Science Birds Novelty: an Open-world Learning Test-bed for Physics Domains

Cheng Xue, Vimukthini Pinto, Peng Zhang, Chathura Gamage, Ekaterina Nikonova, Jochen Renz

School of Computing, The Australian National University, Canberra, Australia
cheng.xue, vimukthini.inguruwattage, p.zhang, chathura.gamage, ekaterina.nikonova, jochen.renz@anu.edu.au

Abstract

Successfully operating in open worlds is a hallmark of human intelligence but still remains a major challenge to modern Artificial Intelligence (AI) systems. With the increasing reliance on autonomous systems (e.g. self-driving vehicles, vacuuming robots), being able to handle unforeseen situations has become a crucial ability for any AI agent that can safely and effectively operate alongside humans. As to facilitate the research in developing agents that are capable of reacting to unexpected events, we propose a test-bed named *Science Birds Novelty* based on the Angry Birds domain. We also demonstrate a use-case of our test-bed, the AIBIRDS Competition Novelty Track and present the results of the competition.

1 Introduction

One of the ultimate goals in the AI field is to have systems that can safely work alongside humans in real-world environments. With the ongoing applications and transitions of AI systems from constrained lab environments to much messier real-world environments, the ability to handle unexpected events (novelties) has taken on new importance in recent years. As a field focusing on developing systems that can operate in such open worlds, open-world learning (OWL) has been proposed (Langley 2020) recently; A successful open-world system has been defined as one that not only deals with in-distribution inputs, but also rapidly adapts to out-of-distribution inputs.

It is usually not feasible to develop and test OWL systems directly in real-world environments due to the limited opportunities to inject novelties (Langley et al. 1981; Choi et al. 2007); therefore, simulated test-beds that allows AI agents to learn and enables to evaluate agents are essential to advance the research in OWL (Langley 2020). We believe an ideal test-bed for OWL should have 5 characteristics. 1) it should be **simple** enough to allow agents to specifically focus on dealing with novelties in the domain. This ensures agents are not affected by problems that involve other bodies of AI research (e.g. object tracking in crowed scenes). 2) The test-bed should be **versatile** enough. This means the test-bed should enable to inject different kinds of novelty ((Langley 2020; Boult et al. 2021)). 3) The test-bed should be **well-controlled**. That is, in order to systematically evaluate the performance of OWL systems, users of the test-bed should be able to decide precisely what, when, where, and how novelties will appear in the environment. 4) It should **allow agents to report novelty detection and characterization, and record them together with the performance of the agents**. Although the novelty adaption performance is what we care about, it is still useful to understand if the agent adapts to the novelty because it successfully detects and characterize the novelty or just because of luck. 5) It is desirable for an OWL test-bed to come with **a set of baseline agents that have the expertise or a large enough normal training data-set to support acceptable agent performance in the normal environment**. These baseline agents would help OWL agent developers to get started by studying the baseline agents and they help to evaluate how good OWL agents can adapt to novelties by comparing the performance of OWL agents with the baseline agents.

2 Angry Birds and AIBIRDS Competition

Angry Birds, a simple and intuitive game with realistic physical simulation, has been one of the most popular testing domains for physical reasoning among the AI community (Renz et al. 2019). The goal of Angry Birds is to destroy all pigs in a game level by shooting birds from a slingshot. Pigs are normally protected by physical structures made of blocks with varied sizes, shapes, and materials. Some birds have special powers that can be activated after being released from the slingshot. The only actions available to the players are to select a bird trajectory by pulling the bird back in the slingshot to the release coordinates (x, y) and then tapping the screen at time t after release to activate the special power.

The Angry Birds AI Competition (https://aibirds.org) encourages the AI community to develop agents that can deal with a large action space, that don't have complete knowledge about the physical parameters of objects and where, therefore, the consequences of possible actions can only be estimated. No forward model is available. Humans deal with such situations all the time and quickly build up experience to estimate outcomes within the physical world. Future AI needs to have the same capabilities: robots that are not aware of the consequences of their physical actions will be unsafe in a human environment. The AIBIRDS competition has been organized annually since 2012, mostly collocated with the International Joint Conference on Artificial Intelligence. 30 minutes are allowed for each elimination round where agents play eight new levels per round, until a winner is determined. Over 60 teams have participated so far. Different approaches, such as advanced simulation, reasoning, planning, heuristic search, various machine learning approaches (including deep learning), and combinations thereof are presented during the competition.

3 Test-bed: *Science Birds Novelty*

After extensively modifying the original Science Birds (Ferreira and Toledo 2014), which is an open-source Angry Birds research clone using Unity, we present our test-bed: the *Science Birds Novelty* [1].

3.1 Simplified Inputs

Recent research (Bear et al. 2021) shows that physical reasoning performance of AI agents can be significantly affected owing to the errors coming from computer vision components. As *Science Birds Novelty* is a test-bed for physics domains, we encourage agents to develop physical reasoning abilities when interacting with novelties. Therefore, in addition to standard screenshot state representation, we also provide ground truth state representation to avoid the need for computer vision.

A screenshot state representation is a 480 x 640 coloured image and the ground truth representation is in JSON format containing all foreground objects in a screenshot. Each object in the ground truth representation is represented as a polygon of its vertices (provided in order) and its respective colour map containing a list of 8-bit quantized colours that appear in the game object with their respective percentages. An agent can request screenshots and/or a ground truth representations of the game level at any time while playing. To further save agents from the object tracking task, we provide an object ID in the ground truth for each object. Hence, an agent can request a batch of ground truth with a desired frequency after a bird has been released and calculate the trajectories of objects of interests without using any advanced computer vision techniques.

3.2 Versatile Possible Novelty Types

As an open-source project written in C# using Unity, Science Birds allows us to modify the game in almost any way we want and hence to introduce a large variate of novel situations. This can be as easy as changing physical parameters and colours of existing game objects, to more difficult modifications such as introducing a new class of objects (e.g. hostile external agents that hinders the agent) or changing the game goals (e.g. instead of killing the pigs, the goal changes to destroy all wood blocks).

3.3 Controlled Novelty Injection

Controlled novelty injection is particularly important when it comes to systematically evaluating novelty detection and reaction performance. For example, being able to decide when a novel situation is appearing in the testing environment helps researchers to measure the timeliness of OWL agents' responses: how long does it take to detect the novelty and how long does it take to adapt to the novelty. Moreover, controlling what, where, and how novelties appear in the environment can help a more detailed evaluation of OWL agents and hence provides clear insights regarding which further improvements are required. For example, an agent might adapt to tasks much faster where the novelty is a reduced mass of objects, rather than to tasks that require the agent to dealing with increased friction of objects as novelty. This may suggest that the agent's ability to reasoning with friction should be improved.

In *Science Birds Novelty*, users can specify the exact type and location of novelty to be included and decide the exact order when the novel game level will appear in a trial (discussed in section 3.7).

3.4 Baseline Agents

Together with the test-bed, we include 3 heuristic-based agents: Eagle Wings, Datalab, Naive Agent, and 1 deep-learning based agent, the DQ-Birds.

- **Eagle Wings:** This is the winner of 2017 and 2018 competitions. The agent selects action based on strategies including shoot at pigs, destroy most blocks, shoot high round objects, and destroy structures (Wang 2017).
- **Datalab:** Datalab is the winner of the 2014 and 2015 competitions. The agent uses the following strategies: destroy pigs, destroy physical structures, and shoot at round blocks. The agent selects the strategy based on the game states, possible trajectories, bird types, and the remaining birds (Borovička, Špetlík, and Rymeš 2014).
- **Naive Agent:** The strategy of the Naive Agent is to directly shoot at the pigs. The agent shoots the bird on the slingshot by randomly selecting a pig and a trajectory to shoot the pig (Stephenson et al. 2018b).
- **DQ-Birds**: A Double Dueling Deep Q-network trained on over 115,000 of Angry Birds game frames using greedy epsilon and partially random policy (Nikonova and Gemrot 2019).

3.5 Integrating Game Level Generation

To help agents to develop acceptable level of performance in normal game envrionment, we integrated the winner of the 2017 and 2018 Angry Birds Level Generation Competitions (Stephenson et al. 2018a), Iratus Aves (Stephenson and Renz 2017), into our test-bed to make it possible to create unlimited number of new game levels for training. Other procedural game level generators can be used as well.

3.6 Example Novelties

We develop a comprehensive set of 12 sample novelties covering the first four novelty levels of the Open World Novelty Hierarchy (Senator 2019).

Novelty level 0 is known as the *Instance Novelty*, which covers previously unseen instances. This corresponds to previously unseen new game levels.

[1] https://gitlab.com/aibirds/sciencebirdsframework

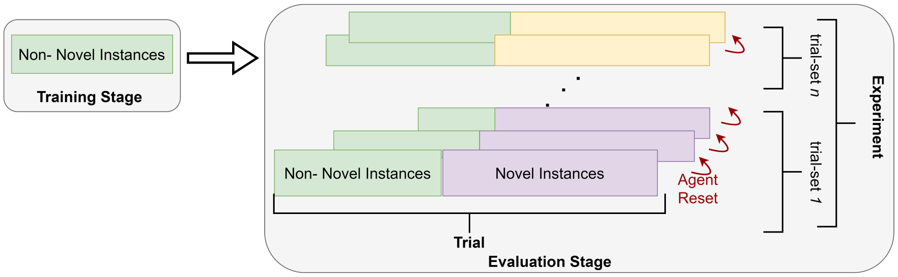

Figure 1: Training stage followed by the evaluation stage. An experiment contains *trial-sets* that are drawn from different novel distributions. A trial contains variable lengths of problem instances drawn first from a pre-novelty distribution and then from a post-novelty distribution. We assume the agent is an online learner, and we record its performance throughout a trial. The agent's model is reset at the end of each trial to eliminate model transfer. Reproduced from (Pinto et al. 2020).

The next novelty level (novelty level 1) is the *Class Novelty*, where novelties in the level are represented as previously unseen classes of objects or entities. This corresponds to new game objects with new properties, such as a new type of block that behaves differently to previous block types. These new game objects can be visually distinguished from known objects, but at first sight it is unknown how they behave. For class novelty level, we developed 5 different sample novelties: 1) new block type with increased linear drag (from 1 to 25); all the other parameters are same as wood blocks, 2) new block type with doubled score compared to the wood blocks; all the other parameters are same as wood blocks, 3) new block type with tripled health points compared to ice blocks; all the other parameters are same as ice blocks, 4) new bird type looks like dark pigs; all the other parameters are same as red birds, and 5) a pig with new appearance; all the other parameters are same as small pigs.

Novelty level 2 is the *Attribute Novelty*. This novelty level focus on changes in a feature of an object or entity, such as color, shape, or orientation not previously relevant to classification or action. Many of these novelties cannot be seen, but lead to a different game play behaviour. For this novelty level, we introduce as sample novelties 1) red bird's bounciness changed to 0.9 from 0.3, 2) red bird's linear drag changed to 0.2 from 0, 3) health points of wood blocks tripled, 4) score of wood blocks doubled, and 5) linear drag of wood blocks changed to 25 from 1.

The last novelty level, novelty level 3, is the *Representation Novelty*. The novelty level mainly include changes in how entities or features are specified, corresponding to a transformation of dimensions or coordinate systems. We develop two sample novelties for this level: 1) grayscale the colours in screenshot and ground truth representation, and 2) rotate the screenshot and ground truth by 180 degree against the centre point (420, 240) of the image.

1. Novelty Level 1: (Class)	1.1. New egg-shaped object which gives $-10,000$ points when hit. 1.2. New bird with low friction and low bounciness that can slide on the ground.
2. Novelty Level 2: (Attributes)	2.1. Pig color changed to red. 2.2. Launch force of red bird increased.
3. Novelty Level 3: (Representation)	3.1. The game is flipped upside down, agents need to shoot downwards. 3.2. Changed color map from RGB to BGR, a different color.

Table 1: Novelties in the AIBIRDS 2021 Novelty Track

3.7 Evaluation Protocol

In *Science Birds Novelty*, we follow the same evaluation protocols (Fig. 1) that is described in (Pinto et al. 2020).

1. Agents are first exposed to a sequence of normal (i.e., no novelties presented) game levels. The number of normal game levels is not known to the agent. Agents can attempt to solve each game level only once in the given order.
2. At some point, the novelty occurs, and all subsequent game levels after that point include a certain type of novelties. The number of novel game levels is unknown to the agent. Agents can attempt to solve game level only once in the given order as well.
3. For every instance (normal or novel) i an agent attempts to solve, we record its task performance (e.g., score) TSP_i and p_i, the probability that the agent believes novelty has occurred. Task performance reflects how well the agent solves a game level.

We refer to the above sequence of normal and novel game levels as a single *trial*. T_j^a is the j^{th} *trial* for a given novelty a. We refer a set of trials with the same post-novelty distribution as a *trial-set*. An *experiment* is a set of trial-sets. When an agent completes a trial, it is reset to its initial state before it begins the next trial (i.e., agents are permitted to learn throughout a trial, but learned models are not transferred between trials. The agent also reports a *detection threshold* where each p_i exceeding the threshold indicates a predicted distribution change (i.e., the agent predicts that a novelty has occurred).

4 Use-case: AIBIRDS Novelty Track

In this section we provide a demonstration of how our testbed can be used to evaluate OWL agents' performance with the AIBIRDS Competition Novelty Track.

At the AIBIRDS competition at IJCAI 2021, we introduced for the first time the **AIBIRDS Novelty Track**. We focus on the same four novelty levels described in section

3 and developed a new set of evaluation novelties for each of the level 1-3 (Table 1) which are unknown to the participants. For every novelty we introduce, we generate game levels that each include this one particular novelty. Game levels can include more than one of the same novel object.

We evaluate the competition using the same evaluation protocol mentioned in section 3.7. This means that the competition consists of multiple trials. Each trial T_i^a is dedicated to one specific novelty a and consists of a sequence of n different Angry Birds games. The first m_i games of a trial are standard games without novelty, the following $n-m_i$ games are games with novelty. Neither n nor m_i are known to participants, m_i can be between 0 and n, that is, a trial might consist of only standard games, only novel games, or a sequence of standard games followed by novel games. Each game in a trial can only be played once, the games in a trial have to be played in the given order. There is a time limit per trial. Agents are required to report for every game in a trial if they believe the novelty switch has happened or not. This is a value between 0 and 1, where any value above 0.5 is interpreted as the trial has switched to novel games. For each game we record the solved score that has been achieved by the agent. If a game has been solved (=all pigs have been killed) the solved score is equal to the game score. If a game has not been solved, the solved score is 0.

For each agent we measure the following:

- For each novelty level, we use the aggregated solved score as *task performance measure*. This is the sum of the solved scores of each game that contains novelty (or of all games for novelty level 0).
- For each novelty level 1-3, we measure the *percentage of correctly detected trials* (%CDT). These are trials where the agent reports that novelty has been detected for the first time for a novel game (no false positives and at least one true positive, if novel games were present).
- For each novelty level, we also measure the *average number of novel games needed* to detect novelty (#NGN). For a given trial, if game #5 is the first novel game, but novelty has only been detected in game #10, then the number of novel games needed for this trial is 6. This is only recorded for correctly detected trials with novel games.

For each novelty level we determine the agent with the *highest aggregated solved score*, as well as the agent with the *highest novelty detection score* $= \%CDT * (MAX\_NGN - \#NGN)$, where MAX_NGN is the NGN value when novelty is detected at the last game of a trial, averaged over all trials. The winner of the competition is the agent with the highest aggregated solved score across all the novelty levels 1-3. There are subcategories for the best performing agent for each of the four novelty levels, and a special award for the agent with the best novelty detection score across all novelty levels 1-3, as well as subcategories for each of the three novelty levels.

The task of the agents remains to solve each level, i.e., to kill all the pigs with as few birds as possible. How to solve this task can change significantly when novelty is introduced, and it is possible that agents unable to deal with novelty cannot solve any games anymore. Agents need to detect the novelty and adjust to it. Each trial contained between 0 and 10 non-novel games, followed by exactly 40 novel games. These settings were unknown to participants. In addition to solving the games, agents also had to report when they believe novelty has been introduced. Therefore, agents are evaluated on two aspects: (1) their novelty detection performance, which is based on the percentage of trials where they correctly detect novelty and on the number of novel games they need before they can detect it. (2) their novelty reaction performance, which is the overall game score they received in the novel games. Given that we used 6 novelties, ten trials per novelty, plus ten trials for no-novelty, and each trial consists of around 50 games, i.e., each agent had to play around 3500 games.

Agent	#NGN	%CDT	Detection Score	Total Solved Score
BamBirds	1	6%	2.44	36192880
CIMARRON	3.23	67%	24.52	**49653730**
Dongqing 1	3.13	67%	24.58	41231190
HYDRA	9.86	83%	25.12	37766360
OpenMIND	1.49	68%	**26.32**	28427980
Shiro	1	5%	1.95	19488940

Table 2: AIBIRDS 2021 Novelty Track Overall Results

Agent	Level 1	Level 2	Level 3
BamBirds	1.95 , **25265200**	2.79 , 6563240	2.79 , 4364440
CIMARRON	**36.40** , 24020150	18.25 , **16386620**	18.90 , **9246960**
Dongqing 1	21.00 , 24139240	**34.75** , 9674290	18.00 , 7417660
HYDRA	21.45 , 19042980	27.05 , 10504580	26.85 , 8218800
OpenMIND	19.5 , 17651480	20.45 , 6919110	**39.00** , 3857390
Shiro	1.95 , 13247470	1.95 , 4335210	1.95 , 1906260

Table 3: AIBIRDS 2021 Novelty Track Detection and Reaction Results by Novelty Level. Detection performance is followed by the reaction performance after the comma ",".

4.1 Competition Results

We had six teams who participated in this extremely challenging competition: **BamBirds** from the University of Bamberg, who was the winner of the standard track of the previous competition in 2019. **CIMARRON** from the University of Massachusetts Amherst, **Dongqing 1** from Bytedance and Monash University, **HYDRA** from the Palo Alto Research Center and the University of Pennsylvania, **OpenMIND** from Smart Information Flow Technologies, and **Shiro** from NIAD-QE.

Table 2 shows the agent with the best novelty reaction performance is **CIMARRON**. Second place went to **Dongqing 1**, third place to **HYDRA**. The agent with the best novelty detection performance is **OpenMIND**.

It is also interesting to notice that **CIMARRON** dominating the novelty reaction performance over all three novelty levels with achieving 3rd place in level 1 and 1st place in level 2 and 3, while there's no dominating winner in novelty detection (Table 3). Different agents were good at detecting each different novelty level. For example, **CIMARRON** is much better than others in detecting level 1 novelties while **Dongqing 1** and **OpenMIND** are better in level 2 and 3.

Acknowledgments

This research was sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO) and was accomplished under Cooperative Agreement Number W911NF-20-2-0002. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the DARPA or ARO, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Bear, D. M.; Wang, E.; Mrowca, D.; Binder, F. J.; Tung, H.-Y. F.; Pramod, R. T.; Holdaway, C.; Tao, S.; Smith, K.; Sun, F.-Y.; Fei-Fei, L.; Kanwisher, N.; Tenenbaum, J. B.; Yamins, D. L. K.; and Fan, J. E. 2021. Physion: Evaluating Physical Prediction from Vision in Humans and Machines. arXiv:2106.08261.

Borovička, T.; Špetlík, R.; and Rymeš, K. 2014. DataLab Angry Birds AI. http://aibirds.org/2014-papers/datalab-birds.pdf. [Accessed: July. 31, 2021].

Boult, T.; Grabowicz, P. A.; Prijatelj, D.; Stern, R.; Holder, L.; Alspector, J.; Jafarzadeh, M.; Ahmad, T.; Dhamija, A. R.; Cli; Cruz, S.; Shrivastava, A.; Vondrick, C.; and Scheirer, W. 2021. Towards a Unifying Framework for Formal Theories of Novelty. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35: 15047–15052.

Choi, D.; Morgan, M.; Park, C.; and Langley, P. 2007. A testbed for evaluation of architectures for physical agents.

Ferreira, L.; and Toledo, C. 2014. A Search-based Approach for Generating Angry Birds Levels. In *Proceedings of the 9th IEEE International Conference on Computational Intelligence in Games*, CIG'14.

Langley, P. 2020. Open-World Learning for Radically Autonomous Agents. In *AAAI*.

Langley, P.; Nicholas, D.; Klahr, D.; and Hood, G. 1981. A Simulated World for Modeling Learning and Development. 274–276. Berkeley, CA.

Nikonova, E.; and Gemrot, J. 2019. Deep Q-Network for Angry Birds. *CoRR*, abs/1910.01806.

Pinto, V.; Renz, J.; Xue, C.; Doctor, K.; and Aha, D. 2020. Measuring the Performance of Open-World AI Systems.

Renz, J.; Ge, X.; Stephenson, M.; and Zhang, P. 2019. AI meets Angry Birds. *Nature Machine Intelligence*, 1.

Senator, T. 2019. Science of Artificial Intelligence and Learning for Open-world Novelty (SAIL-ON). https://www.darpa.mil/program/science-of-artificial-intelligence-and-learning-for-open-world-novelty.

Stephenson, M.; and Renz, J. 2017. Generating varied, stable and solvable levels for Angry Birds style physics games. In *2017 IEEE Conference on Computational Intelligence and Games (CIG)*, 288–295. IEEE.

Stephenson, M.; Renz, J.; Ge, X.; Ferreira, L.; Togelius, J.; and Zhang, P. 2018a. The 2017 aibirds level generation competition. *IEEE Transactions on Games*, 11(3): 275–284.

Stephenson, M.; Renz, J.; Ge, X.; and Zhang, P. 2018b. The 2017 AIBIRDS Competition. *ArXiv*, abs/1803.05156.

Wang, T. J. 2017. AI Angry Birds Eagle Wing. https://github.com/heartyguy/AI-AngryBird-Eagle-Wing. [Accessed: July. 31, 2021].