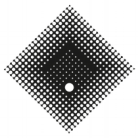

A New AI Evaluation Cosmos: Ready to Play the Game?

Hernández-Orallo, J, Baroni, M, Bieger, J, Chmait, Nader, Dowe, DL, Hofmann, K, Martínez-Plumed, F, Strannegård, C and Thórissons, KR (2017) A New AI Evaluation Cosmos: Ready to Play the Game? AI Magazine, 38 (3). pp. 66-69. ISSN 0738-4602

A New AI Evaluation Cosmos: Ready to Play the Game?

José Hernández-Orallo, Marco Baroni, Jordi Bieger, Nader Chmait, David L. Dowe, Katja Hofmann, Fernando Martínez-Plumed, Claes Strannegård, Kristinn R. Thórisson

■ *We report on a series of new platforms and events dealing with AI evaluation that may change the way in which AI systems are compared and their progress is measured. The introduction of a more diverse and challenging set of tasks in these platforms can feed AI research in the years to come, shaping the notion of success and the directions of the field. However, the playground of tasks and challenges presented there may misdirect the field without some meaningful structure and systematic guidelines for its organization and use. Anticipating this issue, we also report on several initiatives and workshops that are putting the focus on analyzing the similarity and dependencies between tasks, their difficulty, what capabilities they really measure and – ultimately – on elaborating new concepts and tools that can arrange tasks and benchmarks into a meaningful taxonomy.*

AI experimentation and evaluation

Through the integration of more and better techniques, more computing power, and the use of more diverse and massive sources of data, AI systems are becoming more flexible and adaptable, but also more complex and unpredictable. There is thus increasing need for a better assessment of their capacities and limitations, as well as concerns about their safety [1]. Theoretical approaches might provide important insights, but only through experimentation and evaluation tools will we achieve a more accurate assessment of how an actual system operates over a series of tasks or environments.

Several AI experimentation and evaluation platforms have recently appeared, setting a new "cosmos" of AI environments. These facilitate the creation of various tasks for evaluating and training a host of algorithms. The platform interfaces usually follow the reinforcement learning (RL) paradigm, where interaction takes place through incremental observations, actions, and rewards. This is a very general setting and seemingly every possible task can be framed under it.

These platforms are different from the Turing Test - and other more traditional AI evaluation benchmarks proposed to replace it – as summarized by an AAAI 2015 workshop and a recent special issue of the AI Magazine. Actually, some of these platforms can integrate any task and hence in principle they supersede many existing AI benchmarks [2] in their aim to test "general problem solving ability".

This is also a new arena for the Tech Giants, which has attracted mainstream attention. For instance, the *Nature* journal has recently featured a news article on the topic [3]. All in all, a new and uncharted territory for AI is emerging, which deserves more attention and effort within AI research itself.

In this report, we first give a short overview of the new platforms, and briefly report about two 2016 events focusing on (general-purpose) AI evaluation (using these platforms or others).

New playground, new benchmarks

Many different general-purpose benchmarks and platforms have recently been introduced, and they are increasingly adopted in research and competitions to drive and evaluate AI progress.

The **Arcade Learning Environment**, is a platform for developing and evaluating general AI agents using a variety of Atari 2600 games. The

platform is used to compare, among others, approaches such as RL (see, e.g., [4]), model learning, model-based planning, imitation learning and transfer learning. A limitation of this environment is the reduced number of games, leading to overspecialization. The **Video Game Definition Language** (VGDL) follows a similar philosophy, but new 2D arcade games can be generated using a flexible set of rules.

OpenAI **Gym** [5] provides a diverse collection of RL tasks and an open-source interface for agents to interact with them, as well as tools and a curated web service for monitoring and comparing RL algorithms. The environments, formalized as partially observable Markov decision processes, range from classic control and toy text to algorithmic problems, 2D and 3D robots, as well as Doom, board and Atari games.

OpenAI **Universe** is a software platform intended for training and measuring the performance of AI systems on any task where a human can complete with a computer, and in the way a human does: looking at screen pixels and operating a (virtual) keyboard and mouse. In Universe, any program can be turned into a Gym environment, including Flash games, browser tasks, and games like *slither.io* and *GTA V*. The current release consists of 1000 environments ready for RL.

Microsoft's **Project Malmo** [6] gives users complete freedom to build complex 3D environments within the block-based world of the Minecraft video game. It supports a wide range of experimentation scenarios for evaluating RL agents and provides a playground for general AI research. Tasks range from navigation and survival to collaboration and problem solving.

GoodAI's **Brain simulator** and school is a collaborative platform to simulate artificial brain architectures using existing AI modules, like image recognition and working memory.

DeepMind Lab is a highly customisable and extensible 3D game-like platform for agent-based AI research. Agents operate in 3D environments using a first-person viewpoint and can be evaluated over a wide range of planning and strategy tasks, from maze navigation to playing laser tag. Somewhat similarly, the **ViZDoom** [7] research platform allows RL agents to interact with customizable scenarios in the world of the 1993 first-person shooting video game "Doom" using only the screen buffer.

Facebook's **TorchCraft** [8] is a library enabling machine learning research on real-time strategy games. These games' high-dimensional action space is quite different from those previously investigated in RL research and provides a useful bridge to the richness of the real-world. To execute something as simple as "attack this enemy base", one must coordinate mouse clicks, camera, and available resources. This makes actions and planning hierarchical, which is challenging in RL. TorchCraft's current implementation connects the Torch machine learning library to StarCraft: Brood War, but the same idea can be applied to any video game and library. Meanwhile, DeepMind is also collaborating with Blizzard Entertainment to open up **StarCraft II** as a testing environment for AI research.

Facebook's **CommAI-env** [9] is a platform for training and evaluating AI systems from the ground up, to be able to interact with humans via language. An AI learner interacts in a communication-based setup via a bit-level interface with an environment that asks the learner to solve tasks presented with incremental difficulty. Some tasks currently implemented include counting problems, memorizing lists and answering questions about them, and navigating from text-based instructions.

The introduction of all these platforms offers many new possibilities for AI evaluation and experimentation, but it also poses many questions about how benchmarks and competitions can be created using these platforms, especially if the goal is to assess more general AI. Two new venues were set up to explore these issues in 2016, as we discuss next.

1st Workshop on Evaluating General-Purpose AI 2016

EGPAI 2016 was the first workshop focusing on the Evaluation of General-Purpose Artificial Intelligence. It was a satellite workshop of ECAI, the 22nd European Conference on AI, held in August 2016. EGPAI 2016 promoted several discussions on (General) Artificial Intelligence and looked into state-of-the-art research questions such as: "Can the various tasks and benchmarks in AI provide a general basis for evaluation and comparison of a broad range of such systems?", "Can there be a theory of tasks, or cognitive abilities, enabling a more direct comparison and characterization of AI systems?", and "How does the specificity of an AI agent relate to how fast it can approach optimal performance?"

The most relevant outcome of this workshop was the identification of the challenging and urgent demands relevant to general-purpose AI evaluation, such as understanding the relation between tasks (or classes of tasks), the notion of (task and environment) difficulty, and the relevance of how observations are presented to AI agents, including rewards and penalties. The workshop also served to illustrate how several algorithms compare in terms of their generality.

The Machine Intelligence Workshop at NIPS 2016

The Machine Intelligence Workshop at NIPS (December 2016) focused on the parallel questions of what is general AI and how to evaluate it. Concerning evaluation, there was a general agreement that we need to test systems for their ability to tackle new tasks that they did not encounter in their training phase. The speakers also agreed that an important characteristic to be tested is the degree to which systems are "compositional", in the sense that they can creatively recompose skills that they have learned in previous tasks to solve a new problem.

Some speakers argued for tasks to be defined from first principles in a top-down manner, whereas others suggested looking at nature (humans and other intelligent beings) for inspiration in formulating the tasks (with further discussion on whether the inspiration should come from ontogenesis or phylogenesis).

The role of human language was also debated, with some speakers stressing that it is hard to conceive of useful AI without a linguistic communication channel, while others pointed to animal intelligence as a more realistic goal, and to possible applications for non-linguistic AI.

AI and evaluation — the future

A recurrent issue in general intelligence evaluation is based on the old view of intelligence as the capability to succeed in a range of tasks or, ultimately, performing relatively well in all possible tasks. Nevertheless, the notion of "all tasks" is meaningless if the concept is not accompanied by a probability distribution. While [10] advocate a distribution based on Solomonoff's universal prior on task descriptions (higher probability to tasks of short encoding), [11] advocates a distribution based on task difficulty (measuring difficulty as the complexity of the simplest solution for each task, and ensuring *solution* diversity for each difficulty). Alternative distributions could be derived from the set of tasks that humans and other animals face on a daily basis.

When compared to these theoretical distributions, can we say anything about the distribution of tasks that compose any of the new platforms? Is their actual diversity really covering general abilities? And what about their properties with respect to transfer, or gradual, learning?

As more tasks are integrated, different "universes" of tasks are created and the whole set of tasks in all platforms configure the cosmos for AI. At present, this is just an unstructured collection of tasks with no clear criteria for inclusion, exclusion or relative weight. This bears similarity to the early years of psychometrics (among other disciplines) that have been dealing with behavioral evaluation for over a century, putting some order in the space of tasks and abilities.

To move ahead, the space of tasks must be analyzed. This can be done in terms of a hierarchy linking tasks and abilities [11] or in terms of a "task theory" [12], using theoretical approaches to task similarity and difficulty, or a more empirical strategy, by analyzing the results of a population of AI systems with Item Response Theory (IRT) or other psychometric techniques [13].

In summary, evaluation is becoming crucial in AI and will become much more sophisticated and relevant in the years to come. New events in 2017, including challenges (e.g., the General AI challenge), competitions and workshops (e.g., EGPAI2017 at IJCAI2017), will delve much further into how general-purpose AI should be evaluated now and in the future.

Marco Baroni is Research Scientist in the Facebook Artificial Intelligence Research laboratory.

Jordi Bieger is a PhD student at Reykjavik University, Iceland.

Nader Chmait is a PhD student at Monash University, Australia.

David L. Dowe is Associate Professor at Monash University, Australia.

José Hernández-Orallo is Professor at Universitat Politècnica de València, Spain.

Katja Hofmann is a researcher at Microsoft Research in Cambridge, UK.

Fernando Martínez-Plumed is a postdoctoral researcher at Universitat Politècnica de València, Spain.

Claes Strannegård is Associate Professor at Chalmers University of Technology, Sweden.

Kristinn R. Thórisson is Professor of Computer Science at Reykjavik University in Iceland, and Managing Director of the Icelandic Institute for Intelligent Machines.

References

[1] Amodei, D.; Olai, C.: Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. 2016. "Concrete problems in AI safety", arXiv preprint arXiv:1606.06565.

[2] Hernández-Orallo, J. 2016. Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement, Artificial Intelligence Review.

[3] Castelvecchi, D. 2016., Tech giants open virtual worlds to bevy of AI programs. Nature Vol 540, Issue 7633, pp. 323-324.

[4] Mnih, V. et al. 2015. Human-level control through deep reinforcement learning. Nature Vol 518, pp. 529–533.

[5] Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; Zarembas, W. (2016. OpenAI Gym, arXiv.1606.01540v1.

[6] Johnson, M.; Hofmann K.; Hutton, T.; Bignell, D. 2016. The Malmo platform for artificial intelligence experimentation, IJCAI, pp. 4246-4247.

[7] Kempka, M.; Wydmuch, M.; Runc, G.; Toczek, J.; Jaśkowski, W. 2016. ViZDoom: A Doom-based AI Research Platform for Visual Reinforcement Learning, arXiv:1605.02097.

[8] Synnaeve, G.; Nardelli, N.; Auvolat, A.; Chintala, S.; Lacroix, T.; Lin, Z.; Richoux, F.; Usunier, N. 2016. TorchCraft: a Library for Machine Learning Research on Real-Time Strategy Games. arXiv 1611.00625

[9] Mikolov, T.; Joulin, A.; Baroni, M. 2015. A Roadmap towards Machine Intelligence, arXiv 1511.08130.

[10] Legg S.; Hutter M. 2007. Universal Intelligence: A Definition of Machine Intelligence, Minds and Machines, Vol. 17, Issue 4, pp 391–444.

[11] Hernández-Orallo, J. 2017. *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*, Cambridge University Press.

[12] Thórisson, K.R.; Bieger, J.; Thorarensen, T.; Sigurðardóttir, J.S.; Steunebrink, B.R. 2016. Why Artificial Intelligence Needs a Task Theory – And What It Might Look Like. In Steunebrink B. R. et al. (eds.), AGI-16, July 16-19, New York, LNCS, Vol. 9782, pp 118-128.

[13]De Ayala, R.J. 2009. The theory and practice of Item Response Theory, Guilford.