Proceedings of Machine Learning Research 123:164–176, 2020 NeurIPS 2019 Competition and Demonstration Track

The Animal-AI Testbed and Competition

Matthew Crosby M.CROSBY@IMPERIAL.AC.UK
Benjamin Beyret BB1010@IC.AC.UK
Murray Shanahan M.SHANAHAN@IMPERIAL.AC.UK
Leverhulme Centre for the Future of Intelligence, Imperial College London, UK

José Hernández-Orallo JORALLO@UPV.ES
Leverhulme Centre for the Future of Intelligence, Universitat Politècnica de València, Spain

Lucy Cheke LGC23@CAM.AC.UK

Marta Halina MH801@CAM.AC.UK
Leverhulme Centre for the Future of Intelligence, University of Cambridge, UK

Editors: Hugo Jair Escalante and Raia Hadsell

Abstract

Modern machine learning systems are still lacking in the kind of general intelligence and common sense reasoning found, not only in humans, but across the animal kingdom. Many animals are capable of solving seemingly simple tasks such as inferring object location through object persistence and spatial elimination, and navigating efficiently in out-of-distribution novel environments. Such tasks are difficult for AI, but provide a natural stepping stone towards the goal of more complex human-like general intelligence. The extensive literature on animal cognition provides methodology and experimental paradigms for testing such abilities but, so far, these experiments have not been translated en masse into an AI-friendly setting. We present a new testbed, *Animal-AI*, first released as part of the *Animal-AI Olympics* competition at NeurIPS 2019, which is a comprehensive environment and testing paradigm for tasks inspired by animal cognition. In this paper we outline the environment, the testbed, the results of the competition, and discuss the open challenges for building and testing artificial agents capable of the kind of nonverbal common sense reasoning found in many non-human animals.

1. Introduction

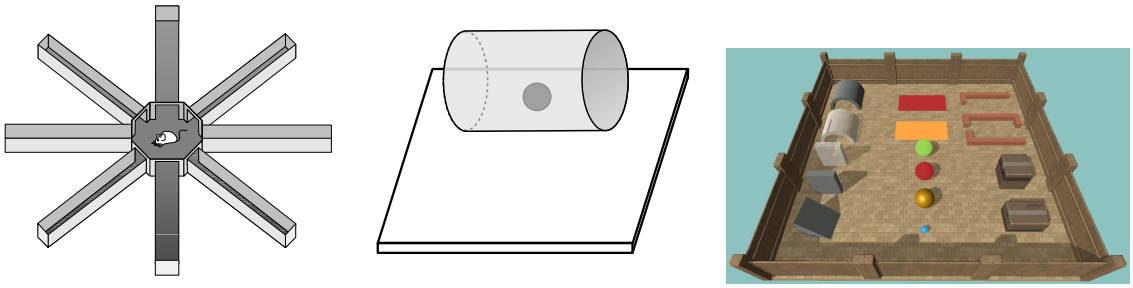
We have recently seen a wide variety of challenging environments where AI now outperforms humans such as Atari games (Bellemare et al., 2012; Mnih et al., 2013), Go (Silver et al., 2016) and Starcraft 2 (Vinyals et al., 2017). Successes such as these have been driven by the introduction of game environments and physics simulators as testing arenas (Todorov et al., 2012), and have even resulted in the transfer of trained agents to the 'real' world for robotic manipulations (OpenAI et al., 2018). While these results are impressive, they are still limited in many ways (Geirhos et al., 2020) and are only a first step towards agents that can robustly interact with their environments, apply common sense reasoning, and adapt to truly novel situations.

Meanwhile, over the last century, comparative psychologists have refined a multitude of experimental paradigms through which to probe animals' cognitive and behavioural capacities (Thorndike, 1911; Shettleworth, 2009). As a result, a wide range of standardised tests now exist, designed to minimise confounding factors, noise, and other non-cognitive elements that may interfere with identifying the targeted skill (Shaw and Schmelz, 2017). Animals have been tested for a wide variety of abilities by exploiting their intrinsic motivation to retrieve food. The food is placed in cleverly designed apparatus (see Figure 1) such that, as much as is possible, retrieval demonstrates the capability in question. Such abilities include object permanence (tracking objects that go briefly out of sight) (Chiandetti and Vallortigara, 2011), spatial memory (remembering previously taken paths) (Hughes and Blight, 1999), and using simple objects as tools (Bluff et al., 2010). These tasks are made especially hard because the inputs are low-level noisy sensory information and, whilst it has been a dream of AI to recreate biological intelligence for decades, it is only with the recent advances mentioned above that it is feasible to even consider testing agents on similar tasks with pixel-based visual inputs.

In this paper we describe *Animal-AI*, a testbed inspired by the comparative cognition paradigm. This includes a novel environment, a 3D-simulated arena using the Unity ML-Agents framework (Juliani et al., 2018), with a simple simulated physics and a set of objects that can be combined to build the kinds of environments and apparatus found in animal experiments. The environment and objects are designed to be as simple as possible whilst still maintaining the possibility to build a wide range of tasks. The testbed contains 12 different categories of tasks over a range of difficulties so that it can be used as both a research path and a measure of AI progress. Easier tasks involve navigation towards food in otherwise empty arenas and choosing between positive and negative rewards. Harder tasks involve working out which of multiple objects can be used to retrieve the food from an inaccessible area and then correctly manipulating them to do so. The benchmark was used for the first Animal-AI Olympics competition, held in 2019, whose results are analysed here, and is intended to be maintained and updated to continue to provide the next challenge on the step towards robust agents with animal-like general intelligence.

We believe this testbed is an important stepping stone towards building agents that can robustly interact with, not only predefined intellectual problems, but also the messy, multifaceted challenges of the 'real-world'. Animal cognition tasks are simplified, abstracted variants of the natural challenges that biological entities evolved to overcome, and often require the use of intuitive physics, accurate representation of the properties of food and other objects, and predictions of the effects of causal interactions. In using these tasks, Animal-AI tests for the basic cognitive abilities that form the foundation of our common sense understanding of the everyday world, with the idea to first focus on the 'simple', currently overlooked, problems and build up from there. To solve the tasks, we must develop agents that can model noisy sensory data at an object level, predict the consequences of their actions, and react appropriately in a wide range of novel situations. It turns out that these 'simple' comprise a considerable challenge. There are many who have argued for a similar shift in focus for AI (Lake et al., 2017; Pearl and Mackenzie, 2018; Chollet, 2019), and we hope to have provided a useful resource for work in this area.

2. Animal and AI testing paradigms



(*a*) 8-Arm Maze (*b*) Cylinder Test (*c*) Animal-AI

Figure 1: Two different apparatus commonly used in comparative cognition alongside the Animal-AI arena and its object building blocks. The objects in Animal-AI can be combined and resized to easily build many different experiments (including the two in the figure).

The study of animal behaviour includes areas such as ethology, behavioural ecology, evolutionary psychology, comparative psychology and more recently, comparative cognition (Wasserman and Zentall, 2006; Shettleworth, 2009). In many of these fields, particularly the latter two, animals are evaluated in carefully designed and controlled conditions, using standardised procedures. There are many issues caused by working with animals that make it difficult to draw any concrete conclusions from experiments (Farrar and Ostojič, 2019). Fortunately, the translation of the tasks to an AI setting mitigates many of these problems.

Experimental designs for animal cognition take a number of considerations into account. All tasks involve some basic functions that are taken for granted; the ability to move, recognise relevant environmental features (e.g., objects, substrates), the motivation to achieve the designated reward (almost always food), and so on. Note that none of these A crucial stipulation is that completing the task must demonstrate application of the cognitive skill or capacity under investigation. First, it is important that the other skills brought to the table cannot be used to solve the tasks by themselves. To facilitate this, cues that would allow subjects to solve the task (e.g., odour-trails in a maze) without relying on the target skill must be controlled for or eliminated. Second, the animal should not have learned or been trained to follow a sequence of actions that happens to solve the task without the target ability. This issue can be controlled for by testing the animal on multiple experimental setups that test the same skill, but involve different stimuli or action patterns (Herrmann et al., 2007; Shaw and Schmelz, 2017). Another standard practice is to record results only for an animal's *first* experience of a new environment. Finally, other environmental or personality factors that may affect behaviour (such as the behaviour of other individuals, individual differences in reward motivation, fear of the testing environment or distraction from irrelevant stimuli) must be minimised or taken into account in the statistical analysis of the results (Shaw and Schmelz, 2017). Note that these principles (especially the prereq-

uisites that can be assumed in animals but not artificial agents), make experimental tasks used in comparative cognition very different from standard AI testbeds.

On the other side, the AI-evaluation landscape has many benefits over testing in animals. It is much easier to test agents, reset and rerun experiments, and generate and store large amounts of data. At the same time, AI has developed common paradigms that hold back progress towards common sense intelligence. **(1)** Many challenges are based on existing games or environments, not deliberately designed for testing specific abilities (e.g. chess, Go, Atari Games, Dota, Starcraft) (Hernández-Orallo, 2017). **(2)** To be considered a real challenge, oversimplified tasks that look like toy problems, such as the cylinder task in Figure 1(*b*), are generally avoided (even if AI cannot currently solve them). **(3)** When many tasks are integrated into a benchmark, it is not always the same trained agent (but the same algorithm) that is evaluated on them (fortunately exceptions to this are becoming more common – see Section 6). **(4)** Many tests, especially those in a supervised or RL setting, disclose lots of information about each task by giving the possibility of training on several episodes of the same task (or slight variations of the task). This facilitates 'shortcut' solutions Geirhos et al. (2020). **(5)** Finally, the overriding metric for achievement is usually a continuous performance score, with other measures such as training time, procedure, and behavioural analysis often overlooked. Animal cognition commonly relies on other measures which can be informative not just on what was done, but how, and with what proficiency, it was achieved. The Animal-AI Testbed presents a new paradigm for AI that combines the positive components of both animal and AI paradigms and runs contrary to (1-5).

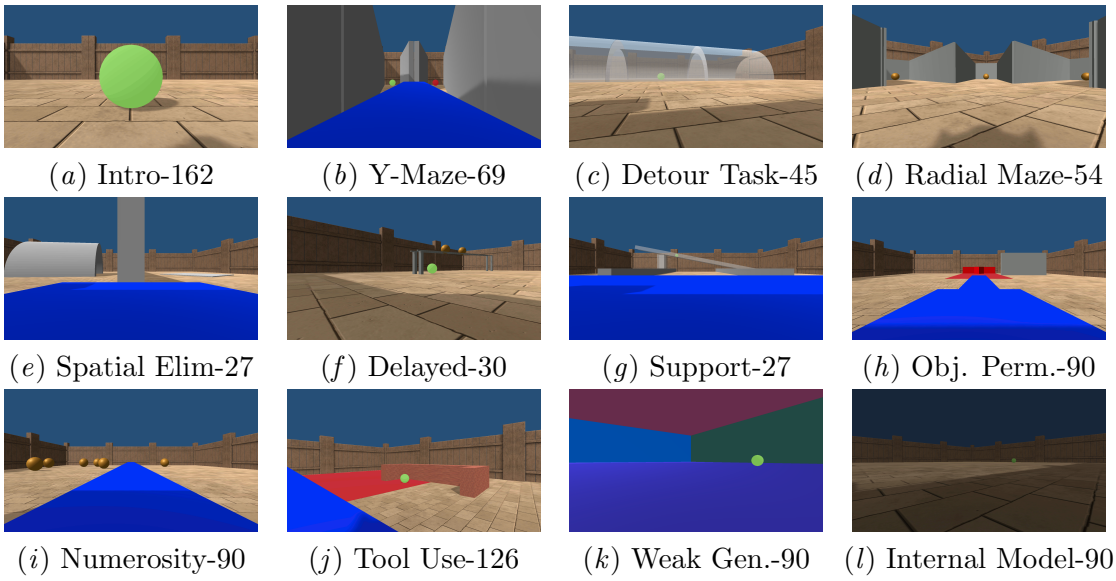
3. The Animal-AI Environment

The Animal-AI environment contains two components: (*i*) a simulator built using the Unity game engine and (*ii*) a training API written in Python. The simulator comprises an arena which is kept deliberately small and can contain a set of relatively simple objects with basic textures so that we exclude as many confounding factors as possible (see Figure 1(*c*)). The tiled floor and wooden walls are included to still give some (task independent) visual cues to the agent. The simulated physics reproduces how objects behave in the real world (e.g., gravity, collisions, friction, etc.). The various experiments presented throughout this paper and on our website are all defined using easy to write configuration files (`YAML` format). Using these, the different object types can be placed in the arena and easily resized, rotated and combined to build complex structures.

In order to train the agent in this arena, the Python API interfaces with the simulator. This is done via a classic reinforcement learning loop where the agent receives pixel inputs, is capable of taking simple actions (move forward/backward and turn left/right), and is rewarded only for retrieving designated reward (food) objects. The framework allows for large parallel training with inbuilt RL algorithms. The environment, along with the testbed, documentation, and tutorials for getting started are available at `https://github.com/beyretb/AnimalAI-Olympics`. The environment has both Gym (Brockman et al., 2016) and ML-Agents APIs (Juliani et al., 2018). The Unity simulator is also available open source `https://github.com/beyretb/ml-agents`. This allows researchers to modify the environment as they wish, for example, to add commonly requested features such as extra

cameras, raycasting, multiple agents, or new object types. For more information about the environment see Beyret et al. (2019).

4. The Testbed



(*a*) Intro-162 (*b*) Y-Maze-69 (*c*) Detour Task-45 (*d*) Radial Maze-54

(*e*) Spatial Elim-27 (*f*) Delayed-30 (*g*) Support-27 (*h*) Obj. Perm.-90

(*i*) Numerosity-90 (*j*) Tool Use-126 (*k*) Weak Gen.-90 (*l*) Internal Model-90

Figure 2: Example problems from each of the 12 task types. (a) contains familiarisation tasks. (b-j) have direct links to animal tests. (k,l) are AI-specific. Number of problems per category shown in captions (900 total).

The testbed is a set of 300 tasks, each with 3 minor variations (900 configurations in total) split into 12 categories. Each configuration has a set of objects including at least one positive reward item (food), a starting location for the agent, a time limit, and a reward threshold. The agent passes the test if its cumulative reward at the end of the episode is above the threshold. Thresholds are generally set so that if the agent retrieves the most possible food items within the time limit it passes the task. More details can be found at animalaiolympics.com where all tasks can be played online and it is also possible to view all the agents' solutions from the competition.

The tests follow certain conventions to make solving them easier. Objects of a particular variety always appear in a particular colour. Immovable walls are grey, ramps are pink, and all other objects use the same skins in Figure 1(c). There are no interventions that change the layout or properties of the environment during a test, everything is set up in the initial configuration file and plays out from there based on the physics and agent actions. This is potentially limiting compared to animal cognition, where humans often intervene to change elements of the environment mid-test or suspend tests early. One example intervention is to set up a forced choice, whereby the experiment is stopped after an animal picks one of multiple options. To replicate this in the environment we use walls as platforms. Blue

platforms denote areas that can not be climbed - the agent can move off them, but not get back up. Using such methods we can implement many standard animal experiments.

a: **Introductory** There are a large number of tasks in the testbed that are used to assess interaction with each of the elements in the environment without interference from other elements. Crucially, these tasks allow assessment of baseline behaviour for elements required to interpret performance on more complicated tests.

b: **Y-Mazes** These are very simple mazes in the shape of the letter 'Y' which present an animal with two simultaneously visible choices and are commonly used in animal studies to assess preference (Pajor et al., 2003; Pollard et al., 1994).

c: **Detour Tasks** These include detour and cylinder tasks such as have been performed in (Smith and Litchfield, 2010) and discussed in (MacLean et al., 2014). In detour tasks, food is placed behind a barrier that the animal must detour around. In the cylinder task, food is placed in a transparent cylinder orientated such that the entrances are perpendicular to the animal. In both cases the animal must suppress the urge to move directly towards the food, and instead take a longer route which involves moving away from the food in order to eventually retrieve it.

d: **Radial Mazes** Radial mazes are commonly used in animal cognition as they can easily be varied in dimensions such as number of arms and landmarks which can be used to help guide navigation. In common setups the food at the end of each arm is hidden (and masked from producing odors) so that memory is required to avoid revisiting previously visited arms (Hughes and Blight, 1999).

e: **Spatial Elimination** In these tasks success requires the ability to reason about the location of a reward based on eliminating possibilities where it cannot possibly be. For example, if the reward is not visible, it cannot be in a location within the visual field. Similar tasks were performed in the Primate Cognition Test Battery (PCTB) with food items hidden (whilst out of sight) underneath cloth or boards such that the location is visually apparent due to bumps or inclines caused by the food (Herrmann et al., 2007).

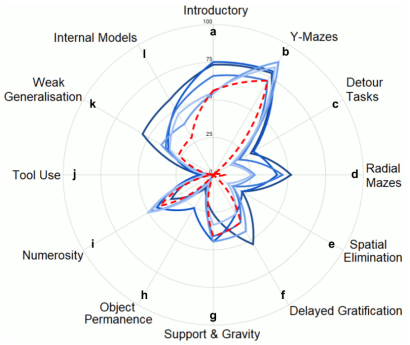
f: **Delayed Gratification** These tasks require the ability to forgo an immediate, less preferred reward for a future, more preferred reward. This has been tested in a number of species, including great apes (Beran, 2002), who have been shown to delay taking a bowl of food whilst multiple items are slowly added. In the Animal-AI testbed we recreate these experiments by combining green and yellow food. An easily accessible green food will terminate the episode, but waiting for the yellow food, which is initially unaccessible but rolling towards the agent down a rail setup (shown in Figure 2(f)) or ramp will increase the overall reward achieved during the episode.

g: **Support and Gravity** These tasks involve objects affected by gravity. In a standard paradigm used in animal cognition, subjects must predict the final location of food that is seen falling. A common result is that animals will often attempt to retrieve food directly below a location it is dropped from, ignoring intervening objects (such as an inclined pipe) which lend the object horizontal velocity (Hauser et al., 2001).

h: **Object Permanence** This involves maintaining knowledge of the existence of objects when they go out of sight. Object permanence has been observed in many animals, including few-days old chicks (Chiandetti and Vallortigara, 2011). Here, the reward object (for chicks this is an object on which they are imprinted) is briefly moved out of sight, and the behaviour of the animal is analysed under different settings to see if the they explore

	a	b	c	d	e	f
Baseline (%)	56.7	72.5	0.0	7.4	0.0	32.1
Max Score (%)	75.6	88.4	31.1	51.9	25.9	50.0
Mean Score (%)	61.0	76.3	17.5	30.5	14.3	29.0
	g	**h**	**i**	**j**	**k**	**l**
Baseline (%)	40.7	0.0	41.1	0.8	26.7	28.9
Max Score (%)	44.4	25.6	50.0	9.5	54.4	57.8
Mean Score (%)	30.6	8.0	40.7	3.4	35.0	40.5

(*a*) Table of results (tests passed by category)



(*b*) Radar plot top 4

Figure 4: Left: Baseline compared to the maximum and mean scores for agents that surpassed it (total score) by the 12 categories introduced in section 4. Right: Radar plot showing performance profiles of the top 4 entries compared to red baseline.

the environment to "search" for the object, and if so, whether they do so in a manner that implies they understand where it must be. For example, their arena can include two occluding objects, one of which is too small for the reward to fit behind.

i: **Numerosity** Many animals have been shown to differentiate between quantities. For example, the PCTB contains tests in which a primate is offered multiple choices of plates of food and is counted as successful if it chooses the one with more (Herrmann et al., 2007). In our environment yellow food adds reward but does not terminate the episode so we can set up forced choice platforms with different amounts of food on each side.

j: **Tool Use** Tool use has been found across multiple species. We focus on that which is easily reproducible within our environment and by a simple non-embodied agent, and adapt experiments from the string pulling paradigm (Jacobs and Osvath, 2015), versions of the 'box and banana' test that utilise ramps and boxes (Kohler, 2018), the swing door task from the PCTB (Herrmann et al., 2007), and trap tube tasks (Mulcahy and Call, 2006).

k: **Weak Generalisation** The final two categorise are more AI-inspired than animal inspired. In weak generalisation we take versions of the previous experiments and change the colours of the objects or present unexpected situations, such as a wall object that acts as a roof over the arena.

l: **Internal Models** Internal Models tasks are also mainly variations of earlier experiments, but here the 'lights' in the environment (the visual input) are set to go on and off at regular intervals, or to go off after a certain period of time. These are designed to test the capacity of agents to build accurate predictive models.

5. Results

The testbed was presented as a competition, *The Animal-AI Olympics,* where 60 teams worked on training an agent over 4 months. The competition format ensured it was possible to assess performance on completely hidden tasks (see Section 2 for why this is important).

Participants were given full access to the environment, the ability to generate any configurations they liked for training, and feedback in terms of number of tests passed only when they submitted an agent to a private testing server. However, they were only given limited information about the contents of the tasks. This was, of course, a very hard challenge, with many of the tests designed to require long-term research. Nevertheless, we can draw some important conclusions from the results.

The majority of the entries used Deep Reinforcement Learning in some form. The winning entry (by Denys Makoviichuk) used an iterated process of building training environments, training an agent, and a validation step involving behavioural analysis. The DRL algorithm used was PPO, with a CNN architecture feeding into an LSTM layer. The training environments were hand-designed configurations that made use of the possibility to specify certain values to be randomised. In the behavioural analysis step, the agent's performance on a custom-built test set was analysed and then the algorithm, training set, and reward shaping were used to encourage more robust behaviour. For example, a small positive reward was given for achieving vertical velocity which made the agent seek out ramps - a useful skill in some of the problems.

Figure 4 (top table) shows the results of the baseline, a simple hand-coded agent that moves towards positive rewards (food) and away from negative rewards, based on summing pixel colours on the left and right sides of the visual input. As expected, its performance is 0 on the detour task, which involves navigating around an object and not just moving towards it. The table also shows the maximum and mean scores of the competitors that are better than the baseline (in terms of tests passed for all categories). The competition entries do significantly better here, but do not solve any of the more complex tasks. Figure 4 (left) shows a radar plot of the top four agents. The best agents did not show robust solutions (solving all variations of a subtask) except in the Introductory category and Maze variations. On tool use, where even the easiest tasks required some manipulation of objects, even the best agents scored close to 0. Also of note is category (g) - support and gravity - where mean score was below the baseline. Behavioural analysis showed that many agents failed to recognise food that was in the top half of their field of vision, presumably because primarily trained on configurations with food on the ground and had not learned to associate food (no matter its location) with reward.

We provide further analysis of the results in the Appendix and more detailed competition results (including the ability to watch the agents) can be found at animalaiolympics.com. As the majority of tasks were intended as long-term challenges and unsolved in the competition, a thorough analysis of agent performance will only be fruitful once further progress has been made. In the meantime, we are looking at developing the testbed to allow for better discrimination between existing algorithms and are also performing experiments to compare directly between animal, agent, and human performance.

6. Related Benchmarks in AI

Progress in DRL in recent years has been fuelled by the use of games and game-inspired simulated environments (Castelvecchi, 2016; Hernández-Orallo et al., 2017). Some important benchmarks are simply collections of existing games, such as the very popular *Arcade Learning Environment* (ALE) (Bellemare et al., 2015; Machado et al., 2018), with dozens

of (2D) Atari 2600 games. In a similar vein, OpenAI Gym (Brockman et al., 2016) provides a common interface to a collection of RL tasks including both 2D and 3D games. There has been a recent trend towards generalisation challenges in AI testing environments, where some skill transfer between training and testing is necessary. *CoinRun* (Cobbe et al., 2018) is a 2D arcade-style game with procedural generation for testing on unseen levels to quantify overfitting. Another example, *Obstacle Tower* (Juliani et al., 2019) is a 3D game based on Montezuma's revenge, one of the harder (for AI) Atari games, whose stages are generated in a procedural way to ensure that the agent is tested on unseen room and puzzle layouts.

Other platforms are designed, like ours, to be customisable. For instance, the video game definition language (VGDL) has led to several *General Video Game AI* (GVGAI) competitions, with new games for each edition (Pérez-Liébana et al., 2016). *ViZDoom* (Kempka et al., 2016) is a research platform with customisable scenarios based on the 1993 first-person shooting video game Doom that has been used to make advancements in model-based DRL (Ha and Schmidhuber, 2018). Microsoft's *Malmo* (Johnson et al., 2016), which is based on the block-based world of Minecraft, also makes it possible to create new tasks, ranging from navigation and survival to collaboration and problem solving. Finally, *DeepMind Lab* (Beattie et al., 2016) is an extensible 3D platform with simulated real-world physics built upon id Software's Quake III Arena. Each of these is useful for a different type of tests, but none has everything we needed.

Notable ability-oriented approaches include *bsuite*, which presents a series of reinforcement learning tasks designed to be easily scalable and to provide a measure for a number of core capabilities (Osband et al., 2019). These tests are deliberately simple to allow for a more accurate measure of the ability being tested. A key benchmark is 'Abstraction and Reasoning' (Chollet, 2019), which, like us, aims at common sense reasoning with tasks easy for humans to solve. This provides a concise and important version of the kinds of tasks we ultimately want to be solvable in the Animal-AI environment. It differs from this work by not using an agent situated and acting within an environment.

7. Conclusions

The Animal-AI testbed is a new AI experimentation and evaluation platform that implements ideas from animal cognition. It is designed to allow for cognitive testing built up from perception and navigation. We start with simple, yet crucial, tasks that many animals are able to solve, and also include more complex reasoning tasks. The testbed has highlighted many open challenges for AI which will take new ideas in order to solve. The 900 tests are now publicly available which means they lose the 'hidden' factor. We are therefore adding a new set of hidden tasks that contain unseen variations and can be used to measure progress and avoid some of the inevitable overfitting. We have seen incredible results in AI in recent years. We hope this momentum can translate to begin to solve the kind of problems that animals solve on a daily basis when navigating their environment or foraging for food and that this will be an important milestone on the path towards general intelligence.

Acknowledgments

This work was supported by the Leverhulme Centre for the Future of Intelligence, LeverhulmeTrust, under Grant RC-2015-067.

References

Charles Beattie, Joel Z Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, et al. Deepmind lab. *arXiv preprint arXiv:1612.03801*, 2016.

Marc Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. In *24th Int. Joint Conf. on Artificial Intelligence*, 2015.

Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *CoRR*, abs/1207.4708, 2012.

Michael J Beran. Maintenance of self-imposed delay of gratification by four chimpanzees (pan troglodytes) and an orangutan (pongo pygmaeus). *The Journal of General Psychology*, 129(1):49–66, 2002.

Benjamin Beyret, José Hernández-Orallo, Lucy Cheke, Marta Halina, Murray Shanahan, and Matthew Crosby. The animal-ai environment: Training and testing animal-like artificial cognition, 2019.

Lucas A Bluff, Jolyon Troscianko, Alex AS Weir, Alex Kacelnik, and Christian Rutz. Tool use by wild new caledonian crows corvus moneduloides at natural foraging sites. *Proceedings of the Royal Society B: Biological Sciences*, 277(1686):1377–1385, 2010.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Davide Castelvecchi. Tech giants open virtual worlds to bevy of AI programs. *Nature News*, 540(7633):323, 2016.

Cinzia Chiandetti and Giorgio Vallortigara. Intuitive physical reasoning about occluded objects by inexperienced chicks. *Proceedings of the Royal Society B: Biological Sciences*, 278(1718):2621–2627, 2011.

François Chollet. On the measure of intelligence, 2019.

Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. *arXiv preprint arXiv:1812.02341*, 2018.

Benjamin G Farrar and Ljerka Ostojič. The illusion of science in comparative cognition. *PsyArXiv. October*, 2, 2019.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Marc D Hauser, Travis Williams, Jerald D Kralik, and Damian Moskovitz. What guides a search for food that has disappeared? experiments on cotton-top tamarins (saguinus oedipus). *Journal of Comparative Psychology*, 115(2):140, 2001.

J. Hernández-Orallo. Evaluation in artificial intelligence: From task-oriented to ability-oriented measurement. *Artificial Intelligence Review*, 48(3):397–447, 2017.

José Hernández-Orallo, Marco Baroni, Jordi Bieger, Nader Chmait, David L Dowe, Katja Hofmann, Martínez-Plumed, et al. A new AI evaluation cosmos: Ready to play the game? *AI Magazine*, 38(3):66–69, 2017.

Esther Herrmann, Josep Call, María Victoria Hernández-Lloreda, Brian Hare, and Michael Tomasello. Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *science*, 317(5843):1360–1366, 2007.

Roger N Hughes and Christine M Blight. Algorithmic behaviour and spatial memory are used by two intertidal fish species to solve the radial maze. *Animal Behaviour*, 58(3): 601–613, 1999.

Ivo F Jacobs and Mathias Osvath. The string-pulling paradigm in comparative psychology. *Journal of Comparative Psychology*, 129(2):89, 2015.

Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. The malmo platform for artificial intelligence experimentation. In *IJCAI*, pages 4246–4247, 2016.

Arthur Juliani, Vincent-Pierre Berges, Esh Vckay, Yuan Gao, Hunter Henry, Marwan Mattar, and Danny Lange. Unity: A general platform for intelligent agents. *CoRR*, abs/1809.02627, 2018.

Arthur Juliani, Ahmed Khalifa, Vincent-Pierre Berges, et al. Obstacle tower: A generalization challenge in vision, control, and planning. *arXiv preprint arXiv:1902.01378*, 2019.

Michał Kempka, Marek Wydmuch, Grzegorz Runc, Jakub Toczek, and Wojciech Jaśkowski. Vizdoom: A doom-based ai research platform for visual reinforcement learning. In *Conference on Computational Intelligence and Games (CIG)*, pages 1–8. IEEE, 2016.

Wolfgang Kohler. *The mentality of apes*. Routledge, 2018.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

Marlos C Machado, Marc G Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *J. of Artificial Intelligence Research*, 61:523–562, 2018.

Evan L MacLean, Brian Hare, Charles L Nunn, et al. The evolution of self-control. *Proceedings of the National Academy of Sciences*, 111(20):E2140–E2148, 2014.

Fernando Martinez-Plumed and Jose Hernandez-Orallo. Dual indicators to analyse AI benchmarks: Difficulty, discrimination, ability and generality. *IEEE Transactions on Games*, 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Nicholas J Mulcahy and Josep Call. How great apes perform on a modified trap-tube task. *Animal cognition*, 9(3):193–199, 2006.

OpenAI, Marcin Andrychowicz, Bowen Baker, et al. Learning dexterous in-hand manipulation. *CoRR*, abs/1808.00177, 2018.

Ian Osband, Yotam Doron, Matteo Hessel, John Aslanides, Eren Sezener, Andre Saraiva, Katrina McKinney, Tor Lattimore, Csaba Szepezvari, and otherst. Behaviour suite for reinforcement learning. *arXiv preprint arXiv:1908.03568*, 2019.

EA Pajor, J Rushen, and AMB De Passillé. Dairy cattle's choice of handling treatments in a y-maze. *Applied Animal Behaviour Science*, 80(2):93–107, 2003.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

Diego Pérez-Liébana, Spyridon Samothrakis, Julian Togelius, et al. The 2014 general video game playing competition. *IEEE Trans. on Computational Intelligence and AI in Games*, 8(3):229–243, 2016.

JC Pollard, RP Littlejohn, and JM Suttie. Responses of red deer to restraint in a y-maze preference test. *Applied Animal Behaviour Science*, 39(1):63–71, 1994.

Rachael C Shaw and Martin Schmelz. Cognitive test batteries in animal cognition research: evaluating the past, present and future of comparative psychometrics. *Animal cognition*, 20(6):1003–1018, 2017.

Sara J Shettleworth. *Cognition, evolution, and behavior*. Oxford university press, 2009.

David Silver, Aja Huang, Christopher J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–503, 2016.

Bradley Philip Smith and Carla Anita Litchfield. How well do dingoes, canis dingo, perform on the detour task? *Animal Behaviour*, 80(1):155–162, 2010.

Edward Thorndike. *Animal Intelligence: Experimental Studies*. The Macmillan Company, 1911.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IROS*, pages 5026–5033. IEEE, 2012. ISBN 978-1-4673-1737-5.

Oriol Vinyals, Timo Ewalds, Sergey Bartunov, et al. Starcraft II: A new challenge for reinforcement learning. *CoRR*, abs/1708.04782, 2017.

Edward A Wasserman and Thomas R Zentall. *Comparative cognition: Experimental explorations of animal intelligence*. Oxford University Press, USA, 2006.

Appendix (Further Results)

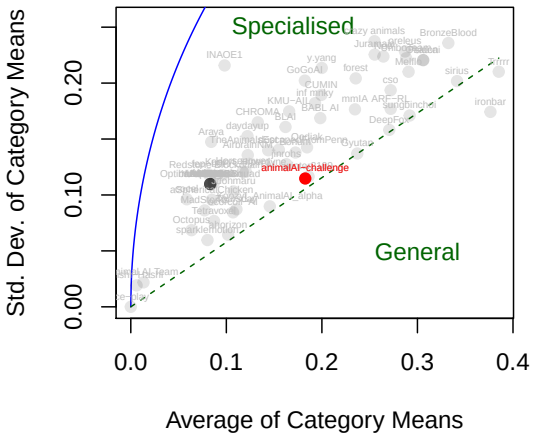


Figure 5: Comparison of average results (except the category "Introductory") and standard deviation. Blue is the theoretical maximum given by a Bernoulli distribution. More general agents appear below the diagonal.

We also performed a straightforward analysis of generality following (Martinez-Plumed and Hernandez-Orallo, 2018), where a *general* agent is understood as one that gets similar (good) scores for a wide range of categories, in front of more *specialised* agents that may do very well on a few categories but very poorly on others. In particular, for the same overall performance, a more general agent should have a lower variance over the category scores. Accordingly, in figure 5 each grey circle represents a participant, with overall performance shown on the x-axis and the standard deviation of the category means on the y-axis. A maximally general agent would show at the bottom, with 0 standard deviation. The blue solid curve is the standard deviation of a maximally specialised agent (being perfect on some categories but totally failing on others, which corresponds to the standard deviation of a Bernoulli distribution) and the green dashed diagonal shows the standard deviation of a relatively general agent (with the standard deviation of a uniform distribution). Note that the baseline agent ('animalAI-challenge', in red) falls very near to this diagonal. We also observe that only two agents are on the general side of the diagonal. These are the best two participants, 'Trrrrr' and 'ironbar'.