

Preface for the IJCAI-22 Workshop on AI Evaluation Beyond Metrics (EBeM)

José Hernández-Orallo^{1,2,5}, Lucy Cheke¹, Joshua Tenenbaum³, Tomer D. Ullman⁴, Fernando Martínez-Plumed¹, Danaja Rutar², John Burden^{2,5}, Ryan Burnell² and Wout Schellaert¹

¹Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, Valencia, Spain

²Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, United Kingdom

³Massachusetts Institute of Technology, Department of Brain and Cognitive Sciences, Cambridge, United States of America

⁴Harvard University, Department of Psychology, Cambridge, United States of America

⁵Centre for the Study of Existential Risk, University of Cambridge, Cambridge, United Kingdom

Abstract

We summarize the IJCAI-22 Workshop on Artificial Intelligence Evaluation Beyond Metrics, held at the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-ECAI) on July 24.

1. Introduction

Traditional approaches to AI evaluation lack the necessary robustness to analyse the capabilities of complex AI systems. Many AI systems solve a task or excel at a particular benchmark, but then fail at other tasks or instances that putatively represent the same capability. This problem becomes more blatant as performance in many AI benchmarks ramps up rapidly, while concomitant improvement in capability is moderate, or limited to specific areas (e.g., language models), which are assessed informally and with low robustness. This generates a large disparity between the “hype” of AI achievement and the reality, and contributes to a feeling of distrust about what AI is really capable of. It also makes it difficult to predict the potential operational applications of AI in the future, especially as AI systems become less task-specific and more general-purpose, as it is happening with language models.


For metrics to be really useful they have to be meaningful. Meaningful assessment must measure well defined capabilities that translate into predictable skills and applications. Here many lessons can be taken from psychology. Current performance indicators are used to compare AI systems for the same benchmark or domain, and leaderboards are used to extrapolate progress in the field. But because most of these benchmarks are not driven by theory, they have very limited construct validity or generalisability. For this reason, there is still much uncertainty

EBeM'22: Workshop on AI Evaluation Beyond Metrics, July 24, 2022, Vienna, Austria

✉ jorallo@upv.es (J. Hernández-Orallo); lgc23@cam.ac.uk (L. Cheke); jbt@mit.edu (J. Tenenbaum); tullman@fas.harvard.edu (T. D. Ullman); fmartinez@dsic.upv.es (F. Martínez-Plumed); dr571@cam.ac.uk (D. Rutar); jjb205@cam.ac.uk (J. Burden); rb967@cam.ac.uk (R. Burnell); wschell@vrain.upv.es (W. Schellaert)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

over how to assess and monitor the state, development, uptake and impact of AI as a whole, including its future evolution, progress and capabilities.

The IJCAI-ECAI-22 Workshop on Artificial Intelligence Evaluation Beyond Metrics (EBeM 2022) seeks to challenge the widespread yet limited approach of evaluating the performance of intelligent systems with aggregated metrics over a benchmark or distribution of tasks. In particular, we discuss further alternative approaches that draw on ideas and recent progress in cognitive and developmental psychology, psychometrics, software testing, and other areas. The following are the main topics of the workshop:

- Evaluation methods founded on cognitive, developmental or comparative psychology
- Measurement of skills, capabilities, or cognitive abilities
- Evaluation methods based on software testing or other engineering practices
- Meta-analysis or comparisons of evaluation instruments
- The role of evaluation in AI development, policy making, and modeling of social impact
- Measurements of generality or common-sense
- Capture and use of evaluation data
- Analysis of the task space and its relation to corresponding capabilities
- The role of causality in evaluation
- Topics complementary to evaluation such as documentation or auditing
- Alternative evaluation methods with added benefits
- Discussion and progress in hard to evaluate scenarios

These topics aim to achieve a holistic view of AI evaluation, targeting people from cognitive science, comparative psychology, neuroscience, psychometrics, philosophy of science and technology, measurement theory, policy, etc., with the ultimate goal of encouraging cross-disciplinary approaches and theoretical and experimental analysis of how AI evaluation should be done at present and in the future.

2. Program

The Program Committee (PC) received 16 submissions. Each paper was peer-reviewed by at least two PC members (with an average of 3 reviewers), by following a single-blind reviewing process. The committee decided to accept 13 full papers, of which 3 papers elected not to be published in this proceedings. The EBeM 2022 program was organized in to four sessions, with two invited speakers, two panels, and a special session by the OECD.

- Technical Session 1
 - Invited speaker: Amanda Seed
 - Paper presentations
- Technical Session 2
 - Paper presentations
 - Panel: Cognitive Evaluation with the Animal AI Environment

- Technical Session 3
 - Invited speaker: Adina Williams
 - Special OECD Session: Artificial Intelligence and the Future of Skills (AIFS)
- Technical Session 4
 - Paper presentations
 - Panel: Evaluating pre-trained, generative, and prompted systems

3. Acknowledgements

We thank all researchers who submitted papers to EBeM 2022 and congratulate the authors whose papers were selected for inclusion into the workshop program and proceedings. We thank all invited speakers and panel members, and our distinguished PC members for reviewing the submissions and providing useful feedback to the authors:

- Atia Cortés - Barcelona Supercomputing Center
- Alex Taylor - University of Auckland
- Alex Wang - New York University
- Celeste Kidd - University of California Berkeley
- Craig S. Greenberg - NIST
- David Fernández-Llorca - European Commission, JRC
- Deborah Raji - Mozilla
- Ellen Voorhees - NIST
- Ernest Davis - New York University
- Guillaume Avrin - Lab. Nat. de Métrologie et d'Essais
- Isabelle Hupont-Torres - European Commission, JRC
- an Feyereisl - GoodAI
- Joel Leibo - DeepMind
- Kevin Smith - MIT
- Koustuv Sinha - McGill University
- Ljerka Ostojic - University of Rijeka
- Melanie Mitchell - Santa Fe Institute
- Moira Dillon - New York University
- Naman Shukla - Deepair Solutions
- Panos Ipeirotis - New York University
- Peter Flach - University of Bristol
- Raul Santos-Rodriguez - University of Bristol
- Ricardo Prudencio - Informatics Center, UFPE
- Ricardo Vinuesa - KTH Royal Institute of Technology
- Richard Mallah - Future of Life Institute
- Rotem Dror - University of Pennsylvania

- Sean Holden - University of Cambridge
- Sebastian Gehrmann - Google Research
- Songul Tolan - European Commission, JRC
- Tadahiro Taniguchi - Ritsumeikan University
- Vicky Charisi - European Commission, JRC