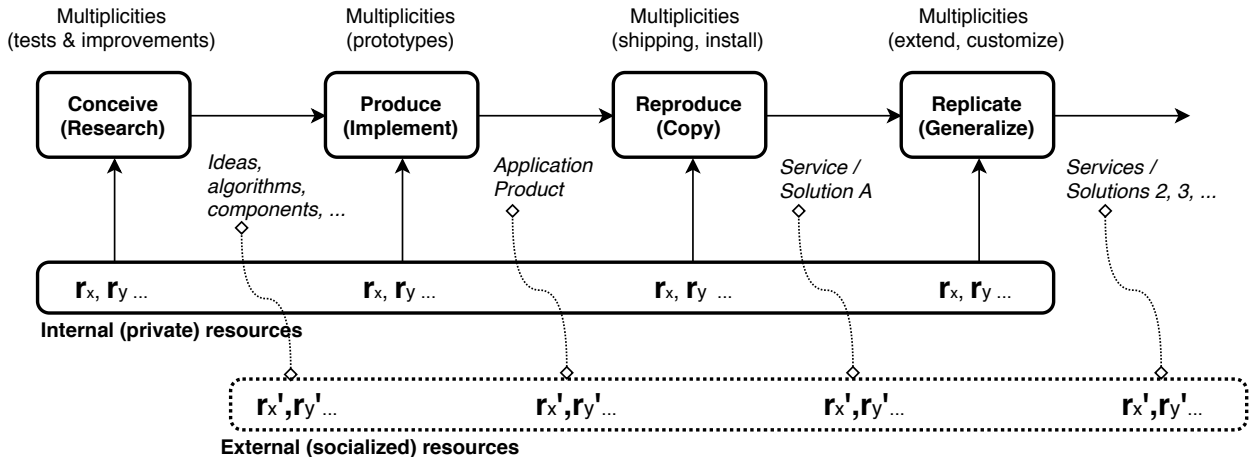


Figure 2: Illustrative representation of stages of the AI system life cycle where resources might be required.

thinking, there will be a great effort in terms of failed experiments, different libraries used, users testing the early systems, etc. If a company takes these ideas and builds a prototype, the tests, software, hardware, and compute will concentrate on production. When the system is reproduced (installed or shipped) to users, additional resource costs will be incurred. Further, if the idea can be adapted for other applications (e.g., adapting a voice recognition system to other languages), depending on its generality and reproducibility, the initial contribution can provide further value, at some further adaptation cost including the need for new corpora, training, semantic knowledge, etc.

At each stage of the life cycle, the contribution may be deployed a multiplicity of times (represented above the boxes in Fig. 2). The total value of the contribution thus needs to take into account the scale of its deployment. For instance, some early speech recognition systems were pre-trained once (the *system cost*, denoted by C , covering the “conceive” and “produce” stages in Fig. 2) and then adapted to thousands of users, with extra hours of customization per user (the *application cost*, denoted by C^j with j indexing each of the n applications, or users, covering the “reproduce” and “replicate” stages). More recent general speech recognition systems do not need such customization. Consequently, the application cost C^j is lower per user. In both cases, the total cost C is $C + \sum_{j=1}^n C^j$. As the number of applications increases, the average cost will converge to the average application cost as the system cost is amortized. For this reason, for contributions that have many possible applications, it is worth paying additional system costs so as to make the contribution more general, adaptable, and reusable, and thereby bring down the application costs. Since AI often has

broad potential applicability, contributions that are general, adaptable, and reusable are likely to have high utility, so having significant economic and social impact.

Fig. 2 not only covers direct “internal” costs (r_x, r_y, \dots) but also some external “debts” or “societal” costs (r'_x, r'_y, \dots). For instance, automated customer service systems (call centers) clearly were not a Pareto improvement relative to previous systems, even though they may be a profit maximizing improvement or can represent a baseline of automation for further improvement that is finally assumed as the standard service. In the end, companies reduce their labor costs for customer service by substituting in phone-trees and voice recognition, but in the process impose time, frustration, and other costs onto the customer. Some navigators and personal assistants can make users more dependent on them, atrophying some capabilities or leading to a simplification of language. In other words, the user adapts to the AI system, and assumes part of the effort or cost. In general, technological innovation both involves developing technology to fit a given conception of the task, and adapting conceptions of the task to fit the capabilities of technology Martínez-Plumed et al. [2021b, 2020b]. In the process of adapting work processes, customer expectations, relationship norms, and even urban design to what is technologically convenient, there can be consequences for society that are not internalized by the designers and deployers of these systems. This footprint of AI is not usually acknowledged in benchmarking, and can have more societal impact than the technology itself.

From the previous sections, we conclude that the contribution of an AI development should, in principle, be given a full accounting of the costs and benefits, across the contribution’s full life cycle. The current emphasis on targeting and reporting performance benchmarks, however, poses an obstacle to a full accounting. Reproducibility and replicability are two traditional tools for addressing this. More precisely:

- *Specific reproducibility* refers to whether the *same result* can be obtained from the same conditions and procedures. In AI, this requires that all the necessary code and data are given. This also assumes the same cost functions as well: $\sum_{j=1}^n \sum_x \mathbf{C}_x^j(r_x^j) = n \sum_x \mathbf{C}_x(r_x)$.
- *General replicability* will check whether the AI technique can be *applied to other problems*, a set of n tasks, applications, or users indexed by j , with an overall cost $\sum_{j=1}^n \sum_x \mathbf{C}_x^j(r_x^j)$ that must consider the adaptation effort, with different resources r_x^j and cost functions \mathbf{C}_x^j per user.

Especially for replicability, we can experiment with different hardware architectures, change some of the software and get different computational costs, apart from different performance. That means that the partial results for each \mathcal{B}^j and $\mathbf{C}_x^j(r_x^j)$ might be different, but we still have something replicable with similar utility. A clear example of this notion of replicability is “approximate computation” in deep learning, where one can get much smaller computational costs without a significant change in accuracy Reagen et al. [2017].

5 Exploring the Pareto-frontier of AI research

Corporations, governments, startups, NGOs, personal users, and contemporary and future AI researchers are the intended recipients, or *receivers*, of the AI technologies being developed, and they each have different preferences, resources and constraints, or in other words different operating characteristics Martínez-Plumed et al. [2021a]. The familiar concept of the ROC curve plots true positive rates (TPR) and false positive rates (FPR) for binary classifiers, and emphasizes the importance of comparing multi-dimensional surfaces, rather than single metrics.

For instance, Fig. 3 (left) just shows a single metric, performance, as a function of time. This plot does not explain what the cluster of attempts after 2014 really contribute, when they have more error than the already obtained human level. Other dimensions are neglected in this plot, limiting insight about progress.

Before analyzing the case studies, we have to understand how to build and work with the Pareto frontier. When resources are included, the analysis of optimal Pareto surfaces might be slightly different than the traditional triangulation approach. When showing performance metrics such as TPR and FPR for two models, any point in between can be obtained by interpolation, connecting any two points by a straight segment. However, we should note that these points require the implementation of both models. While some of the resources can be interpolated, others (e.g., software) will simply sum up, and the points between two points will not be achievable with a straight line, but by an axis-parallel route.

For instance, Fig. 3 (right) shows performance against one hypothetical resource. For each method, A, B, C, D, and E, the numbers represent the extremes when varying their parameters. E1 represents a random (or baseline) model. Assuming interpolation is possible by changing the parameters of a method but not possible between different methods, the Pareto surface here is shown in blue. Method C can be discarded (as it is covered by A), but method B could also be

discarded, as its region is always dominated on the two dimensions by other methods (the dashed Pareto frontier), even if there are unreachable regions in between.

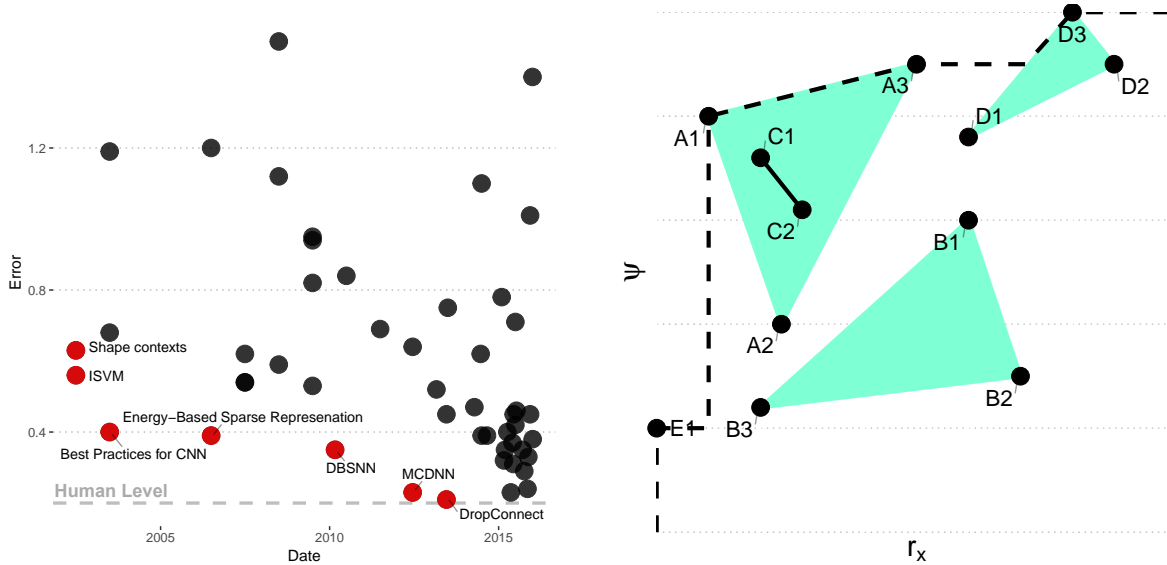


Figure 3: Left: Performance for the MNIST benchmark (data from EFF). Best state-of-the-art results in red. Right: A schematic representation of techniques A, B, C, D, E, with variants, the areas they cover, and the Pareto frontier (dashed black line).

The diversity of receivers —with different subjective utilities— and the number of dimensions suggest that a single utility metric is simplistic. This operating condition translates into a vector, or gradient, in the multidimensional space. For example, large technology corporations may gain significant utility from a discovery that allows modest speed-ups in exchange for significantly increased compute demands, whereas individual researchers, personal users and startups may find little value in such a discovery. Conversely, the existence of real recipients whose preferences can be known in advance allows us to prioritize exploration of those configurations. From the above, we derive a few criteria to identify progress events, where all contributions below the Pareto frontier should not be necessarily discarded:

- *Improving the Pareto frontier for a known group of recipients* (A1, A3 or D3 in Fig. 3, right). This would include all-else-being-equal improvements in performance, but also reductions in computation, data, manipulation or other resources in Table 1. This would not, however, consider extreme regions no recipient assigns value to.
- *Covering a location slightly under the Pareto frontier with more flexibility* (B3 in Fig. 3, right). Instead of reaching some areas by combining existing approaches, a new technique can reach there easily with a trade-off between its own parameters, allowing more receivers to easily find their subjectively optimal trade-offs.
- *Covering a location slightly under the Pareto frontier with more diversity* (C in Fig. 3, right, if it is very different from A). The current dominant technique or paradigm can push the Pareto frontier for some time, but slightly suboptimal approaches, especially if they are radically different (i.e., alternative “research programs”), should not be discarded because they may lead to potential improvement in the frontier if the current paradigm stalls.

Receivers can be incentivized to generate and communicate their gradients (though in some cases, countervailing considerations may exist such as commercial secrecy). It is also in the interests of discoverers to show the recipients benefited by their discovery. Brokers of such information (peer-review, surveys, competitions, etc.) are in a position to meet the incentives (and gradients) of both researchers and recipients by ensuring such discoveries are properly rewarded.

6 Case studies

In this section we will examine a number of representative case studies of progress in AI: Alpha*, ALE, imageNet, personal assistants and some others.

6.1 Alpha*

Alpha* refers to a series of papers and associated techniques by DeepMind to play board games. We analyzed the whole series, from AlphaGo Silver et al. [2016] (including the Fan and Lee versions, used against Fan Hui and Lee Sedol, respectively, and its latest version, AlphaGo Master, which won 60 straight online games against professional Go players), AlphaGo Zero Silver et al. [2017a] (a version created without using data from human games) and AlphaZero Silver et al. [2017b] (which uses an approach similar to AlphaGo Zero to master not just Go, but also chess and shogi).

	<i>AlphaGo_{Fan}</i>	<i>AlphaGo_{Lee}</i>	<i>AlphaGo_{Master}</i>	<i>AlphaGo_{Zero}</i>	<i>AlphaZero</i>
r_d (Data)	✓	✓	✓	✓	✓
r_k (Knowledge)	○	○	○	○	○
r_s (Software)	○	×	×	○	×
r_h (Hardware)	×	×	×	×	×
r_m (Manipulation)	✓	✓	✓	✓	✓
r_c (Computation)	✓	○	○	✓	○
r_n (Network)	—	—	—	—	—
r_t (Time)	—	—	—	—	—
r_l (Load)	×	×	×	×	×
r_e (Energy)	×	×	×	×	×
ψ (Performance)	✓	✓	✓	✓	○

Table 2: Dimensions (resources and performance) reported in the Alpha* papers. Systems from Silver et al. [2016, 2017a,b]

Table 2 shows whether the dimensions were reported in the papers (✓), reported in different sources (possibly from different authors) (✓), only partially accounted for (○), not mentioned but relevant (×) and not applicable (—). Many dimensions are relevant for the analysis: the data, the knowledge, the software, the hardware, manipulation, computation and, of course, performance, etc. However, only some of them are provided, which makes a comprehensive comparison of the whole space difficult. Still, we will represent three dimensions: performance (in ELO ranking, which can only be partially estimated for AlphaZero), computational resources (using the equivalence: $1\ TPU_{v2} \simeq 3\ TPU_{v1} \simeq 36\ GPU \simeq 180\ CPU$ Jouppi et al. [2017]) and human manipulation resources (as represented quantitatively by the ELO ranking of the player or players the system learns from)³. Other dimensions (like knowledge⁴ about Go, software, etc.) are not included because of insufficient information from some papers.

What we see in Fig. 4 is that the Pareto frontier at the moment is represented by AlphaGo Master and AlphaGo Zero. AlphaGo Fan and AlphaGo Lee are discarded because AlphaGo Zero needs less compute⁵, requires no manipulation and gets better performance. Why is AlphaZero seen as a breakthrough if it is not Pareto optimal? The answer is generality. AlphaGo* only solved one task (Go) and AlphaZero can solve several tasks. Note that the computation times shown in Fig. 4 include both training and deployment (system and application costs). Hence, a model that is half way between models A and B (choosing between them with equal probability), denoted by \overline{AB} , has performance $\psi(\overline{AB}) = 0.5\psi(A) + 0.5\psi(B)$, but has a computational cost of $r_c(\overline{AB}) = r_c(A) + r_c(B)$. This is why the Pareto frontier in Fig. 4 has parallel segments, as in Fig. 3 (right). Finally, if we look chronologically at the plot, we see that the main gradient that has been followed has been performance.

6.2 ALE

The second case study is ALE Bellemare et al. [2013], a collection of Atari games that has become popular for the evaluation of general-purpose RL algorithms learning from screen shots. We selected all the papers (systems) from EFF’s AI Progress Measurement Project Eckersley and Yomna [2017] and the papers introducing Rainbow Hessel et al. [2017a] and REACTOR Gruslys et al. [2017]. Table 3 shows what information we found about the resources and performance⁶. Again, many dimensions are relevant, but only a few are systematically reported: data, computation and performance. Fig. 4 represents computation and performance. Computation time (whenever the authors do not provide this information explicitly) is roughly estimated from the kind of approach used, whether it is follow-up work,

³Complete information regarding compute can be found in Table 6 in the supplementary material.

⁴We have the constructed features: stones to be captured or escaped, legal moves, ‘liberties’, etc. While this knowledge is crucial, there is no cost for a new match (reproduction), but the adaptation of AlphaZero to other games (replication) may be important.

⁵The compute used for generating the training data, i.e. for the self-play games, has not been considered as it is unclear from some of the Alpha* papers (only AlphaZero makes it explicit).

⁶Complete information regarding compute can be found in Table 7 in the supplementary material.

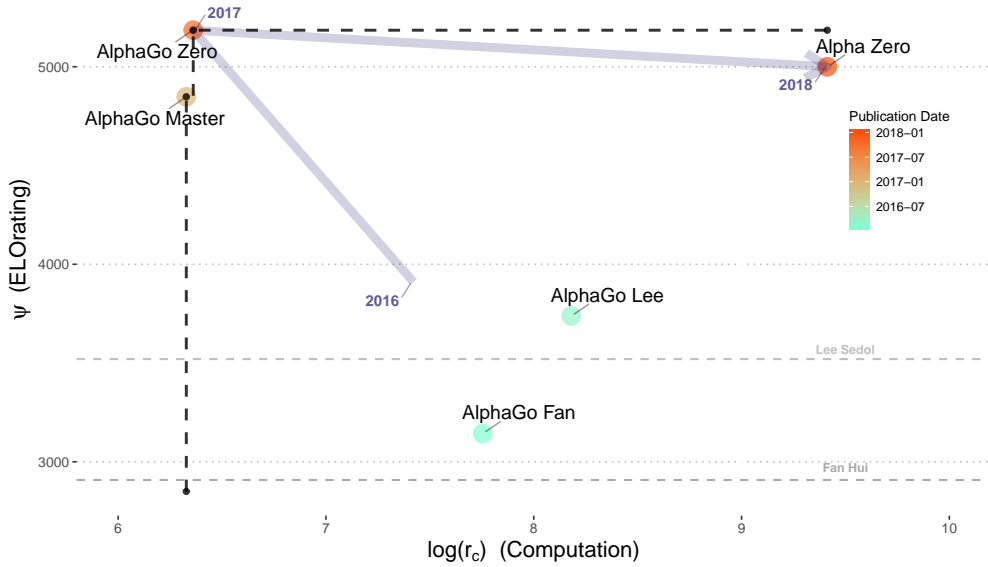


Figure 4: Multidimensional utility space for Alpha*. Research gradient evolution from 2015 to 2018 represented with a segmented gray arrow. The Pareto frontier (dashed black) does not include other resources (software, and humans used for training) that duplicate for connecting segments.

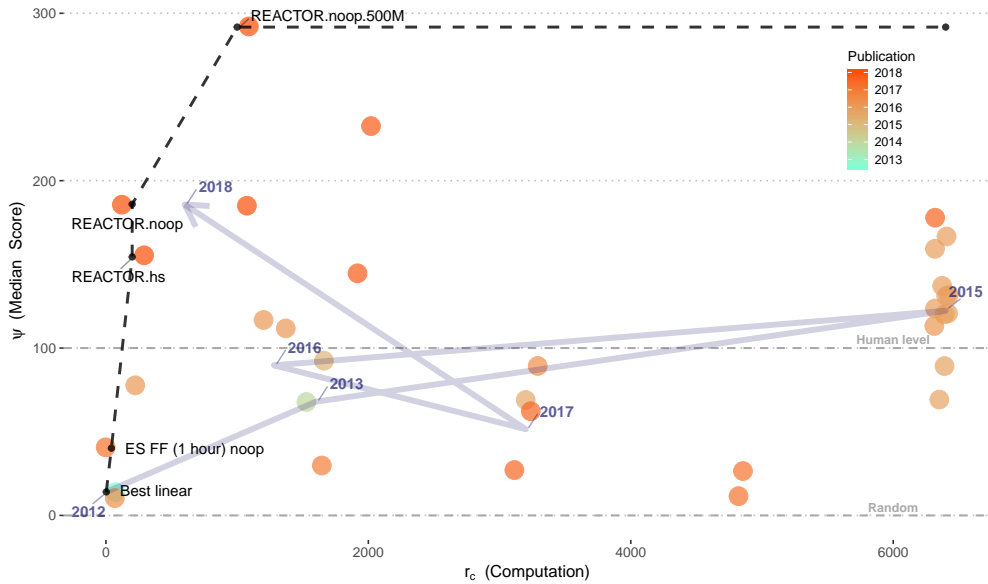


Figure 5: Multidimensional utility space for ALE. Research gradient evolution from 2013 to 2018 represented with a segmented gray arrow. The Pareto frontier (dashed black) does not include other resources (software, and humans used for training) that duplicate for connecting segments.

the training setting used, etc., or from figures in more recent papers making explicit comparisons between them and the state of the art Hessel et al. [2017a], Gruslys et al. [2017].

What we see in Fig. 5 is a current Pareto frontier dominated by REACTOR variants, ES FF and Best Linear. In this case, the computation times in Fig. 5 includes just training time. If we select a model \overline{AB} that is half way between two models A and B (choosing between them with equal probability), we may have A train and play for half of the ALE games and B train and play for the rest. As we average for the whole set of games, we can actually have $r_c(\overline{AB}) = 0.5r_c(A) + 0.5r_c(B)$, at least if there is no transfer effort between games. This is why the Pareto frontier is

	<i>Sarsa</i>	<i>Best Linear</i>	<i>DQN Best</i>	<i>NatureDQN</i>	<i>Gorila</i>	<i>DQN_{noop} & hs</i>	<i>DUEL_{noop} & hs</i>	<i>DDQN_{tuned} hs</i>	<i>PRIOR_{noop} & hs</i>	<i>P.DUEL_{noop} & hs</i>	<i>AC₃LSTM, FF, FF1d</i>	<i>DDQN_{Pop-Art} noop</i>	<i>AC₃CTS</i>	<i>Sarsae & f-EB</i>	<i>TRPO_{hash}</i>	<i>DQN_{CTS} & PixelCNN</i>	<i>C51_{noop}</i>	<i>ES FF_(1h) noop</i>	<i>RAINBOW</i>	<i>REACTOR</i>
r_d (Data)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
r_k (Knowledge)	○	○	×	✓	×	○	×	○	○	○	○	○	×	×	○	○	○	×	✓	✓
r_s (Software)	×	×	×	✓	×	×	×	×	×	×	×	×	×	×	×	×	×	×	✓	×
r_h (Hardware)	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
r_m (Manipulation)	×	✓	×	✓	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
r_c (Computation)	○	○	○	○	○	○	○	○	○	○	✓	○	○	○	○	○	○	○	✓	✓
r_n (Network)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
r_t (Time)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
r_l (Load)	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
r_e (Energy)	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
ψ (Performance)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 3: Dimensions (resources and performance) reported in the ALE papers (from EFF Eckersley and Yomna [2017] and Gruslys et al. [2017], Hessel et al. [2017a])

shown with direct straight segments. Regarding the research gradient (in gray), it has changed over the years, with some disregard of compute initially and more concern in efficiency recently.

For this benchmark, it is common to find “learning curves” in the papers (e.g., Machado et al. [2017]), which show performance varying on the number of episodes. This is clearly the r_d (data) but it also influences directly on computation. These learning curves give information of full regions of the multidimensional space, as we saw in Fig. 2.

For some papers, the comparison was not possible (e.g., due to different subsets of games). It is important to note, however, that some approaches based on genetic programming Kelly and Heywood [2017] and on planning Bandres et al. [2018] are valuable in terms of diversity.

6.3 ImageNet

The third case study is the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) Russakovsky et al. [2015]. Specifically, we focused on the ImageNet classification challenge, a multi-class classification problem framework with about 1.2 million images for training (having 1,000 leaf-node categories in the Imagenet hierarchy), 50,000 for validation and 100,000 images for testing. Each image is associated with one ground truth category, and performance is usually reported as: the top-1 error rate (comparing the ground truth against the first predicted class), and the top-5 error rate (comparing the ground truth against the first 5 predicted classes). Since the breakthrough in 2012 achieved by the first Deep Neural Network (DNN) system AlexNet Krizhevsky et al. [2012], several other DNNs with increasing complexity have been submitted to the challenge in order to achieve better performance. We analyzed the following DNNs which obtained the highest performance in these past six years, including those more recent approaches developed for environments with a small computational budget (e.g., mobile devices): AlexNet Krizhevsky et al. [2012], BN-AlexNet Zagoruyko [2016], BN-NiN Lin et al. [2013], ENet Paszke et al. [2016], GoogLeNet Szegedy et al. [2015], VGG Simonyan and Zisserman [2014], ResNet He et al. [2016], Inception-v3 Szegedy et al. [2016], Inception-v4 Szegedy et al. [2017], Shufflenet Zhang et al., Mobilenet-v1 Howard et al. [2017], Mobilenet-v2 Sandler et al. [2018], Xception Chollet [2017], Densenet Huang et al. [2017], Squeezenet Iandola et al. [2016], fd-MobileNet Qin et al. [2018], AmoebaNet Real et al. [2018], SENet Hu et al. [2018], Shufflenet v2 Ma et al. [2018], GPipe Huang et al. [2018] and PolyNet Zhang et al. [2017a].

Table 4 shows the information we found about the resources and performance⁷. In this case, many dimensions are relevant and, although some of them are often reported (data, computation, performance, volume), the different hardware/software used as well as the variations in data and training/evaluation techniques precludes a direct comparison of resource utilisation between systems. Furthermore, there is a marked inconsistency in the figures found in different

⁷Complete information regarding compute, volume and power consumption can be found in Table 8 in the supplementary material.

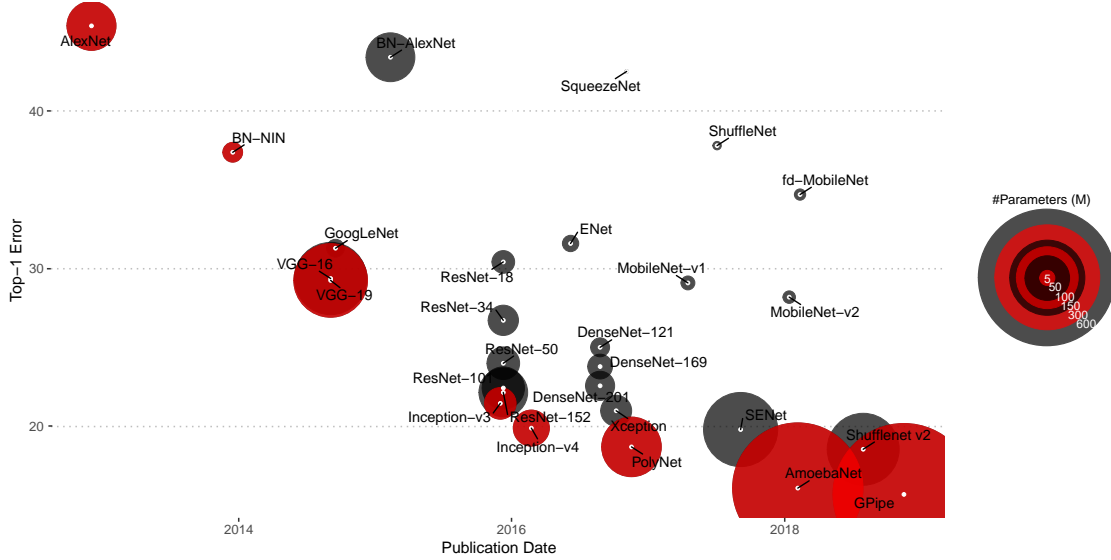


Figure 6: Performance (Top-1 Error) and number of learnable parameters in each DNN for the ImageNet benchmark. Systems are plotted against their publication date.

paper. For instance, for the computation dimension, authors provide either training time, multiply-accumulate (MAC) operations or FLOPs in a single forward pass for input image, etc., which are therefore not easily comparable between systems. Still, we can compare some independent dimensions such as *load* (in terms of learnable parameters in each DNN) and performance as a function of time (see Figure 6). Again, this plot does not entirely explain what the cluster of attempts after 2017 really contributes to the progress in this domain, when they have more error than the already best result. From the literature we can extract that, in the last few years, authors have increasingly focused on developing highly efficient and accurate networks for very limited computational budgets, which explain why systems such as ENet, MobileNet or fd-MobileNet have reduced their size (compact networks with a significant smaller number of parameters). However, other interesting dimensions are neglected in this plot, limiting insight about progress in this benchmark.

In this regard, and given the relevance of the challenge and the wide variety of systems approaching it, some authors have also tried to provide and compare some figures about the quality of different networks in more controlled (and thus comparable) environments in terms of model sizes Real et al. [2018], number of operations (see Real et al. [2018], Huang et al. [2018]) or evaluation procedures Zhang et al. [2017a]. A much more comprehensive analysis in terms of computational requirements and performance can be found in Canziani et al. [2016]. From the latter, we can obtain metrics related to memory footprint, number of operations and power consumption for a number of systems that can then be used to compare resource utilisation. This can be seen in Fig. 7, which is a much more insightful and comprehensive plot showing that the current Pareto frontier in imagenet is currently dominated by Inception variants, Densenet and those more efficient networks such as MobileNet variants. For the same reason indicated in the previous section for Fig. 5, the Pareto frontier is shown with direct straight segments. The research gradient (in gray) has changed over the years, with some disregard of compute initially and more concern in efficiency recently. Note that not all the papers in Fig. 6 appear in Fig. 7. For some of them the comparison was neither possible nor acceptable (e.g., mainly due to lack of information but also due to the use of different testing procedures, hardware, software, etc.)

6.4 Intelligent Personal Assistants

Another case study we analyze is AI-powered intelligent personal assistants (PA). We focus on a few big players: Siri Apple [2019], Alexa Amazon [2019a,b], Cortana Microsoft [2019] and Google Assistant Google [2019]. PAs are mainly based on conversational AI Cassell et al. [2000], natural language processing and knowledge-base systems. They may not represent a ‘leap’ in particular technologies, but an important progression in the integration of the current state-of-the-art techniques to power and improve sophisticated apps and services in terms of latency, automatic speech recognition accuracy, question answering, UI/UX, etc. Within the past few years, all these PAs have been incorporated into a myriad of new physical devices.

However, rarely, if ever, is there evidence regarding how all these PAs have been trained, tested and developed, how knowledge, conversational rules or utterance matching slots are defined, compiled and updated, how data is acquired

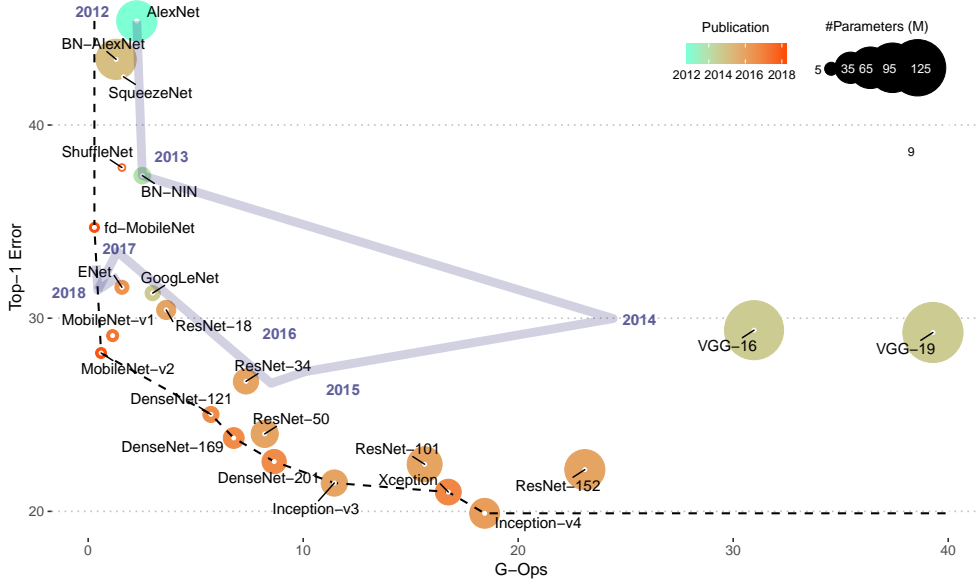


Figure 7: Multidimensional utility spaces for ImageNet. Top-1 Error vs. operations and size of the architectures in terms of millions of parameters. Research gradient evolution from 2012 to 2018 represented with a segmented gray arrow. The Pareto frontiers (dashed black) do not include other resources (software, hardware, etc.) that are duplicated for connecting segments.

	<i>AlexNet</i>	<i>BN - AlexNet</i>	<i>SqueezeNet</i>	<i>ShuffleNet</i>	<i>BN - NIN</i>	<i>fd - MobileNet</i>	<i>ENet</i>	<i>GoogLeNet</i>	<i>ResNet</i>	<i>VGG</i>	<i>MobileNet_{v1}</i>	<i>MobileNet_{v2}</i>	<i>DenseNet</i>	<i>Inception_{v3}</i>	<i>Inception_{v4}</i>	<i>Xception</i>	<i>AmoebaNet</i>	<i>SENet</i>	<i>ShuffleNet_{v2}</i>	<i>GPipe</i>	<i>PolyNet</i>	
r_d (Data)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
r_k (Knowledge)	✓	✓	✓	✓	○	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
r_s (Software)	-	-	✓	✓	-	✓	✓	○	✓	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	✓
r_h (Hardware)	✓	-	-	○	-	○	○	○	✓	-	○	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
r_m (Manipulation)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
r_c (Computation)	○	-	-	✓	-	✓	✓	○	○	✓	✓	✓	✓	✓	-	✓	○	✓	✓	✓	✓	○
r_n (Network)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
r_t (Time)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
r_l (Load)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
r_e (Energy)	✓	✓	-	-	✓	-	✓	✓	✓	✓	-	-	-	✓	✓	-	-	-	-	-	-	-
ψ (Performance)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 4: Dimensions (resources and performance) reported in ImageNet papers (from Krizhevsky et al. [2012], Zagoruyko [2016], Lin et al. [2013], Paszke et al. [2016], Szegedy et al. [2015], Simonyan and Zisserman [2014], He et al. [2016], Szegedy et al. [2016, 2017], Zhang et al., Howard et al. [2017], Sandler et al. [2018], Chollet [2017], Huang et al. [2017], Iandola et al. [2016], Qin et al. [2018], Real et al. [2018], Hu et al. [2018], Ma et al. [2018], Huang et al. [2018], Zhang et al. [2017a]). Note that ✓ indicates that a dimension may be reported in sources different from the original (possibly from different authors).

(from users), which models are trained to understand natural language, reason and interact with human beings, etc. Apart from this, the physical/cloud infrastructure needed is also presumably high with respect to GPU/CPU-based hardware for building the models, data collection/storage/manipulation, testing and scalability tools, architectural and design choices/models, etc.

This is all about internalities, but there are also a number of (neglected) externalities (“societal” costs) in terms of privacy (e.g., PAs usually collect information about the services that are used (and how and when they are used), or the impact of weariness and distrust towards a sometimes unpolished technology (e.g., misunderstood phrases, incorrect

answers and other mistakes, challenging configurations, etc). In this regard, unlike the internalities, there is a number of studies analysing different psychological aspects such as elderly engagement Reis et al. [2017], user experience Jiang et al. [2015], cognitive workload (mental effort) in voice-based interactions Strayer et al. [2017] or privacy Pellungrini et al. [2017], Zhang et al. [2017b] and ethical implications Hoy [2018], Manikonda et al. [2018].



Figure 8: Personal assistants performance over the years 2017 and 2018 Enge [2018]. While *Answered* refer to the percentage of questions in which the PA has attempted to answer obtaining a correct, partially correct or a wrong response, *Correct* refers to the percentage of the questions attempted answered fully correct.

Focusing on the more neglected internalities, we find that many dimensions are relevant for the analysis: data, knowledge, software, hardware, manipulation, computation, time, load, energy and, of course, performance. However, as we can see in Table 5, none of them are directly provided through their documentation (user guides, service manual, datasheets, websites, etc.) and, when this is the case (e.g., vendor-reported transcription errors Protalinski [2017]), these figures are not reliably reported in the literature, or cannot be compared due to companies not following the same standards of evaluation. This means having to make do with external documentation such as reviews, analyses and studies from agencies and other media outlets.

Accordingly, using several external sources we have been able to obtain information regarding energy requirements and consumption for personal assistants in some scenarios Lloyd [2018], Williams [2018] (Figure 9), as well as some performance values from a set of ~5,000 questions asked to each PA Enge [2018] (Figure 8). For each question, the authors checked whether the PA answered (i.e., the PA thinks that it understands the question, and makes an overt effort to provide a correct, a partial correct or an incorrect response to what the user asked for), whether the PA provides a direct and full correct answer to the question answered, whether the answer was wrong, and whether the answers were sourced from a database or a third-party source (e.g., Wikipedia).

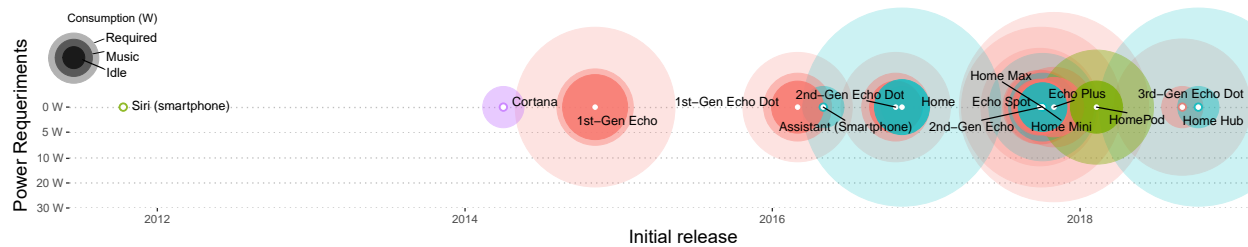


Figure 9: Personal assistant power consumption (minimum power supply requirement, idle and reproducing music).

What we see in Figures 8 and 9 is that, given the incomplete character of the information gathered, we can only represent the evolution of performance of those different devices/IPAs from 2017 to 2018, as well as the power

consumption required for some of them (although the latter is not very illustrative as the energy requirements of these devices are minimal). The most immediate result in terms of performance is that Microsoft’s Cortana outperforms Google on Google Home, where the "Home" version of Google Assistant is not as ‘smart’ as on mobile devices. Google Assistant for smartphones can be found in 2018 (no data in 2017) the most accurate PA, attempting to answer almost 80% of the questions presented with over 90% in accuracy (correct responses obtained). Cortana also surpasses Alexa and Apple’s Siri by a significant margin.

Regarding the temporal evolution, if we look chronologically at the plot in terms of performance, we see that Amazon Alexa is growing faster than any other PA: while it attempted to answer only 20% of questions in 2017, in 2018 Alexa attempted to answer over 50%. Finally, with respect to energy consumption, although there is an obvious difference of how much electricity various PA devices pull when standby compared to playing music, the average cost per month will not really make a dent in the electricity bill (with costs being under a dollar per month Lloyd [2018]). However, as they have millions of users, the global impact may be less negligible, especially if we include the consumption on the server side, which is rarely disclosed by the PA companies.

From these plots (and data), we cannot extract a clear Pareto frontier for any dimensions, or the research gradient over the years, making it difficult (if not impossible) to assess the different contributions or their economic and social impact. Having into account that these systems are thought to be ubiquitous in the future, this at least worrying, and more effort (or regulations) should be done so that some of the dimensions could be analysed and compared by users, governments and regulators.

	<i>Amazon Echo Dot</i>	<i>Amazon Echo</i>	<i>Amazon Echo Plus</i>	<i>Amazon Echo Spot</i>	<i>Google Home</i>	<i>Google Home Mini</i>	<i>Google Home Max</i>	<i>Google Home Hub</i>	<i>Apple HomePod</i>	<i>Apple Siri</i>	<i>Google Assistant</i>	<i>Microsoft Cortana</i>
r_d (Data)	×	×	×	×	×	×	×	×	×	×	×	×
r_k (Knowledge)	×	×	×	×	×	×	×	×	×	×	×	×
r_s (Software)	×	×	×	×	×	×	×	×	×	×	×	×
r_h (Hardware)	×	×	×	×	×	×	×	×	×	×	×	×
r_m (Manipulation)	×	×	×	×	×	×	×	×	×	×	×	×
r_c (Computation)	×	×	×	×	×	×	×	×	×	×	×	×
r_n (Network)	–	–	–	–	–	–	–	–	–	–	–	–
r_t (Time)	×	×	×	×	×	×	×	×	×	×	×	×
r_l (Load)	×	×	×	×	×	×	×	×	×	×	×	×
r_e (Energy)	○	○	○	○	○	○	○	○	✓	×	×	×
ψ (Performance)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 5: Dimensions (resources and performance) for virtual personal assistants: Siri Apple [2019], Alexa Amazon [2019a,b], Cortana Microsoft [2019] and Google Assistant Google [2019].

6.5 Others

Further case studies can be analyzed, although with even much less detail and absence of data, so we just briefly comment on some important cases that are considered influential or impactful in AI.

IBM Watson Ferrucci [2012], which uses cognitive computing technology together with information retrieval support and natural language processing, has been successfully applied in different domains, such as health care and life sciences, education or business analytics. However, it presumably needs a large degree of adaptation effort across domains. It is not always explicit how much this effort is, as this is also part of the business model for IBM. As in many other commercial products, full replicability is not possible for obvious reasons, and only some general architectural aspects of IBM Watson are known.

Another interesting (yet again obscure) domain to explore is self-driving cars. Here, reliability, safety Hernández-Orallo et al. [2019, 2020] and social acceptance are only a few of the most well-known costs, but there are a lot more. Autonomous vehicles are not single devices but a collection of hardware and software pieces Wei et al. [2013] applied in complex and novel ways, so involving new research, development and production costs. For instance, with regard to hardware, while radars and sensor are already cheap and robust enough to be incorporated into mass-market cars, laser-shooting *LIDAR* or 3D photometry are still expensive solutions. The same happens with software: safe

driving requires more than state-of-the-art computer vision, it is also necessary to identify blind spots, use artificial intuition, anticipate driving behaviour, etc. All of this requires breakthrough end-to-end solutions with high costs in terms of data collection, model training, and testing in both simulation and real environments. Again, it is not always clear and transparent how much of this effort comes from high-level competition in the market between these players (e.g. Waymo⁸, General Motors⁹, Uber¹⁰, Tesla¹¹, etc.). On the other hand, we can also think about the not fully-explored societal costs (or benefits) self-driving cars might entail in terms of unemployment Poczter and Jankovic [2014], pollution Peter Fox-Penner and Gorman [2018], traffic World Economic Forum [2018], infrastructures and urban design Urmson et al. [2008], fatalities Sivak and Schoettle [2015], etc. In the same way, regulations are now compulsory for cars in terms of declaring their emissions, some other indicators of the AI self-driving technology that is incorporated in a car should also be disclosed and approved, so that we could plot several dimensions. Otherwise the cars of the future may end up having very efficient engines, but very inefficient ‘brains’.

7 Conclusions

The interest in more comprehensive evaluation protocols to assess the potential impact of new AI technologies and the progress they represent, going beyond performance alone, was illustrated by some of the references we included in the background section. However, in order to rigorously evaluate the impact a new contribution in AI can have more broadly, we need an explicit enumeration of all the dimensions (as represented by Table 1) and their integration into utility functions or their representation in a multidimensional space, with a clear delimitation of the extent of accounting. This represents a novel model to help anticipate the impact of a particular AI technology, bringing dimensions that are usually part of other disciplines or still not sufficiently technical or developed to be considered as parameters to optimize or evaluate. The several scenarios we have analyzed in this paper illustrate how the evaluation techniques can be applied in practice, but they also show that more transparency (through accountability, openness and replicability) has to be applied to AI research in the first place, to assess the contributions and potential economic and social impact more scientifically.

Of course, there can be resistance from AI researchers and reviewers, as more dimensions and indicators in papers, products and competitions can be seen as a counter-productive burden. Also, the lack of these dimensions in many papers today make illustrative examples, such as those we have included here, more challenging.

While we share some of these concerns, we have to look retrospectively to areas that were completely neglected a few years ago Hager et al. [2017], such as fairness, with a wide range of technical metrics that can be used in utility functions or in trade-offs against other dimensions. Similarly, this is what happened in cost-sensitive learning more than 15 years ago Elkan [2001], Turney [2002], leading to a wide range of techniques that covered different operating conditions. While all these costs are nowadays integrated into the measures of performance, many other resources are not, as we have surfaced here. Within this framework, we make a series of recommendations:

- Benchmarks and competitions should be defined in terms of a more comprehensive utility function, considering as many dimensions as possible, or recognize the value of all contributions that have any of the positive effects on the Pareto frontier identified previously, in short or long terms.
- Papers presenting or evaluating algorithms should generally try to report the whole region they cover, and how to navigate the region by modifying parameters or resources. There are many partial examples nowadays: learning curves, plots comparing the number of models vs. performance, planning performance vs. lookahead, etc.
- These utility functions and multidimensional spaces must also be seen in terms of replicability, for variants of the problems and at different stages of the AI life cycle. The multiplicities are more difficult to plot graphically, but we can still define operating conditions depending on the adaptation (or transfer) effort for m problems, or n users.

Frequently, we will not be able to say that one technique is ‘better’ than another: they just cover different regions of the multidimensional space. It is the receiver who will choose the system that best fits their needs. Having a representation of the Pareto frontier may hugely facilitate this choice for other researchers and industry, as simply as moving the gradient until touching the Pareto surface. Also, small players in AI could focus on those areas that require less resources and still contribute to the Pareto frontier or to diversity. Finally, the Pareto surface can help detect some

⁸<https://waymo.com/>

⁹<https://getcruise.com/>

¹⁰<https://www.uber.com/info/atg/technology/>

¹¹<https://www.tesla.com/autopilot>

societal risks, and unexpected huge social impact, especially if we see that a powerful capability in AI can be achieved with very few resources, becoming available to malicious actors.

This view of the operating condition as a gradient may suggest clever approaches to push the frontier for some resources, as gradient descent is increasingly being used at a meta-level Andrychowicz et al. [2016]. In general, we hope this paper will help change perceptions, promote more general and versatile techniques, highlight the trade-offs, and raise awareness of the overall “AI footprint”, well beyond performance.

Acknowledgments

This work has been partially supported by the Norwegian Research Council grant 329745 Machine Teaching for Explainable AI, also by the EU (FEDER) and Spanish MINECO grant RTI2018-094403-B-C32 funded by MCIN/AEI/10.13039/501100011033 and by “ERDF A way of making Europe”, Generalitat Valenciana under grant PROMETEO/2019/098, EU’s Horizon 2020 research and innovation programme under grant agreement No. 952215 (TAILOR), US DARPA HR00112120007 (RECoG-AI), and the UPV (Vicerrectorado de Investigación) grant PAI-10-21

References

- P Eckersley and N Yomna. Measuring the progress of AI research, 2017. URL <https://www.eff.org/ai/metrics>.
- Yoav Shoham, Raymond Perrault, Erik Brynjolfsson, Jack Clark, and Calvin LeGassick. AI Index, 2017. URL <http://cdn.aiindex.org/2017-report.pdf>.
- Fernando Martínez-Plumed, Pablo Barredo, Sean O Heigeartaigh, and José Hernández-Orallo. Research community dynamics behind popular ai benchmarks. *Nature Machine Intelligence*, 3(7):581–589, 2021a.
- Fernando Martínez-Plumed and José Hernández-Orallo. Ai results for the atari 2600 games: difficulty and discrimination using irt. *EGPAI, Evaluating General-Purpose Artificial Intelligence*, 33, 2016.
- Fernando Martínez-Plumed, Ricardo BC Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. Making sense of item response theory in machine learning. In *ECAI 2016*, pages 1140–1148. IOS Press, 2016.
- Fernando Martínez-Plumed and Jose Hernandez-Orallo. Dual indicators to analyze ai benchmarks: Difficulty, discrimination, ability, and generality. *IEEE Transactions on Games*, 12(2):121–131, 2018.
- Fernando Martínez-Plumed, Ricardo BC Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. Item response theory in ai: Analysing machine learning classifiers at the instance level. *Artificial intelligence*, 271:18–42, 2019.
- José Hernández-Orallo, Wout Schellaert, and Fernando Martínez-Plumed. Training on the test set: Mapping the system-problem space in ai. 2022.
- Fernando Martínez-Plumed, David Castellano-Falcón, Carlos Monserrat, and José Hernández-Orallo. When ai difficulty is easy: The explanatory power of predicting irt difficulty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.
- José Hernández-Orallo. *The measure of all minds: evaluating natural and artificial intelligence*. Cambridge University Press, 2017a.
- José Hernández-Orallo, Bao Sheng Loe, Lucy Cheke, Fernando Martínez-Plumed, and Seán Ó hÉigeartaigh. General intelligence disentangled via a generality metric for natural and artificial intelligence. *Scientific reports*, 11(1):1–16, 2021.
- José Hernández-Orallo. Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. *Artificial Intelligence Review*, 48(3):397–447, 2017b.
- Jose Hernández-Orallo, Marco Baroni, Jordi Bieger, Nader Chmait, David L Dowe, Katja Hofmann, Fernando Martínez-Plumed, Claes Strannegård, and Kristinn R Thórisson. A new ai evaluation cosmos: Ready to play the game? In *AI Magazine*, volume 38, pages 66–69, 2017.
- Carl Benedikt Frey and Michael A Osborne. The future of employment: how susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114:254–280, 2017.
- Erik Brynjolfsson and Tom Mitchell. What can machine learning do? Workforce implications. *Science*, 358(6370):1530–1534, 2017.
- Philippe Aghion, Benjamin F Jones, and Charles I Jones. Artificial intelligence and economic growth. *National Bureau of Economic Research*, 2017.
- Anton Korinek and Joseph E Stiglitz. Artificial intelligence and its implications for income distribution and unemployment. *National Bureau of Economic Research*, 2017.

- Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *BBS*, 40, 2017.
- Gary Marcus. Deep learning: A critical appraisal. *CoRR abs/1801.00631*, 2018.
- Nihar B Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike Von Luxburg. Design and analysis of the nips 2016 review process. *CoRR abs/1708.09794*, 2017.
- Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4):391–444, 2007.
- David H Wolpert. What the no free lunch theorems really mean; how to improve search algorithms. In *Santa fe Institute Working Paper*, page 12. 2012.
- C. Drummond. Replicability is not reproducibility: nor is it good science. *Evaluation Methods for ML (ICML)*, 2009.
- Fabio Bonsignorio and Angel P Del Pobil. Toward replicable and measurable robotics research. *IEEE Robotics & Aut. M.*, 22(3):32–35, 2015.
- Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. *CoRR abs/1709.06560*, 2017.
- Rolf A Zwaan, Alexander Etz, Richard E Lucas, and M Brent Donnellan. Making replication mainstream. *Behavioral and Brain Sciences*, pages 1–50, 2017.
- J. Spohrer. Opentech AI workshop, 2017. URL <https://opentechai.blog/2017/12/29/opentech-ai-workshop/>.
- Brandon Reagen, Robert Adolf, Paul Whatmough, Gu-Yeon Wei, and David Brooks. Deep learning for computer architects. *SL on Comp. Architecture*, 12(4):1–123, 2017.
- Tim Hwang. Computational power and the social impact of artificial intelligence. *CoRR abs/1803.08971*, 2018.
- Miles Brundage. Modeling progress in ai. *AAAI Workshop on AI, Ethics, and Society*, arXiv preprint arXiv:1512.05849, 2016.
- Katja Grace. Trends in algorithmic progress, 2017. URL <https://aiimpacts.org/trends-in-algorithmic-progress/>.
- Cody Coleman, Deepak Narayanan, Daniel Kang, Tian Zhao, Jian Zhang, Luigi Nardi, Peter Bailis, Kunle Olukotun, Chris Ré, and Matei Zaharia. Dawnbench: An end-to-end deep learning benchmark and competition, 2017. URL <http://dawn.cs.stanford.edu/benchmark/>.
- MLPerf. Benchmark suite. <https://mlperf.org/>, 2018.
- Nicolas Bougie and Ryutaro Ichise. Deep reinforcement learning boosted by external knowledge. *CoRR abs/1712.04101*, 2017.
- Charles Elkan. The foundations of cost-sensitive learning. In *IJCAI*, volume 17, pages 973–8, 2001.
- Peter D Turney. Types of cost in inductive concept learning. *CoRR abs/cs/0212034*, 2002.
- D Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In *NIPS*, pages 2503–2511, 2015.
- Chris Olah and Shan Carter. Research debt. *Distill*, 2(3):e5, 2017.
- Radosvet Desislavov, Fernando Martínez-Plumed, and José Hernández-Orallo. Compute and energy consumption trends in deep learning inference. *arXiv preprint arXiv:2109.05472*, 2021.
- Fernando Martínez-Plumed, Bao Sheng Loe, Peter Flach, Seán Ó hÉigeartaigh, Karina Vold, and José Hernández-Orallo. The facets of artificial intelligence: A framework to track the evolution of ai. In *IJCAI*, pages 5180–5187, 7 2018. doi:10.24963/ijcai.2018/718.
- Fernando Martínez-Plumed, Songül Tolan, Annarosa Pesole, José Hernández-Orallo, Enrique Fernández-Macías, and Emilia Gómez. Does ai qualify for the job? a bidirectional model mapping labour and ai intensities. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 94–100, 2020a.
- Songül Tolan, Annarosa Pesole, Fernando Martínez-Plumed, Enrique Fernández-Macías, José Hernández-Orallo, and Emilia Gómez. Measuring the occupational impact of ai: tasks, cognitive abilities and ai benchmarks. *Journal of Artificial Intelligence Research*, 71:191–236, 2021.
- Lawrence H. Putnam. A general empirical solution to the macro software sizing and estimating problem. *IEEE Trans. on Software Engineering*, (4):345–361, 1978.

- Ian Sommerville. *Software engineering*. Addison-wesley, 2015.
- Daniel Ragan, Peter Sandborn, and Paul Stoaks. A detailed cost model for concurrent use with hardware/software co-design. In *Design Automation Conference*, pages 269–274. ACM, 2002.
- MathWorks. Estimate computation costs. <https://www.mathworks.com/help/physmod/simscape/ug/estimate-computation-costs.html>, 2018.
- Richard Veryard. *The economics of Information Systems and software*. Butterworth-Heinemann, 2014.
- James C Anderson, James A Narus, and Wouter Van Rossum. Customer value propositions in business markets. *Harvard business review*, 84:1–4, 2006.
- Sorelle A. Friedler and Christo Wilson. Preface. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 1–2, New York, NY, USA, 23–24 Feb 2018. PMLR. URL <http://proceedings.mlr.press/v81/friedler18a.html>.
- Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, 2019. ACM. URL <https://dl.acm.org/citation.cfm?id=3287588>.
- Martínez-Plumed Fernando, Ferri Cèsar, Nieves David, and Hernández-Orallo José. Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems*, 36(7):3217–3258, 2021.
- Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.
- Tianshi Chen, Zidong Du, Ninghui Sun, Jia Wang, Chengyong Wu, Yunji Chen, and Olivier Temam. Diannao: A small-footprint high-throughput accelerator for ubiquitous machine learning. In *Sigplan Not.*, volume 49, pages 269–284. ACM, 2014.
- Fernando Martínez-Plumed, Emilia Gómez, and José Hernández-Orallo. Futures of artificial intelligence through technology readiness levels. *Telematics and Informatics*, 58:101525, 2021b.
- Fernando Martínez-Plumed, Jose Hernández-Orallo, and Emilia Gómez. Tracking ai: The capability is (not) near. In *ECAI 2020*, pages 2915–2916. IOS Press, 2020b.
- David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.
- David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550: 354, 2017a.
- David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dhharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *CoRR abs/1712.01815*, 2017b.
- Norman P Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, et al. In-datacenter performance analysis of a tensor processing unit. In *Computer Architecture*, pages 1–12. ACM, 2017.
- Audrunas Gruslys, Mohammad Gheshlaghi Azar, Marc G. Bellemare, and Rémi Munos. The reactor: A sample-efficient actor-critic architecture. *CoRR*, abs/1704.04651, 2017.
- Matteo Hessel, Joseph Modayil, Hado van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Daniel Horgan, Bilal Piot, Mohammad Gheshlaghi Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. *CoRR*, abs/1710.02298, 2017a.
- M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. The arcade learning environment: An evaluation platform for general agents. *JAIR*, 47:253–279, jun 2013.
- Marlos C. Machado, Marc G. Bellemare, Erik Talvitie, Joel Veness, Matthew J. Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *CoRR*, abs/1709.06009, 2017. URL <http://arxiv.org/abs/1709.06009>.
- Stephen Kelly and Malcolm I Heywood. Emergent tangled graph representations for atari game playing agents. In *EuroGP*, pages 64–79. Springer, 2017.
- Wilmer Bandres, Blai Bonet, and Hector Geffner. Planning with pixels in (almost) real time. *CoRR abs/1801.03354*, 2018.

- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- Sergey Zagoruyko. imagenet-validation.torch. <https://github.com/szagoruyko/imagenet-validation.torch>, 2016.
- Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.
- Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016.
- Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
- Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017.
- X Zhang, X Zhou, M Lin, and J Sun. Shufflenet: an extremely efficient convolutional neural network for mobile devices (2017). arxiv preprint. *arXiv preprint arXiv:1707.01083*.
- Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520. IEEE, 2018.
- François Chollet. Xception: Deep learning with depthwise separable convolutions. *arXiv preprint*, pages 1610–02357, 2017.
- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.
- Zheng Qin, Zhaoning Zhang, Xiaotao Chen, Changjian Wang, and Yuxing Peng. Fd-mobilenet: Improved mobilenet with a fast downsampling strategy. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1363–1367. IEEE, 2018.
- Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. Regularized evolution for image classifier architecture search. *arXiv preprint arXiv:1802.01548*, 2018.
- Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.
- Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 116–131, 2018.
- Yanping Huang, Yonglong Cheng, Dehao Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, and Zhifeng Chen. Gpipe: Efficient training of giant neural networks using pipeline parallelism. *arXiv preprint arXiv:1811.06965*, 2018.
- Xingcheng Zhang, Zhizhong Li, Chen Change Loy, and Dahua Lin. Polynet: A pursuit of structural diversity in very deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 718–726, 2017a.

- Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*, 2016.
- Apple. Siri. <https://www.apple.com/siri/>, 2019.
- Amazon. Amazon Echo and Alexa Devices. <https://www.amazon.com/Amazon-Echo-And-Alexa-Devices/b?ie=UTF8&node=9818047011>, 2019a.
- Amazon. Alexa skills kit. <https://developer.amazon.com/public/solutions/alexa/alexa-skills-kit>, 2019b.
- Microsoft. Cortana. <https://www.microsoft.com/en-us/cortana>, 2019.
- Google. Google assistant. <https://assistant.google.com/>, 2019.
- Justine Cassell, Joseph Sullivan, Elizabeth Churchill, and Scott Prevost. *Embodied conversational agents*. MIT press, 2000.
- Arsénio Reis, Dennis Paulino, Hugo Paredes, and João Barroso. Using intelligent personal assistants to strengthen the elderlies' social bonds. In *International Conference on Universal Access in Human-Computer Interaction*, pages 593–602. Springer, 2017.
- Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. Automatic online evaluation of intelligent assistants. In *Proceedings of the 24th International Conference on World Wide Web*, pages 506–516. International World Wide Web Conferences Steering Committee, 2015.
- David L Strayer, Joel M Cooper, Jonna Turrill, James R Coleman, and Rachel J Hopman. The smartphone and the driver's cognitive workload: A comparison of apple, google, and microsoft's intelligent personal assistants. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 71(2):93, 2017.
- Roberto Pellungrini, Luca Pappalardo, Francesca Pratesi, and Anna Monreale. A data mining approach to assess privacy risk in human mobility data. *ACM Trans. Intell. Syst. Technol.*, 9(3):31:1–31:27, December 2017. ISSN 2157-6904. doi:10.1145/3106774. URL <http://doi.acm.org/10.1145/3106774>.
- Ruide Zhang, Ning Zhang, Changlai Du, Wenjing Lou, Y. Thomas Hou, and Yuichi Kawamoto. From electromyogram to password: Exploring the privacy impact of wearables in augmented reality. *ACM Trans. Intell. Syst. Technol.*, 9(1):13:1–13:20, September 2017b. ISSN 2157-6904. doi:10.1145/3078844. URL <http://doi.acm.org/10.1145/3078844>.
- Matthew B Hoy. Alexa, siri, cortana, and more: An introduction to voice assistants. *Medical reference services quarterly*, 37(1):81–88, 2018.
- Lydia Manikonda, Aditya Deotale, and Subbarao Kambhampati. What's up with privacy?: User preferences and privacy concerns in intelligent personal assistants. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 229–235. ACM, 2018.
- Eric Enge. Rating the smarts of the digital personal assistants in 2018. <https://www.stonetemple.com/digital-personal-assistants-study/>, 2018.
- Emil Protalinski. Google's speech recognition technology now has a 4.9rate. <https://venturebeat.com/2017/05/17/googles-speech-recognition-technology-now-has-a-4-9-word-error-rate/>, 2017.
- Craig Lloyd. How much electricity does the amazon echo use? <https://www.howtogeek.com/348219/how-much-electricity-does-the-amazon-echo-use/>, 2018.
- Andrew Williams. How much power does your smart home tech really use? <https://www.the-ambient.com/features/power-smart-home-tech-yearly-cost-374>, 2018.
- D. A. Ferrucci. Introduction to "this is watson". *IBM J. Res. Dev.*, 56(3):235–249, May 2012. ISSN 0018-8646.
- José Hernández-Orallo, Fernando Martínez-Plumed, Shahar Avin, et al. Surveying safety-relevant ai characteristics. In *SafeAI@ AAAI*, 2019.
- Jose Hernández-Orallo, Fernando Martínez-Plumed, Shahar Avin, Jess Whittlestone, et al. Ai paradigms and ai safety: mapping artefacts and techniques to safety issues. 2020.
- Junqing Wei, Jarrod M Snider, Junsung Kim, John M Dolan, Raj Rajkumar, and Bakhtiar Litkouhi. Towards a viable autonomous driving research platform. In *2013 IEEE Intelligent Vehicles Symposium (IV)*, pages 763–770. IEEE, 2013.
- Sharon L Poczter and Luka M Jankovic. The google car: driving toward a better future? *Journal of Business Case Studies (Online)*, 10(1):7, 2014.

- Jennifer Hatch Peter Fox-Penner and Will Gorman. Spread of self-driving cars could cause more pollution – unless the electric grid transforms radically. <https://theconversation.com/spread-of-self-driving-cars-could-cause-more-pollution-unless-the-electric-grid-transforms-radically> 2018.
- World Economic Forum. Reshaping urban mobility with autonomous vehicles lessons from the city of boston. http://www3.weforum.org/docs/WEF_Reshaping_Urban_Mobility_with_Autonomous_Vehicles_2018.pdf, 2018.
- Chris Urmson et al. Self-driving cars and the urban challenge. *IEEE Intelligent Systems*, 23(2):66–68, 2008.
- Michael Sivak and Brandon Schoettle. Road safety with self-driving vehicles: General limitations and road sharing with conventional vehicles. 2015.
- Gregory D Hager, Randal Bryant, Eric Horvitz, Maja Mataric, and Vasant Honavar. Advances in AI require progress across all of computer science. *CoRR abs/1707.04352*, 2017.
- Marcin Andrychowicz, Misha Denil, Sergio Gomez Colmenarejo, Matthew W. Hoffman, David Pfau, Tom Schaul, and Nando de Freitas. Learning to learn by gradient descent by gradient descent. *CoRR*, abs/1606.04474, 2016. URL <http://arxiv.org/abs/1606.04474>.
- Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.
- Arun Nair, Praveen Srinivasan, Sam Blackwell, Cagdas Alcicek, Rory Fearon, Alessandro De Maria, Vedavyas Panneershelvam, Mustafa Suleyman, Charles Beattie, Stig Petersen, et al. Massively parallel methods for deep reinforcement learning. *arXiv preprint arXiv:1507.04296*, 2015.
- Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *AAAI*, volume 2, page 5. Phoenix, AZ, 2016.
- Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.
- Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Van Hasselt, Marc Lanctot, and Nando De Freitas. Dueling network architectures for deep reinforcement learning. *arXiv preprint arXiv:1511.06581*, 2015.
- Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937, 2016.
- Hado P van Hasselt, Arthur Guez, Matteo Hessel, Volodymyr Mnih, and David Silver. Learning values across many orders of magnitude. In *Advances in Neural Information Processing Systems*, pages 4287–4295, 2016.
- Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, pages 1471–1479, 2016.
- Marc G Bellemare, Joel Veness, and Michael Bowling. Investigating contingency awareness using atari 2600 games. In *AAAI*, 2012.
- Haoran Tang, Rein Houthoofd, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2753–2762, 2017.
- Georg Ostrovski, Marc G Bellemare, Aaron van den Oord, and Rémi Munos. Count-based exploration with neural density models. *arXiv preprint arXiv:1703.01310*, 2017.
- Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.
- Audrunas Gruslys, Will Dabney, Mohammad Gheshlaghi Azar, Bilal Piot, Marc Bellemare, and Remi Munos. The reactor: A fast and sample-efficient actor-critic agent for reinforcement learning. 2018.
- Jarryd Martin, Suraj Narayanan Sasikumar, Tom Everitt, and Marcus Hutter. Count-based exploration in feature space for reinforcement learning. *arXiv preprint arXiv:1706.08090*, 2017.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. *arXiv preprint arXiv:1707.06887*, 2017.

Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. *arXiv preprint arXiv:1710.02298*, 2017b.

A Supplementary materials

A.1 Case study 1: Alpha* series

In this section, we showcase the resources used for the development of the Alpha* systems to play board games. Table 6 shows the computational resources for the different Alpha* systems according to the information disclosed in the original papers. It should be noted that most of the computational effort for the development of AlphaGo Master, AlphaGo Zero and the Alpha Zero was used for generating the training data (self-play games). However this data is not explicitly specify for all the systems with the exception of Alpha Zero. Therefore, the compute power used for generating this training data has not been considered for generating the multidimensional utility space for Alpha*.

		Self-Play HW	Self-Play HW (CPU)	Training HW	Training HW (CPU)	Playing HW	Playing HW (CPU)	ELO rating	Publication Date
AlphaGo Fan	Silver et al. [2016]	-	-	50 GPU	250	176 GPU + 1202 CPU	2082	3144	Jan 2016
AlphaGo Lee	Silver et al. [2016]	-	-	50 GPU	250	1920 CPU + 280 GPU	3320	3739	March 2016
AlphaGo Master	Silver et al. [2016]	-	-	64 GPU	320	4 TPU_{v1}	240	4848	Dec 2016
AlphaGo Zero	Silver et al. [2017a]	-	-	64 GPU + 19 CPU	339	4 TPU_{v1}	240	5185	Oct 2017
Alpha Zero	Silver et al. [2017b]	5000 TPU_{v1}	300000	64 TPU_{v2}	11520	4 TPU_{v2}	720	5000	Dec 2017

Table 6: Computational resources for Alpha* systems. Normalized hardware (in CPUs) has been calculated using the equivalence: $1 TPU_{v2} \simeq 3 TPU_{v1} \simeq 36 GPU \simeq 180 CPU$ from Jouppi et al. [2017])

A.2 Case study 2: ALE papers

In this section we examine the resources used for producing the RL-based systems to play the collection of Atari games in ALE benchmark. Table 7 shows the computational resources for the different systems addressing ALE benchmark according to the information reported in the original papers. In this case, most of the computational effort for the development of the systems fall within the training procedures (CPUs per day and millions of frames used to learn). Whenever the authors do not provide this information explicitly, (e.g., training time or hardware used), this is roughly estimated from the kind of approach used, whether it is follow-up work, the training setting used, etc., or from figures in more recent papers, which make explicit comparisons between them and the state of the art (such as in Hessel et al. [2017a], Gruslys et al. [2017]).

A.3 Case study 3: ImageNet papers

In this section we disclose the resources used by systems addressing the third case study: the Imagenet Large Scale Visual Recognition Challenge (ILSVRC). Table 8 shows a number of metrics related to the computational resources used by the different approaches (training hardware, number of layers, number of training images, training time and number of operations) as well as their load/size (number of parameters learned) and the energy consumption, according to the information disclosed in the original papers or in different sources (possibly from different authors).

A.4 Case study 4: Personal Assistants

In this section we display some of the costs (power requirements in deployment) and performance of those virtual personal assistants analysed in the fourth case study. Table 9 shows the values of this limited number of resources, mostly collected from external sources.

Algorithm	Training Time (days)	Training HW	Training HW (CPU/days)	#Workers	Training Frames (Millions)	Games Tested	Frames Test	Test Procedure	Score norm. (median)	Publication Date
Best linear Bellemare et al. [2013]	17	CPU	17	1	0.15	50	18000	-	14.04	19/07/2012
DQN best Mnih et al. [2013]	8	GPU	40	1	50	7	-	-	67.96	19/12/2013
Nature DQN Mnih et al. [2015]	8	GPU	40	1	50	49	18000	noop	92.67	26/02/2015
Gorilla Nair et al. [2015]	4	GPU	20	100	200	49	108000	hs	68.56	15/07/2015
DDQN (tuned) hs Van Hasselt et al. [2016]	8	GPU	40	1	200	57	108000	hs	121.04	22/09/2015
DQN hs Van Hasselt et al. [2016]	8	GPU	40	1	200	49	108000	hs	68.52	22/09/2015
DQN noop Van Hasselt et al. [2016]	8	GPU	40	1	200	49	18000	noop	89.30	22/09/2015
Prior hs Schaul et al. [2015]	8	GPU	40	1	200	49	108000	hs	119.74	18/11/2015
Prior noop Schaul et al. [2015]	8	GPU	40	1	200	49	108000	noop	137.66	18/11/2015
DDQN (tuned) noop Wang et al. [2015]	8	GPU	40	1	200	57	18000	noop	132.52	20/11/2015
Duel hs Wang et al. [2015]	8	GPU	40	1	200	57	108000	hs	131.66	20/11/2015
Duel noop Wang et al. [2015]	8	GPU	40	1	200	57	18000	noop	166.38	20/11/2015
Prior+Duel hs Wang et al. [2015]	8	GPU	40	1	200	57	108000	hs	123.48	20/11/2015
Prior+Duel noop Wang et al. [2015]	8	GPU	40	1	200	57	18000	noop	159.73	20/11/2015
A3C FF (1 day) hs Mnih et al. [2016]	1	CPU	1	16	320	57	108000	hs	78.53	04/02/2016
A3C FF hs Mnih et al. [2016]	4	CPU	4	16	320	57	108000	hs	117.44	04/02/2016
A3C LSTM hs Mnih et al. [2016]	4	CPU	4	16	320	57	108000	hs	112.63	04/02/2016
DDQN+Pop-Art noop van Hasselt et al. [2016]	8	GPU	40	1	200	49	108000	noop	112.83	24/02/2016
A3C-CTS Bellemare et al. [2016]	4	CPU	4	16	200	60	-	-	89.67	06/06/2016
SARSA Bellemare et al. [2012]	30	CPU	30	1	2	46	-	-	10.09	06/06/2016
TRPO-hash Tang et al. [2017]	8	GPU	40	1	50	6	-	-	28.90	15/11/2016
DQN-CTS Ostrovski et al. [2017]	8	GPU	40	1	150	57	-	-	11.98	03/03/2017
DQN-PixelCNN Ostrovski et al. [2017]	8	GPU	40	1	150	57	-	-	27.45	03/03/2017
ES FF (1 hour) noop Salimans et al. [2017]	0.0416	CPU	0.0416	1	1000 M	51	108000	noop	40.18	10/03/2017
REACTOR hs Gruslys et al. [2018]	2	CPU	2	10 + 1 ¹	200	57	108000	hs	154.42	15/04/2017
REACTOR 500M hs Gruslys et al. [2018]	4	CPU	4	10 + 1 ¹	500	57	108000	hs	185.56	15/04/2017
REACTOR noop Gruslys et al. [2018]	2	CPU	2	10 + 1 ¹	200	57	18000	noop	185.95	15/04/2017
REACTOR 500M noop Gruslys et al. [2018]	4	CPU	4	10 + 1 ¹	500 M	57	18000	noop	291.74	15/04/2017
Sarsa-e Martin et al. [2017]	8	GPU	40	1	100 M	5	18000	noop	28.07	25/06/2017
Sarsa-f-EB Martin et al. [2017]	8	GPU	40	1	100 M	5	18000	noop	62.86	25/06/2017
C51 noop Bellemare et al. [2017]	8	GPU	40	1	200 M	57	18000	noop	177.71	21/07/2017
Rainbow hs Hessel et al. [2017b]	10	GPU	40	1	200 M	54	108000	hs	144.96	06/10/2017
Rainbow noop Hessel et al. [2017b]	10	GPU	40	1	200 M	54	18000	noop	232.46	06/10/2017

¹ 10 actor-learner workers (CPUs) and 1 parameter server.

Table 7: Computational resources for the systems addressing ALE benchmark Bellemare et al. [2013]. Systems from Eckersley and Yomna [2017], Hessel et al. [2017a], Gruslys et al. [2017]. Normalized training time (in CPUs) has been calculated using the equivalence: 1 GPU \approx 5 CPU from Jouppi et al. [2017]. #Workers represents the number of parallel machines used. We calculated median human normalised scores across all games according to Nair et al. [2015].

System	Training HW	#Layers	#Params (Millions)	Batch Size	Training time	G-Ops*	Net power (Watts)*	Top-1 ACC*	Date
AlexNet Krizhevsky et al. [2012]	2 \times nVidia GTX580	8	60	128	6 days	2.26	11.20	54.61	03/12/2012
BN-NIN Lin et al. [2013]	-	8	8.6	128	-	2.52	12.60	62.62	16/12/2013
VGG-16 Simonyan and Zisserman [2014]	4 \times nVidia Titan Black	16	138.4	64	2-3 weeks	30.97	12.40	70.62	04/9/2014
VGG-19 Simonyan and Zisserman [2014]	4 \times nVidia Titan Black	19	143.7	128	2-3 weeks	39.29	12.20	70.74	04/9/2014
GoogLeNet Szegedy et al. [2015]	few GPUs	22	7	128	1 week	3.00	11.15	68.70	11/9/2014
BN-AlexNet Zagoruyko [2016]	-	8	60.6	256	-	1.30	11.40	56.60	11/2/2015
Inception-v3 Szegedy et al. [2016]	50 \times NVidia Kepler	42	23.85	256	-	11.45	12.30	78.53	02/12/2015
ResNet-18 He et al. [2016]	few GPUs	18	11.7	256	-	3.63	12.30	69.57	10/12/2015
ResNet-34 He et al. [2016]	few GPUs	34	21.8	256	-	7.34	12.80	73.27	10/12/2015
ResNet-50 He et al. [2016]	few GPUs	50	25.6	256	-	8.21	11.80	75.99	10/12/2015
ResNet-101 He et al. [2016]	few GPUs	101	44.6	256	-	15.65	11.45	77.56	10/12/2015
ResNet-152 He et al. [2016]	few GPUs	152	60.3	256	-	23.10	11.40	77.84	10/12/2015
Inception-v4 Szegedy et al. [2017]	20 \times NVidia Kepler	75	31.6	256	-	18.44	11.60	80.10	23/2/2016
ENet Paszke et al. [2016]	4 nVidia Titan X	29	5.9	128	3-6 hours	1.57	11.60	68.40	7/6/2016
DenseNet-121 Huang et al. [2017]	8 \times nVidia Tesla M40	121	8	256	-	5.71	-	74.98	25/8/2016
DenseNet-169 Huang et al. [2017]	8 \times nVidia Tesla M40	169	14.2	256	-	6.77	-	76.20	25/8/2016
DenseNet-201 Huang et al. [2017]	8 \times nVidia Tesla M40	201	20	256	-	8.65	-	77.42	25/8/2016
Xception Chollet [2017]	60 \times nVidia K80	36	22.9	256	3 days	16.75	-	79.00	07/10/2016
SqueezeNet Iandola et al. [2016]	-	8	1.3	512	-	1.64	-	57.50	04/11/2016
PolyNet Zhang et al. [2017a]	32 \times nVidia Titan X	92	2.9	512	-	-	-	81.29	17/11/2016
MobileNet-v1 Howard et al. [2017]	-	14	4.3	96	-	1.14	-	70.90	17/4/2017
ShuffleNet Zhang et al.	4 \times GPUs	8	1.9	1024	1-2 days	1.57	-	62.20	04/7/2017
MobileNet-v2 Sandler et al. [2018]	16 \times GPU	20	3.5	96	-	0.60	-	71.80	13/1/2018
fd-MobileNet Qin et al. [2018]	4 \times GPUs	12	2.9	256	-	0.29	-	65.30	11/2/2018
AmoebaNet Real et al. [2018]	450 \times nVidia K40	-	469	256	7 days	-	-	83.9	05/02/2018
Shufflenet v2 Ma et al. [2018]	64 \times nVidia Titan Pascal	164	137	4	-	-	-	81.44	30/07/2018
SENet Hu et al. [2018]	8 \times nVidia Titan X	154	145.8	256	-	-	-	80.19	05/09/2018
GPipe Huang et al. [2018]	TPU _{v2}	-	557	256	-	-	-	84.3	16/11/2018

¹ Results over the ImageNet validation set.

Table 8: Computational, load, and energy resources for ImageNet systems. Some of the values for the attributes with * were obtained from Canziani et al. [2016].

System	Power Req. (Watts)	Power Idle (Watts)	Power Playing music (Watts)	Q. Attempted (2017)	Q. Correct (2017)	Q. Attempted (2018)	Q. Correct (2018)	Date
Apple Siri	-	-	9.25	31.40	86.10	40.80	80.00	12/10/2011
Microsoft Cortana	-	-	-	53.90	86.00	64.90	91.60	2/4/2014
Amazon 1st-Gen Echo	21.00	2.95	3.25	19.80	94.50	53.70	86.20	6/11/2014
Amazon 1st-Gen Echo Dot	9.36	1.75	2.25	19.80	94.50	53.70	86.20	1/3/2016
Google Assistant	-	-	-	-	-	77.20	95.20	1/5/2016
Amazon 2nd-Gen Echo Dot	9.36	1.75	2.25	19.80	94.50	53.70	86.20	20/10/2016
Google Home	33.00	2.00	-	65.30	94.50	66.20	87.50	4/11/2016
Amazon Echo Spot	10.92	2.08	2.90	19.80	94.50	53.70	86.20	27/9/2017
Amazon 2nd-Gen Echo	21.00	1.95	2.90	19.80	94.50	53.70	86.20	1/10/2017
Google Home Mini	9.00	1.50	2.25	65.30	94.50	66.20	87.50	4/10/2017
Google Home Max	-	-	-	65.30	94.50	66.20	87.50	4/10/2017
Amazon Echo Plus	30.00	2.40	3.65	19.80	94.50	53.70	86.20	31/10/2017
Apple HomePod	-	1.76	9.25	31.40	86.10	40.80	80.00	9/2/2018
Amazon 3rd-Gen Echo Dot	15.00	-	-	19.80	94.50	53.70	86.20	1/9/2018
Google Home Hub	33.00	-	-	65.30	94.50	66.20	87.50	9/10/2018

Table 9: Energy resources for personal assistants. Performance values (questions attempted and questions correct) are from Enge [2018] where the digital marketing firm *Stone Temple* tested the assistants via an exhaustive list of 4,942 queries. Power requirements have been collected from manuals and other sources. Power consumptions (idle and playing music) are from Lloyd [2018], Williams [2018]