Science Policy Forum

Rethink reporting of evaluation results in AI

Aggregate metrics and lack of access to results limit understanding

By **Ryan Burnell[1], Wout Schellaert[2], John Burden[1,3], Tomer D. Ullman[4], Fernando Martinez-Plumed[2], Joshua B. Tenenbaum[5], Danaja Rutar[1], Lucy G. Cheke[1,6], Jascha Sohl-Dickstein[7], Melanie Mitchell[8], Douwe Kiela[9], Murray Shanahan[10,11], Ellen M. Voorhees[12], Anthony G. Cohn[13,14,15,16], Joel Z. Leibo[10], Jose Hernandez-Orallo[1,2,3]**

[1]Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, UK. [2]Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de Valencia, València, Spain. [3]Centre for the Study of Existential Risk, University of Cambridge, Cambridge, UK. [4]Department of Psychology, Harvard University, Cambridge, MA, USA. [5]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [6]Department of Psychology, University of Cambridge, Cambridge, UK. [7]Brain team, Google, Mountainview, CA, USA. [8]Santa Fe Institute, Santa Fe, NM, USA. [9]Stanford University, Stanford, CA, USA. [10]DeepMind, London, UK. [11]Department of Computing, Imperial College London, London, UK. [12]National Institute of Standards and Technology (Retired), Gaithersburg, MD, USA. [13]School of Computing, University of Leeds, Leeds, UK. [14]Alan Turing Institute, London, UK. [15]Tongji University, Shanghai, China. [16]Shandong University, Jinan, China.

Email: rb967@cam.ac.uk

ABSTRACT

Artificial intelligence (AI) systems have begun to be deployed in high-stakes contexts, including autonomous driving and medical diagnosis. In contexts such as these, the consequences of system failures can be devastating. It is therefore vital that researchers and policy-makers have a full understanding of the capabilities and weaknesses of AI systems so that they can make informed decisions about where these systems are safe to use and how they might be improved. Unfortunately, current approaches to AI evaluation make it exceedingly difficult to build such an understanding, for two key reasons. First, aggregate metrics make it hard to predict how a system will perform in a particular situation. Second, the instance-by-instance evaluation results that could be used to unpack these aggregate metrics are rarely made available (1). Here, we propose a path forward in which results are presented in more nuanced ways and instance-by-instance evaluation results are made publicly available.

Artificial intelligence (AI) systems have begun to be deployed in high-stakes contexts, including autonomous driving and medical diagnosis. In contexts such as these, the consequences of system failures can be devastating. It is therefore vital that researchers and policymakers have a full understanding of the capabilities and weaknesses of AI systems so that they can make informed decisions about where these systems are safe to use and how they might be improved. Unfortunately, current approaches to AI evaluation make it exceedingly difficult to build such an understanding, for two key reasons. First, aggregate metrics make it hard to predict how a system will perform in a particular situation. Second, the instance-by-instance evaluation results that could be used to unpack these aggregate metrics are rarely made available (*1*). Here, we propose a path forward in which results are presented in more nuanced ways and instance-by-instance evaluation results are made publicly available.

Across most areas of AI, system evaluations follow a similar structure. A system is first built or trained to perform a particular set of functions. Then, the performance of the system is tested on a set of tasks relevant to the desired functionality of the system. In many areas of AI, evaluations use standardized sets of tasks known as "benchmarks." For each task, the system will be tested on a number of example "instances" of the task. The system would then be given a score for each instance based on its performance, e.g., 1 if it classified an image correctly, or 0 if it was incorrect. For other systems, the score for each instance might be based on how quickly the system completed its task, the quality of its outputs, or the total reward it obtained. Finally, performance across the various instances and tasks is usually aggregated to a small number of metrics that summarize how well the system performed, such as percentage accuracy.

But aggregate metrics limit our insight into performance in particular situations, making it harder to find system failure points and robustly evaluate system safety. This problem is also worsening as the increasingly broad capabilities of state-of-the-art systems necessitate ever

more diverse benchmarks to cover the range of their capabilities. This problem is further exacerbated by a lack of access to the instance-by-instance results underlying the aggregate metrics, making it difficult for researchers and policy-makers to further scrutinize system behavior.

AGGREGATE METRICS

Use of aggregate metrics is understandable. They provide information about system performance "at a glance" and allow for simple comparisons across systems. But aggregate performance metrics obfuscate key information about where systems tend to succeed or fail (*2*). Take, for example, a system that was trained to classify faces as male or female that achieved classification accuracy of 90% (*3*). Based on this metric, the system appears highly competent. However, a subsequent breakdown of performance revealed that the system misclassified females with darker skin types a staggering 34.5% of the time, while erring only 0.8% of the time for males with lighter skin types. This example demonstrates how aggregation can make it difficult for policymakers to determine the fairness and safety of AI systems.

Compounding this problem, many benchmarks include disparate tasks that are ultimately aggregated together. For example, the Beyond the Imitation Game Benchmark (BIG-bench) for language models includes over 200 tasks that evaluate everything from language understanding to causal reasoning (*4*). Aggregating across these disparate tasks—as the BIG-bench leaderboard does—reduces the rich information in the benchmark to an overall score that is hard to interpret.

It is also easy for aggregation to introduce unwarranted assumptions into the evaluation process. For example, a simple average across tasks implicitly treats every task as equally important—in the case of BIG-bench, a sports understanding task has as much bearing on the overall score as a causal reasoning task. These aggregation decisions have huge implications

for the conclusions that are drawn about system capabilities, yet are seldom considered carefully or explained.

Aggregate metrics depend not only on the capability of the system but also on the characteristics of the instances used for evaluation. If the gender classification system above were re-evaluated by using entirely light-skinned faces, accuracy would skyrocket, even though the system's ability to classify faces has not changed. Aggregate metrics can easily give false impressions about capabilities when a benchmark is not well constructed.

Problems and trade-offs that arise when considering aggregate versus granular data and metrics are not specific to AI, but they are exacerbated by the challenges inherent in AI research and the research practices of the field. For example, machine learning evaluations usually involve randomly splitting data into training, validation, and test sets. An enormous amount of data is required to train state-of-the-art systems, so these datasets are often poorly curated and lack the detailed annotation necessary to conduct granular analyses. In addition, the research culture in AI is centered around outdoing the current state-of-the-art performance, as evidenced by the many leaderboards, competitions, and challenges that offer prestige or monetary prizes (*5*). This research culture emphasizes aggregate metrics and incentivizes immediate publication of new findings at the expense of robust evaluation practices. In addition, the strict space restrictions and fast turnaround times enforced by high-impact AI conferences disincentivize researchers from reporting results in a granular way. Finally, the primary focus of most publications in AI is not the experimental results themselves but the new algorithms or techniques being evaluated. As a result, less attention has been paid in AI to issues around experimental design and reporting than in other fields such as psychology or physics.

INSTANCE-BY-INSTANCE EVALUATION

There are many situations in which the community might want to conduct analyses that go beyond those reported in a paper. For example, researchers often seek to investigate the extent to which AI systems are biased against minority or disadvantaged populations. It is also frequently useful to scrutinize patterns of performance to debug systems or to determine their safety in a particular deployment context. Moreover, in areas such as robotics and reinforcement learning, examining the trajectory of a system or its sequence of actions can help researchers better understand a system's strategy.

These supplemental evaluations often require access to the instance-by-instance evaluation results (i.e., the outputs and scores of the systems for each instance). But these results are rarely made available—one recent analysis found that only 4% of papers in top AI venues fully report the evaluation results (*1*). As systems and benchmarks continue to grow in size and complexity, it is becoming increasingly costly for researchers to recreate results by conducting their own evaluations. The result is that researchers and policy-makers are increasingly forced to take reported results at face value or to incur substantial and unnecessary costs just to recreate them.

A PATH FORWARD

To address these problems, we propose a broad set of solutions (see guidelines text box).

Guidelines for robust evaluation practices

We recommend that researchers and organizations adopt these guidelines to make it easier for the community to understand system capabilities and conduct follow-up analyses.

1. Wherever possible, reporting of system performance should be granular, with breakdowns across features of the problem space that have been either hypothesized or empirically shown to affect performance. Aggregation decisions should be clearly explained, and analyses conducted to explore system performance should be described.

2. Benchmarks should be designed to test specific capabilities and to systematically vary on important features of the problem space. Benchmark instances should be annotated to allow for granular analyses to be conducted.

3. All recorded evaluation results (e.g., success or failure, response time, partial or full trace, cumulative reward) for each system on each instance should be made available. These data can be reported in supplementary materials or uploaded to a public repository. In cases of cross validation or hyper-parameter optimization, results should ideally be reported for each run and validation split separately.

4. To enable researchers to conduct follow-up analyses, information about each test instance used in an evaluation should be made available, including data labels and all annotated features of those instances.

Moving beyond aggregate metrics

It is important that in-depth performance breakdowns are presented instead of, or alongside, aggregate metrics. Breakdowns can be created by identifying features of the problem space that might be relevant to performance and using those features to analyze, visualize, and predict performance (*6*). These kinds of granular analyses are not yet widely employed. However, researchers focused on system robustness and fairness have begun to demonstrate how valuable they can be.

For example, granular analyses can help researchers explore the concepts that a system has learned. Researchers examined the patterns of errors made by systems on a spatial reasoning benchmark designed to vary systematically on important features of the problem space (7). The researchers found that the systems performed much worse on problems involving the concept of "boundary" than on problems involving the concepts of "top" or "bottom," suggesting that the abstract concept of boundary is more difficult for systems to learn.

Instance-level analyses are essential for identifying "Clever Hans" phenomena in which a system can perform well by relying on unintended patterns in the dataset. For example, a computer vision system that was excellent at classifying images into categories such as "ship" or "horse" (8) was ultimately shown to have not really learned to identify ships or horses. Instead, it had learned to distinguish categories based on the surrounding background or watermarks naming the source of the image—features that the system could not rely on in the real world.

Granular analyses can also allow for more meaningful comparisons between AI systems and humans. For example, though an AI system was better overall at breast cancer screening than six human radiologists (*9*), an in-depth error breakdown showed that the AI system failed to detect various cancers that were detected by all six radiologists. These errors could not be easily explained by the researchers, suggesting that further investigation is needed to understand how the system detects cancers and why it failed in these cases. These findings demonstrate the complementary value of human and AI screenings in a way that aggregate metrics could not.

Of course, granular analysis approaches are not without their challenges. Annotating the features of each instance can be labor intensive. Granular reporting usually needs more space in publications, although detailed breakdowns can be provided in supplementary materials or

online. Granular analyses also often involve slicing the data into smaller pieces, so care must be taken to ensure that findings are not simply artifacts of the data. These issues can be avoided by including a large and diverse range of instances, controlling for multiple comparisons, and specifying important features a priori where possible. Finally, deciding which features to include in performance breakdowns takes time, thought, and expertise. Rigorous theoretical and empirical work may be needed to build an understanding of the problem space. Data-driven approaches can help with this process—regression analyses or deep-learning models can be used to identify features that are predictive of performance.

Ultimately, the best way of presenting evaluation results will depend on the context. The costs of performing granular analyses must be weighed against the potential consequences of system failures, and there is no one-size-fits-all solution. But we think that a shift toward more granular reporting would be a win-win situation for developers and the wider community in most subdisciplines of AI. Granular analyses would help developers pinpoint system weaknesses to guide improvements and avoid potentially catastrophic failures. Granular reporting would ease the evaluation burden that is currently placed on external groups such as algorithmic justice organizations, which often lack the resources and expertise to evaluate systems in detail.

These changes to reporting must go hand in hand with changes to how benchmark tasks are constructed. How well a system performs across different situations cannot be evaluated unless the benchmark comprehensively covers the problem space. The commonly used approach of collating a large dataset and randomly splitting it to create a test set does not assure this coverage, so alternative approaches to benchmark construction must be considered. For example, one can design tasks that test for specific concepts or cognitive abilities and ensure that task instances systematically vary on important features of the problem space (*7*, *10, 11*).

In this endeavor, techniques such as procedural or adversarial generation of task instances might be useful.

Bringing about these changes in research culture will require the participation and support of the entire community. Within academia, we recommend that publications report granular analyses wherever possible, and that reviewers and editors ask for performance breakdowns when they are not provided. It might also be valuable to alter the space limits in conference publications to enable in-depth reporting of evaluation results. More broadly, we think the field needs to reckon with the potentially detrimental effects of leaderboards and competitions on evaluation practices.

We also recommend that private organizations consider incorporating guidelines for granularity and aggregation into their wider evaluation and reporting practices (*12*). We are encouraged to see growing support for "model cards" that explain a system's architecture and justify evaluation decisions (*13*). Policy-makers should bear in mind the need for granular analyses when creating guidelines or safety standards for specific applications. For instance, the recently proposed Minimum Information for Medical AI Reporting (MINIMAR) standard could be modified to ask for explanations of aggregation decisions and performance breakdowns across features of the problem space.

Ensuring the availability of instance-by-instance evaluation results

A growing open-science movement has laid the groundwork for moving toward making instance-by-instance evaluation more common. Drawing on lessons from multiple disciplines, many have argued that a lack of transparency and reproducibility in AI research could stifle progress and lead to dangerous misestimations of AI capabilities (*5, 14, 15*). In response, various initiatives have been set up to promote code and data sharing, such as the Hugging Face Hub, the Machine Learning Open Source Software section of the *Journal of Machine Learning*

Research, GitHub, OpenML, Papers With Code, and the Open Science Framework. However, instance-by-instance evaluation results are rarely included on these platforms. There have been few incentives to put in the extra work needed to clean, document, release, and maintain these results. Furthermore, many researchers fear that subsequent analyses might discover flaws or biases in their systems, or that they will be "scooped" by other researchers (*5*).

It is vital to incentivize the release of instance-by-instance results. In other disciplines, various kinds of requirements, incentives, and nudges have been implemented for similar purposes. For example, changes to journal and conference reporting guidelines to include instance-by-instance evaluation results could help encourage the sharing of these results. It should also be possible to ensure that researchers are credited for subsequent uses of their results—perhaps by giving the results a unique identifier that other researchers can cite.

Given that many AI systems are developed or deployed by non-academic organizations, it is important to consider how policy-makers and industry organizations can encourage the sharing of results. For example, funding agencies could require the release of instance-by-instance results as a condition of funding, and private organizations could be encouraged to share instance-by-instance results whenever they publish preprints or press releases involving system evaluations. These solutions would complement wider efforts of groups such as the European Centre for Algorithmic Transparency to encourage transparency in AI. We recognize that there are some situations in which instance-by-instance results cannot be released (e.g., owing to privacy concerns or practical constraints), but in most cases it should be possible to do so. Even if no features were annotated, releasing instance-by-instance results would still allow other researchers to extract features themselves and perform additional analyses as long as the benchmark or test data are obtainable.

Some successful examples of results-sharing in AI give us confidence that these changes in broader research culture are possible. For example, researchers who developed the Holistic

Evaluation of Language Models (HELM) benchmark made instance-by-instance results available for a variety of models across the entire benchmark. If other fields such as psychology and medicine can make progress on these issues even in the face of considerable data privacy challenges, AI should be able to do the same.

REFERENCES AND NOTES

1. O. E. Gundersen, S. Kjensmo, *Proceedings of the AAAI Conference on Artificial Intelligence* (Association for the Advancement of Artificial Intelligence, 2018), vol. 32, pp. 1644–1651.
2. B. Nushi, E. Kamar, E. Horvitz, *Proceedings of the Sixth AAAI Conference on Human Computation and Crowdsourcing* (Association for the Advancement of Artificial Intelligence, 2018), vol. 6, pp. 126–135.
3. J. Buolamwini, T. Gebru, Proceedings of the 1st Conference on Fairness, Accountability and Transparency (PMLR, 2018), vol. 81, pp. 77–91.
4. A. Srivastava *et al*., arXiv:2206.04615 (2022); http:// arxiv.org/abs/2206.04615.
5. J. Pineau *et al*., *J. Mach. Learn. Res*. 22, 1 (2021).
6. R. Burnell et al., in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, Vienna, 23 to 29 July 2022 (International Joint Conferences on Artificial Intelligence), pp. 2827–2835.
7. V. V. Odouard, M. Mitchell, arXiv:2206.14187 (2022); https://arxiv.org/abs/2206.14187.
8. S. Lapuschkin *et al*., *Nat. Commun*. 10, 1096 (2019).
9. S. M. McKinney *et al*., *Nature* 577, 89 (2020).

10. D. Kiela et al., in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Association for Computational Linguistics, 2021), pp. 4110–4124.
11. J. Z. Leibo et al., in *Proceedings of the 38th International Conference on Machine Learning*, 18 to 24 July 2021 (PMLR), vol. 139, pp. 6187–6199.
12. I. D. Raji *et al*., arXiv:2001.00973 (2020); https://arxiv. org/abs/2001.00973.
13. M. Mitchell et al., in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2019), pp. 220–229.
14. B. Haibe-Kains *et al*., *Nature* 586, E14 (2020).
15. M. Hutson, *Science* 359, 725 (2018).

ACKNOWLEDGMENTS

We are grateful to J. Pineau and S. Russell for their feedback on the manuscript and to M. Brundage and L. Ahmad for helpful discussions about the need for instance-by-instance results for foundation models. Funding support was provided by the Future of Life Institute under grant RFP2-152 (J.H.O. and J.B.); US Defense Advanced Research Projects Agency HR00112120007 (RECoG-AI) (J.H.O., L.C., R.B., J.B., and D.K.); grant PID2021-122830OB-C42 (SFERA) funded by MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe" (J.H.O., F.M.P., and W.S.); INNEST/2021/317 [Project cofunded by the European Union with the "Programa Operativo del Fondo Europeo de Desarrollo Regional (FEDER) de la Comunitat Valenciana 2014-2020"] (J.H.O. and W.S.); and UPV (Vicerrectorado de Investigación) grant PAI-10-21 (J.H.O. and W.S.).