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. Thórissons

■ *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.*

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 (Amodei et al. 2016). 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.

  ISSN 0738-4602

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[1] and a recent special issue of the AI Magazine.[2] Actually, some of these platforms can integrate any task and hence in principle they supersede many existing AI benchmarks (Hernández-Orallo 2016) in their aim to test general problem-solving ability.

This topic has also attracted mainstream attention. For instance, the journal *Nature* recently featured a news article on the topic (Castelvecchi 2016). In summary, 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[3] 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, for example, Mnih et al [2015]), 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)[4] follows a similar philosophy, but new two-dimensional (2D) arcade games can be generated using a flexible set of rules.

OpenAI Gym[5] (Brockman et al. 2016) 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 three-dimensional (3D) robots, as well as Doom, board, and Atari games.

OpenAI Universe[6] 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[7] (Johnson et al. 2016) 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[8] and school is a collaborative platform to simulate artificial brain architectures using existing AI modules, like image recognition and working memory.

DeepMind Lab[9] is a highly customizable and extensible 3D gamelike 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 (Kempka et al. 2016) 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 (Synnaeve et al. 2016) is a library enabling machine-learning research on real-time strategy games. The high-dimensional action space of these games 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[10] (Mikolov, Joulin, and Baroni 2015) is a platform for training and evaluating AI systems from the ground up, to be able to interact with humans through language. An AI learner interacts in a communication-based setup through 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 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 such 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.

The Evaluating General-Purpose AI Workshop

The 2016 Workshop on Evaluating General-Purpose AI (EGPAI 2016)[11] was the first workshop focusing on the evaluation of general-purpose artificial intelli-

gence. A satellite workshop of the 22nd European Conference on AI (ECAI) 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

The Machine Intelligence Workshop[12] held at the December 2016 Conference on Neural Information Processing Systems (NIPS 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 nonlinguistic 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 Legg and Hutter (2007) advocate a distribution based on Solomonoff's universal prior on task descriptions (higher probability to tasks of short encoding), Hernández-Orallo (2017) 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 (Hernández-Orallo 2017) or in terms of a task theory (Thórisson et al. 2016), 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 (De Ayala 2009).

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 (such as the General AI challenge[13]), competitions, and workshops, such as the Evaluating General-Purpose AI 2017 workshop[14] at IJCAI 2017), will delve much further into how general-purpose AI should be evaluated now and in the future.

Notes

1. See the AAAI 2015 workshop, Beyond the Turing Test, chaired by Gary Marcus, Francesca Rossi, and Manuela Veloso.

2. See the spring 2016 issue of *AI Magazine*, volume 37, number 1. The 13 special issue articles were edited by Gary Marcus, Francesca Rossi, and Manuela Veloso.

3. See the Arcade Learning Environment website (www.arcadelearningenvironment.org).

4. See www.gvgai.net/vgdl.php.

5. See gym.openai.com.

6. See blog.openai.com/universe.

7. See www.microsoft.com/en-us/research/project/project-malmo.

8. See www.goodai.com/brain-simulator.

9. See deepmind.com/blog/open-sourcing-deepmind-lab.

10. See github.com/facebookresearch/CommAI-env.
11. See www.dsic.upv.es/~flip/EGPAI2016.
12. See mainatnips.github.io.
13. See www.general-ai-challenge.org.
14. See www.dsic.upv.es/~flip/EGPAI2017.

References

Amodei, D.; Olah, C.: Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. 2016. Concrete Problems in AI Safety. arXiv Preprint. arXiv:1606.06565 [cs.AI]. Ithaca, NY: Cornell University Library.

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; Zarembas, W. 2016. OpenAI Gym. arXiv Preprint. arXiv.1606.01540. [cs.LG]. Ithaca, NY: Cornell University Library.

Castelvecchi, D. 2016. Tech Giants Open Virtual Worlds to Bevy of AI Programs. *Nature* 540(7633): 323–324. doi.org/10.1038/540323a

De Ayala, R. J. 2009. *The Theory and Practice of Item Response Theory*. New York: The Guilford Press.

Hernández-Orallo, J. 2016. Evaluation in Artificial Intelligence: From Task-Oriented to Ability-Oriented Measurement. *Artificial Intelligence Review.* Online First Article, 1–51.

Hernández-Orallo, J. 2017. *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge, UK: Cambridge University Press. doi.org/10.1017/9781316594179

Johnson, M.; Hofmann K.; Hutton, T.; Bignell, D. 2016. The Malmo Platform for Artificial Intelligence Experimentation. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (IJCAI-16), 4246–4247. Palo Alto, CA: AAAI Press.

Kempka, M.; Wydmuch, M.; Runc, G.; Toczek, J.; Jasowski, W. 2016. ViZDoom: A Doom-Based AI Research Platform for Visual Reinforcement Learning. arXiv Preprint. arXiv:1605.02097[cs.LG]. Ithaca, NY: Cornell University Library.

Legg S., Hutter M. 2007. Universal Intelligence: A Definition of Machine Intelligence. *Minds and Machines* 17(4): 391–444. doi.org/10.1007/s11023-007-9079-x

Mikolov, T.; Joulin, A.; Baroni, M. 2015. A Roadmap Towards Machine Intelligence. arXiv Preprint. arXiv 1511.08130[cs.AI]. Ithaca, NY: Cornell University Library.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou, I.; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; Hassabis, D. 2015. Human-Level Control Through Deep Reinforcement Learning. *Nature* 518(7540): 529–533.

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 Preprint. arXiv 1611.00625[cs.LG]. Ithaca, NY: Cornell University Library.

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 *AGI 2016: Artificial General Intelligence,* ed. B. Steunebrink, P. Wang, and B. Boertzel. Lecture Notes in Artificial Intellligence volume 9782, 118–128. Berlin: Springer.

2018 AAAI Special Award Nominations

AAAI is pleased to announce the continuation of several special awards in 2018, and is currently seeking nominations for the 2018 AAAI Classic Paper Award, the AAAI Distinguished Service Award, and the AAAI/EAAI Outstanding Educator Award. The 2018 AAAI Classic Paper Award will be given to the author of the most influential paper(s) from the Seventeenth National Conference on Artificial Intelligence, held in 2000 in Austin, Texas. The 2018 AAAI Distinguished Service Award will recognize one individual for extraordinary service to the AI community. The AAAI/EAAI Outstanding Educator Award honors a person (or group of people) who has made major contributions to AI education that provide long-lasting benefits to the AI community. Awards will be presented at AAAI-18 in New Orleans, Louisiana, USA.

Complete nomination information, including nomination forms, is available on the AAAI website (www.aaai.org/Awards/awards.php). The deadline for nomination is September 29, 2017. For additional inquiries, please contact Carol Hamilton at hamilton@aaai.org.

José Hernández-Orallo (PhD Universitat de València) is professor at Universitat Politècnica de València, Spain.

Marco Baroni (PhD, University of California, Los Angeles) is a research scientist at Facebook Artificial Intelligence Research in Paris, France, and an associate professor at the Center for Mind/Brain Sciences of the University of Trento, Italy.

Jordi Bieger is a PhD student in the Center for Analysis and Design of Intelligent Agents at Reykjavik University, Iceland.

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

David L. Dowe (PhD, Monash University) is a professor at Monash University, Australia.

Katja Hofmann (PhD, University of Amsterdam) is a researcher in the Machine Intelligence group at Microsoft Research in Cambridge, UK.

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

Claes Strannegård (PhD, University of Gothenburg) is a professor at Chalmers University of Technology in Gothenburg, Sweden.

Kristinn R. Thórisson (PhD, Massachusetts Institute of Technology) is a professor of computer science at Reykjavik University, Iceland, and Director of the Icelandic Institute of Intelligent Machines.