Measuring (machine) intelligence universally. An interdisciplinary challenge

José Hernández-Orallo
Dep. of Information Systems and Computation
Universitat Politècnica de València
València, Spain
jorallo@dsic.upv.es

Abstract—Artificial intelligence (AI) is having a deep impact on the way humans work, communicate and enjoy their leisure time. AI systems have been traditionally devised to solve specific tasks, such as playing chess, diagnosing a disease or driving a car. However, more and more AI systems are now being devised to be generally adaptable, and learn to solve a variety of tasks or to assist humans and organisations in their everyday tasks. As a result, an increasing number of robots, bots, avatars and 'smart' devices are enhancing our capabilities as individuals, collectives and humanity as a whole. What are these systems capable of doing? What is their global intelligence? How to tell whether they are meeting their specifications? Are the organisations including AI systems being less predictable and difficult to govern? The truth is that we lack proper measurement tools to evaluate the cognitive abilities and expected behaviour of this variety of systems, including hybrid (e.g., machine-enhanced humans) and collectives. Once realised the relevance of AI evaluation and its difficulty, we will survey what has been done in the past twenty years in this area, focussing on approaches based on algorithmic information theory and Kolmogorov complexity, and its relation to other disciplines that are concerned with intelligence evaluation in humans and animals, such as psychometrics and comparative cognition. This will lead us to the notion of universal intelligence test and the new endeavour of universal psychometrics.

Artificial intelligence; intelligence evaluation; universal psychometrics; Kolmogorov complexity.

I. Introduction

Many artificial intelligence (AI) systems have specialised applications: computer vision, speech recognition, music analysis, machine translation, text summarisation, information retrieval, robotic navigation and interaction, automated vehicles, game playing, prediction, estimation, planning, automated deduction, expert systems, etc. (see, e.g., [78]). These applications automate many tasks that were thought to require intelligence, and were previously done only by humans. Some tasks AI systems solve today cannot even be performed by humans, or the best human performance is worse than the best AI system (e.g., chess, information retrieval, etc.). Fortunately, we do not see this as a competition between humans and machines, but rather as a way of enhancing human possibilities. Today, individuals and organisations have access to a large set of AI systems that enhance what they can do, such as filtering the irrelevant information from our inbox, recommending a new film to see or reminding us what to do after a meeting.

Most of the above applications depend on well-defined, specific tasks. The evaluation of AI systems doing these tasks is not always easy, but it is still relatively straightforward to define a performance metric for the task. Many benchmark problems and competitions have arisen in the past decades to evaluate AI systems [10]. From them, we can see the progress of AI in these specialised applications.

However, there is an increasing recent interest in AI systems that do not solve a predefined task, but that are able to solve problems they have never faced before and have never been programmed for. A new plethora of assistants, adaptive robots and other kinds of *intelligent* systems are being designed to cover a wider range of problems. Systems are not programmed to solve a task, but learn to solve a task. As a result, we are beginning to experience a human-machine interaction where machines are not programmed but *taught*. However, this generality has a price. How predictable are these systems? How intelligent are they? What are they able to do?

II. Ability-oriented Evaluation

Ability-oriented evaluation is much more difficult than task-oriented evaluation. First, what is an ability? This question has been addressed in the context of human intelligence by psychometrics. Second, can we measure the cognitive abilities of machines using human cognitive tests? This has been advocated several times in the past [1], even if intelligence tests have been designed for humans and not for machines. In fact, some non-intelligent computer programs have been able to score reasonably well in some IQ tests [19] (for a full discussion about this, see [3]).

A very different approach to intelligence evaluation has been based on algorithmic information theory (AIT) and the related notions of Solomonoff universal probability [20], Kolmogorov complexity [17] and Wallace's Minimum Message Length (MML) principle [21]. These ideas have helped to develop a variant of the Turing Test featuring compression [2], an intelligence test derived from letter sequences of various Kolmogorov complexity [6][15], tests for other cognitive abilities [7][8] and tests where agents are placed in environments with rewards and penalties [16], in a way that resembles animal intelligence evaluation.

All this has led to the notion of *universal* intelligence test, i.e., a test that can be applied to humans, non-human animals, machines, hybrids and collectives [4][9][11]. By extending this to a wider range of cognitive abilities we face an even more difficult and interdisciplinary challenge, known as universal psychometrics [14], borrowing ideas from the approach based on AIT, from psychometrics and from comparative cognition.

III. Discussion

At the end of this talk, we will discuss on the evaluation of systems that have many agents, either artificial or biological, i.e., social environments [5][13]. The evaluation of social intelligence as well as how intelligence and other abilities evolve with time [12] are questions that will be required in a context where organisations and collectives will be composed of both humans and machines of diverse capabilities, as well as hybrids (e.g., machine-enhanced humans). Their behaviour, potential and even personality will have to be evaluated. This will be key to ensure that collectives and organisations will be more predictable and governable.

References

[1] D. K. Detterman. A challenge to Watson. *Intelligence,* 39(2-3):77 – 78, 2011.

[2] D. L. Dowe and A. R. Hajek. A non-behavioural, computational extension to the Turing Test. In *Intl. Conf. on Computational Intelligence & multimedia applications* (ICCIMA'98), Gippsland, Australia, pages 101–106, 1998.

[3] D. L. Dowe and J. Hernández-Orallo. IQ tests are not for machines, yet. *Intelligence*, 40(2):77–81, 2012.

[4] D. L. Dowe and J. Hernández-Orallo. How universal can an intelligence test be? *Adaptive Behavior*, 22(1):51–69, 2014.

[5] D. L. Dowe, J. Hernández-Orallo, and P. K. Das. Compression and intelligence: social environments and communication. In J. Schmidhuber, K.R. Thórisson, and M. Looks, editors, *Artificial General Intelligence*, volume 6830, pages 204–211. LNAI series, Springer, 2011.

[6] J. Hernández-Orallo. Beyond the Turing Test. *J. Logic, Language & Information*, 9(4):447–466, 2000.

[7] J. Hernández-Orallo. On the computational measurement of intelligence factors. In A. Meystel, editor, *Performance metrics for intelligent systems workshop*, pages 1–8. National Institute of Standards and Technology, Gaithersburg, MD, U.S.A., 2000.

[8] J. Hernández-Orallo. Computational measures of information gain and reinforcement in inference processes. *AI Communications,* 13(1):49–50, 2000.

[9] J. Hernández-Orallo. A (hopefully) non-biased universal environment class for measuring intelligence of biological and artificial systems. In M. Hutter et al., editor, *Artificial General Intelligence*, 3rd Intl Conf, pages 182–183. Atlantis Press.

[10] J. Hernández-Orallo "AI Evaluation: past, present and future" *arXiv*:1408.6908

[11] J. Hernández-Orallo and D. L. Dowe. Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence*, 174(18):1508 – 1539, 2010. 20

[12] J. Hernández-Orallo and D. L. Dowe. On potential cognitive abilities in the machine kingdom. *Minds and Machines*, 23:179–210, 2013. 22

[13] J. Hernández-Orallo, D. L. Dowe, S. España-Cubillo, M. V. Hernández-Lloreda, and J. Insa-Cabrera. On more realistic environment distributions for defining, evaluating and developing intelligence. In J. Schmidhuber, K.R. Thórisson, and M. Looks, editors, *Artificial General Intelligence*, volume 6830, pages 82–91. LNAI, Springer, 2011.

[14] J. Hernández-Orallo, D. L. Dowe, and M. V. Hernández-Lloreda. Universal psychometrics: Measuring cognitive abilities in the machine kingdom. *Cognitive Systems Research*, 27:5074, 2014.

[15] J. Hernández-Orallo and N. Minaya-Collado. A formal definition of intelligence based on an intensional variant of Kolmogorov complexity. In Proc. *Intl Symposium of Engineering of Intelligent Systems* (EIS'98), pages 146–163. ICSC Press, 1998.

[16] S. Legg and M. Hutter. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4):391–444, 2007.

[17] M. Li and P. Vitányi. *An introduction to Kolmogorov complexity and its applications* (3rd ed.). Springer-Verlag, 2008.

[18] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2009.

[19] P. Sanghi and D. L. Dowe. A computer program capable of passing IQ tests. In *4th Intl. Conf. on Cognitive Science* (ICCS'03), Sydney, pages 570–575, 2003.

[20] R. J. Solomonoff. A formal theory of inductive inference. Part I. *Information and control*, 7(1):1–22, 1964.

[21] C. S. Wallace and D. M. Boulton. An information measure for classification. *Computer Journal*, 11(2):185–194, 1968.