NovPhy: A Physical Reasoning Benchmark for Open-world AI Systems

Vimukthini Pinto[a,c], Chathura Gamage[a,c], Cheng Xue[a,c], Peng Zhang[a], Ekaterina Nikonova[a], Matthew Stephenson[b], Jochen Renz[a]

[a] *School of Computing, The Australian National University, Canberra, Australia*
[b] *College of Science and Engineering, Flinders University, Adelaide, Australia*
[c] *These authors contributed equally*

Abstract

Due to the emergence of AI systems that interact with the physical environment, there is an increased interest in incorporating physical reasoning capabilities into those AI systems. But is it enough to only have physical reasoning capabilities to operate in a real physical environment? In the real world, we constantly face novel situations we have not encountered before. As humans, we are competent at successfully adapting to those situations. Similarly, an agent needs to have the ability to function under the impact of novelties in order to properly operate in an open-world physical environment. To facilitate the development of such AI systems, we propose a new benchmark, NovPhy, that requires an agent to reason about physical scenarios in the presence of novelties and take actions accordingly. The benchmark consists of tasks that require agents to detect and adapt to novelties in physical scenarios. To create tasks in the benchmark, we develop eight novelties representing a diverse novelty space and apply them to five commonly encountered scenarios in a physical environment, related to applying forces and motions such as rolling, falling, and sliding of objects. According to

our benchmark design, we evaluate two capabilities of an agent: the performance on a novelty when it is applied to different physical scenarios and the performance on a physical scenario when different novelties are applied to it. We conduct a thorough evaluation with human players, learning agents, and heuristic agents. Our evaluation shows that humans' performance is far beyond the agents' performance. Some agents, even with good normal task performance, perform significantly worse when there is a novelty, and the agents that can adapt to novelties typically adapt slower than humans. We promote the development of intelligent agents capable of performing at the human level or above when operating in open-world physical environments. benchmark website: `https://github.com/phy-q/novphy`

Keywords: Physical Reasoning, Open-world Learning, Novelty Benchmark, Novelty Detection, Novelty Adaptation, AI Evaluation

1. Introduction

A key aspect of human intelligence is the ability to reason about the physical behaviour of objects and make decisions in the physical environment [1]. Research suggests that humans develop physical reasoning capabilities within just the first year of birth [2, 3]. Even though it has been challenging to develop AI systems that can do general physical reasoning as good as humans do [4, 5], there is work that shows AI systems that could achieve human performance in some physical reasoning tasks [6, 7]. But the question is: is it adequate to only have physical reasoning capabilities to successfully work in the real physical world?

Encountering novel situations is an inherent characteristic of the real

world (Figure 1). As humans, we are adept at working in such novel situations that we constantly face in our day-to-day lives. For example, consider a person who can ride a bicycle on a normal road. On a rainy day when the roads are slick, the person can still ride the bicycle safely by adjusting the speed and applying the brakes smoothly without slipping. Even though sometimes a human may initially make a suboptimal decision when facing a novel situation, they usually quickly recover successfully after continuing to work for a short period of time under the influence of novelty. Similarly, for an agent that operates in an open-world physical environment, along with physical reasoning capabilities, it is crucial to possess capabilities that are required to handle novel situations [8], i.e., novelty detection and adaptation capabilities.

Figure 1: Example novelties that could be encountered in the real world. Left: self-checkout machines started appearing in supermarkets after the early 2000s, until then customers were used to traditional checkout methods, hence using self-checkout machines was a novelty to the customers [9]. Right: traffic accidents are generally novel situations for self-driving cars as such incidents are rare in the training data and usually visually unique in each incident [10].

In contrast to the intelligence of humans, current AI systems tend to struggle when they are presented with situations that were not available

during their training stage or if the situation was not anticipated by the developers [11]. This could be due to the fact that the research field, Open World Learning (OWL), attempting to address this issue is relatively new [12, 13, 14]. Apart from that, not having adequate testbeds/benchmarks to experiment and evaluate such AI systems also hinders their advancement. There are frameworks such as Monopoly [14], Polycraft [15], Cartpole [16], etc, that treat novelties as first-class citizens and facilitate agent experimentation. Even though some of them are physics based environments [16], none of them specifically focus on introducing novelties to the physical scenarios that an agent would encounter in the real world. Also, it is out of their context to evaluate agents in real-world physical tasks in the presence of novelties.

To fill the above gaps, we propose a new novelty-centric benchmark, NovPhy, where agents need to perform in real-world physical scenarios in the presence of novelties. NovPhy includes a wide variety of novelties applied to different physical scenarios. We implement our benchmark on the physics-based video game Angry Birds as it has realistic physics and is a versatile domain to introduce physics-based novelties. Moreover, Angry Birds is popular in both physical reasoning research [5, 17, 18] and OWL research [19, 20, 21]. The main contributions of this work can be summarized as follows:

- **NovPhy - A benchmark for novelty detection and adaptation in physical environments:** We consider five commonly encountered physical scenarios in NovPhy: applying a single force and mul-

tiple forces, and rolling, falling, and sliding objects. We developed eight novelties representing a diverse novelty space. We designed task templates by applying the eight novelties to the five physical scenarios separately, resulting in 40 novel task templates. A task template is used to generate related tasks by varying task template parameters such as the locations of the objects. We also created 40 corresponding normal task templates without novelties to facilitate our evaluation protocol. Further, we developed a task variation generator that can generate an unlimited amount of tasks from these task templates.

- **Agent evaluation setups for open-world physical environments:** We propose a comprehensive evaluation setup to evaluate the novelty detection and adaptation of AI systems in open-world physical environments. In this setup, the novelties are orthogonal to the physical scenarios, hence facilitating us to evaluate agents in two settings: first, the same novelty is applied separately to the tasks of multiple physical scenarios, and second, multiple novelties are applied separately to the tasks of the same physical scenario. The former is used to evaluate how well an agent can deal with the same novelty in different physical reasoning tasks and the latter is used to evaluate how well an agent can perform the same physical reasoning task under different novel situations.

- **Evaluation measures to ensure reliable novelty detection:** While novelty detection is not the primary focus of this paper, we introduce supplementary evaluation measures and statistical tests to comple-

ment existing detection evaluations. Our new evaluation ensures that agents do not exhibit detection bias or make random detections.

- **Establishing results for baseline agents:** We evaluate 11 baseline agents. Three heuristic-based agents, two standard online learning agents, two standard offline agents, three adaptive learning agents, and a random agent. We report their novelty detection and novelty adaptation performance.

- **Establishing baseline human performance:** In order to show our novelties are adaptable for humans, we conducted an experiment using human players in our benchmark. These results also show that humans can detect and adapt to novelties better and faster compared to AI agents, thus acting as a milestone performance for AI to achieve.

2. Background and Related Work

In this section, we discuss the background and related work regarding physical reasoning research, novelty theories, and the existing novelty-centric benchmarks/testbeds.

2.1. Physical Reasoning Research

Physical reasoning has become an important aspect of AI research due to the increased reliance on autonomous AI systems in day-to-day operations. There are multiple physical reasoning benchmarks and testbeds available to assist AI systems enhance physical reasoning capabilities in order to perform securely in the real world.

The physical reasoning benchmarks such as Physion [7], IntPhys [22], CLEVERER [23] are based on videos while COPHY [24] is an image-based physical reasoning benchmark. Physion comprises eight physical reasoning scenarios, including rolling, sliding, and projectile motion, which are important capabilities to work in the physical world [7]. IntPhys, on the other hand, is concerned with physical reasoning abilities acquired during infancy, such as object permanence, spatio-temporal continuity, and shape consistency [22]. The CLEVERER benchmark presents videos and asks questions inspired by the theory of human causal judgement [23]. COPHY benchmark presents a sequence of images based on physical scenarios and asks to predict the outcome if you make a modification to the initial image. COPHY is developed to test counterfactual reasoning applied to the physical world [24].

To bring physical reasoning abilities of AI systems closer to reality, researchers have created action based benchmarks that require agents to take an action in order to accomplish the goal. Examples of such benchmarks include PHYRE [25] and Virtual Tools [6]. PHYRE is a benchmark consisting of simple 2D physics based tasks, aimed to foster the development of efficient models capable of generalisation across tasks [25]. The Virtual Tools game, focuses on evaluating agents on selecting appropriate tools and taking the correct action using the tool to solve the tasks [6].

Phy-Q [5] is also an action based physical reasoning testbed that consists of a broad variety of 15 physical reasoning scenarios. Some physical scenarios in Phy-Q (applying single or multiple forces, rolling, falling, sliding) are inspired by physical reasoning abilities developed during childhood,

whereas some scenarios (adequate timing, clearing paths, and manoeuvring) are required to overcome challenges for robots to work safely in physical environments. As an improvement to the previously mentioned benchmarks, Phy-Q testbed supports different evaluation settings based on different generalization levels (local generalization and broad generalization) and also they have established human performance on the scenarios in the testbed. Furthermore, this testbed allows us to compute the physical reasoning quotient, which reflects an agent's physical reasoning aptitude.

Even though all the preceding benchmarks support the development of AI systems with advanced physical reasoning capabilities, none of them focuses on physical reasoning under novel circumstances, which is the setting an agent in the real world would frequently encounter. In our benchmark NovPhy, we combine physical reasoning scenarios with novel situations, to create a real-world like setting to evaluate the agent's performance. To introduce novel situations, we consider the first five physical reasoning scenarios from the Phy-Q tesbed: single force, multiple forces, rolling, falling, and sliding.

2.2. Theories of Novelty

AI systems have already shown superhuman performance in a wide range of closed-world domains [26, 27, 28]. However, compared to a closed-world, in an open world, the agents may struggle to perform due to encountering novel situations. Even though some novel situations can be predicted by the developers, some cannot be anticipated, making it impossible to integrate all the possible novel situations into an agent model. DARPA has

launched the Science of Artificial Intelligence and Learning for Open-world Novelty (SAIL-ON) program to investigate and develop the underlying scientific principles, general engineering techniques, and algorithms required to create AI systems that adapt appropriately when a novel situation arises [12, 29].

Researchers have looked at different directions to formalize what it means to be a novelty. The novel situations are sometimes referred to as anomalies or out-of-distribution data by some researchers [30, 15]. SAIL-ON program defines novelty as situations that violate implicit or explicit assumptions in an agent's model of the external world, including other agents, the environment, and their interactions [31, 14]. Alternatively, Langley describes novelty as transformations of the elements in the environment [13]. Examples of such transformations are spatio-temporal transformations, structural transformations, process transformations, and constraint transformations. In addition, Molineaux and Dannenhauer [32] formally define different environmental transformations. On the other hand, Boult et al [30], have introduced a unifying framework of novelty in the context of AI. Boult et al, focus on the world space, observation space, and the agent state to formally define diverse types of novelties.

A working group in the SAIL-ON program, the novelty working group (see [31] for participant list), has also developed a novelty hierarchy that enables the categorization of the novelties considering their properties. This categorization of novelties facilitates a solid novelty evaluation as it helps to design novelties covering a large novelty space and helps to identify different categories of novelties an agent model would fail to perform. In this paper,

we use the novelty hierarchy developed at the SAIL-ON novelty working group. The levels in this novelty hierarchy are objects, agents, actions, relations, interactions, environments, goals, and events [31]. Table 1 provides a description of the novelty hierarchy levels and representative novelties we have designed in NovPhy. In Section 3, we have a detailed discussion of our desiderata on how we designed novelties and novel tasks in a physical environment such that it allows us a comprehensive agent evaluation.

2.3. Novelty-centric Domains

The availability of novelty-centric domains that facilitate agents to evaluate and compare agents' performance is a critical factor that contributes to the advancement of open-world learning. Several testbeds/ frameworks/ benchmarks have been introduced to facilitate OWL by considering novelties as first-class citizens. This section describes related research on domains for OWL and situates NovPhy within them.

GNOME is a novelty-centric simulator tool that facilitates developing and evaluating AI systems in multi-agent environments such as strategic board games [14]. GNOME is applied to the popular board game Monopoly to inject novelties from the first three levels of the novelty hierarchy described above. Example novelties from GNOME are adding more dice to the game, shuffling the order of slots on the board, etc.

NovGrid [33] and NovelGridworlds [34] are two frameworks developed for grid environments. NovGrid is a toolkit developed for novelty generation in MiniGrid environment [35]. NovGrid extends the MiniGrid environment by enhancing the functionality of existing objects and enabling agents to detect

and adapt to novelty. NovelGridworlds implements a Minecraft [36] inspired grid world to study novelties. In NovGrid, authors have introduced novelties based on the first two novelty hierarchy levels while in NovelGridworlds authors introduce novelties based on the first three levels of the novelty hierarchy.

NovelCraft [15] is a benchmark dataset for novelty detection and adaptation based on a modified version of Minecraft. NovelCraft is a 3D environment where an agent needs to select a sequence of actions to turn available resources into a pogo stick [37]. In this work, authors have established agent performance in novelties based on the first level of the novelty hierarchy: objects [15].

None of the above-mentioned novelty domains is based on physics, which is an important characteristic to consider when developing agents that work in the real world. Cartpole with novelty [16] and Science Birds Novelty [19] are two physics based domains developed for novelty detection and adaptation. In Cartpole, an agent must keep the pole balanced by pushing the cart forward, backward, left, or right. In Science Birds Novelty, which is based on the physics game Angry Birds [38], an agent needs to destroy the pigs by shooting birds from a slingshot. NovPhy uses the Science Birds framework [19] and introduces various realistic physical reasoning tasks, as our focus is on evaluating agents' physical reasoning ability under the influence of novelty. As mentioned in Section 2.1, we use five physical reasoning scenarios and combine them with novelties from all eight levels of the novelty hierarchy.

3. Designing Novel Tasks and Agent Evaluations in Open-world Physical Environments

In this section, we discuss the desiderata we satisfied when designing tasks in our benchmark and when setting up agent evaluations for those tasks. We term the tasks that have novelties in them as novel tasks and the tasks without novelties (i.e., tasks in the normal environment) as normal tasks. In our benchmark, we consider a constrained environment setting where an agent's action only applies to the state of the environment when the action is taken, and that action determines the subsequent states of the environment (i.e., the agent cannot control the subsequent states of the environment after the action is taken).

3.1. Designing Novel Tasks

An agent working in a physical environment has to encounter different physical scenarios such as applying a force to an object, moving an object from one place to another, avoiding an obstacle in its path, etc. We introduce novelties to these scenarios that an agent could encounter in the physical environment.

When designing novel tasks, **we ensure that the agent has to work under the effects of the novelty to solve the task**. In other words, in the novel tasks, there are no solutions to the task that skip the effects of the novelty. For example, consider a bowling game where the player has to roll a ball on a surface to knock over the pins. Assume that when novelty occurs, the surface on which the ball rolls becomes slippery. In this scenario, the player is required to interact with the novel element (the surface) in order

to complete the task. Therefore, the only way to successfully perform this task is to adapt to roll the ball on the slippery surface.

To ensure that the agent has to go through the novelty when completing a task, when designing novelties we consider the physical interactions in the solution of a physical scenario. We categorize these physical interactions into three phases: the initial phase, the middle phase, and the final phase. We only design novelties that at least affect one of these three interaction phases, to guarantee that the agent has to work along with the effects of the novelty. The initial phase includes the immediate impact on the objects by the agent's action, the middle phase includes the consequences of the immediate impact of the action, and the final phase includes the interactions that complete the task. In all the phases, the objects that are involved in those physical interactions are also considered.

For example, consider the previously mentioned bowling game. In this scenario, the possible physical interactions are, the player throws the ball giving a starting velocity to the ball, the ball rolls on the surface, and the ball hits the pins knocking them down. In this instance, the initial interaction phase includes the ball and the velocity the player applies to the ball. The middle phase includes the ball and the surface, and the rolling movement of the ball. The final phase includes the ball and the pins, and the collision between the ball and the pins. Therefore, here, the novelty can be applied to the ball, to the surface, to the pins, or to something that affects the rolling of the ball and collision of the ball and pins, in order to make sure the agent has to bowl under the effect of the novelty.

3.2. Designing Agent Performance Evaluations

In OWL, agent evaluations are conducted to measure two capabilities of agents: the novelty detection capability and the novelty adaptation capability [39, 40, 41]. The novelty detection measures evaluate whether an agent could successfully detect a novelty in the environment and the novelty adaptation measures evaluate whether an agent could successfully perform the task in the presence of the novelty. Novelty adaptation is generally more emphasized than novelty detection, as performing the task under the influence of a novelty is more important than merely detecting something is novel for an agent that actually works in an open-world environment. In this work, we also prioritize evaluating agents' novelty adaptation performance.

The standard agent evaluation setup in OWL consists of a set of trials, where each trial consists of a sequence of normal tasks followed by a sequence of novel tasks [39, 41, 42]. We follow the same setup in NovPhy. We believe that, in order to measure whether an agent genuinely adapts to a novelty, **there should be a change in the solution path of the tasks when moving from the normal tasks to the novel tasks**. In an abstract form, we define a solution path as a sequence of physical interactions including the associated objects, initiated by an agent's action, that leads to solving the task. When designing a novel task for a physical scenario we also design a corresponding normal task that has a different solution path compared to the novel task. Then, when defining the trials for the evaluation, we select these normal and novel task pairs to guarantee that there is a solution path change from normal to novel tasks.

In this setting, since there is an obvious change in the solution path from

the normal tasks to the novel tasks, novelty detection becomes trivial. To detect whether there is a novelty, the agent has to simply monitor whether the solution in the normal tasks is no longer working. To avoid this consequence, one could define separate trials that are only used to evaluate the novelty detection performance by including the tasks that have the same solution path in both normal and novel tasks. In this work, we do not include such trials in the evaluation as our main focus is evaluating the novelty adaptation performance of the agents.

We consider another desideratum when evaluating the performance of an agent in an open-world physical environment. From the perspective of OWL, we believe that, **if an agent can truly perform under a novelty, the agent should be able to perform with that novelty when the novelty is applied to different physical scenarios**. Also, from the perspective of physical reasoning, we believe that, **if an agent is robust at performing in a physical scenario, that agent should be able to perform in that scenario under the effect of different novelties**. To achieve these two evaluation setups, we designed the novelties orthogonally to the physical scenarios, such that the same novelty can be applied to multiple scenarios and the same scenario can get affected by multiple novelties.

4. NovPhy benchmark

In this section, we introduce our benchmark, the physical scenarios we consider, the novelties we designed, the tasks in the benchmark, and explain the evaluation settings we have used in the benchmark.

4.1. Introduction to NovPhy

Based on different physical scenarios and novelties, we develop the novelty-centric benchmark NovPhy using Angry Birds. We use an open-source research clone of the game developed in Unity called Science Birds [43]. Our benchmark is adapted from a framework that can be used to inject novelties and conduct agent evaluations, developed from Science Birds [19].

In Angry Birds, the goal of the player is to kill all the pigs in the game level by shooting a given number of birds from a slingshot. In the normal game environment, along with the slingshot, the player will encounter four types of game objects: birds, pigs, blocks, and platforms. Additionally, we have also introduced an external agent to the normal environment called Air Turbulence that applies an upward force to any object that travels through it. An external agent is an agent with goal-oriented behaviour and having external agents enables situations that hinder or support the action that a player takes. In the game, birds, pigs, and blocks are dynamic objects, which behave according to Newtonian physics, while platforms are static and are not affected by external forces. The dynamic objects have health points that get reduced in the collisions and they get destroyed when the health points become zero. The blocks have 12 variations in shape and they are made of one of 3 types of materials: wood, stone, and ice. There are three types of pigs with three different sizes; the larger the size the higher the health points are. All objects in NovPhy are shown in Appendix Figure A.10.

An agent playing the game can request the current game state anytime as a screenshot and/or as a symbolic representation. The screenshot is a

480x640 coloured image of the game. The symbolic representation is in JSON format and contains all the objects in the screenshot. Here, an object is represented as a polygon of ordered vertices along with the percentages of its 8-bit quantized colours. The full world state is not provided to the agent such as the exact positions of the objects and their physical parameters such as mass, coefficient of friction, etc. as they are not directly observable in the real world. The action of an agent is the release point of the bird relative to the slingshot. Sometimes when there is more than one bird in the game level the agent takes a sequence of actions. We provide a trajectory planner that can be used to calculate the release point of the bird to reach a target, under the normal settings in the environment, when the target point is given. The agent passes the game level if it destroys all the pigs with the provided number of birds or fails if not.

The physical scenarios and the novelties we consider in this benchmark are discussed in the next subsections 4.2 and 4.3 respectively.

4.2. Physical Scenarios in NovPhy

As discussed in Section 2, we use the first five physical scenarios introduced in Phy-Q as they are the most basic and frequently encountered scenarios in a physical environment. The scenarios include applying forces directly on target objects - the effect of a single force and the effect of multiple forces [44]. The motion-related scenarios: rolling, falling, and sliding, inspired by the physical reasoning capabilities developed in human infancy [45]. The five scenarios and the corresponding physical rules that can be used to achieve the goal of the associated tasks are:

1. Single force: Target objects have to be destroyed with a single force.
2. Multiple forces: Target objects have to be destroyed with multiple forces.
3. Rolling: Circular objects have to be rolled along a surface to a target.
4. Falling: Objects have to fall onto a target.
5. Sliding: Non-circular objects have to be slid along a surface to a target.

4.3. Novelties used in NovPhy

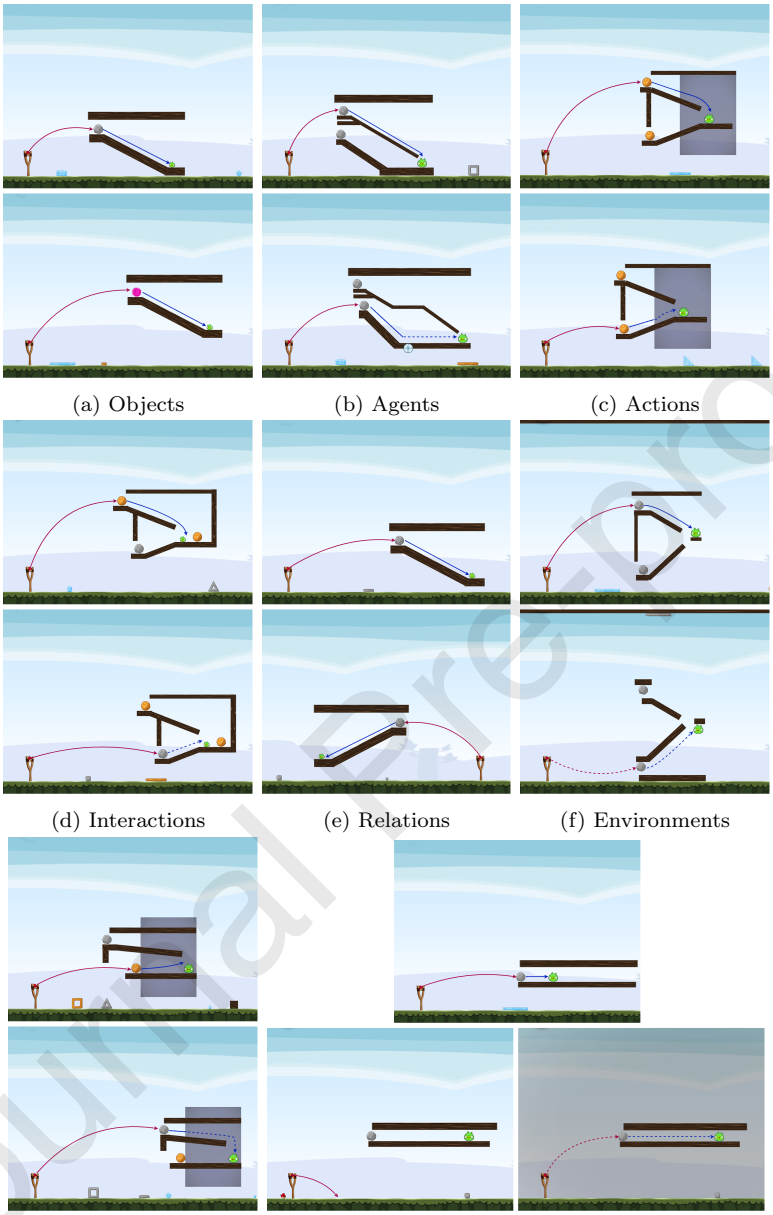
We design a representative novelty for each hierarchy level in the open-world novelty hierarchy proposed by the SAIL-ON program novelty working group. The novelty hierarchy consists of eight novelty levels that cover a wide range of novelty types that could occur in an open-world environment. Table 1 shows the open-world novelty hierarchy and descriptions of representative novelties in NovPhy. Appendix Figure A.11 shows the new game objects that are introduced to the game for the novelties associated with a game object.

4.4. Task Templates

A task template defines a set of related tasks that can be created by varying task template parameters such as the locations of the objects. We design task templates by applying each of the eight novelties to each of the five physical scenarios discussed in the above sections. For example, for the

Novelty Level	Description	Representative Novelty
1. Objects	New classes, attributes, or representations of non-volitional entities.	A new pig/block that has a different colour to the normal pigs/blocks.
2. Agents	New classes, attributes, or representations of volitional entities.	A novel external agent, Fan, that blows air (horizontally from left to right) affecting the moving path of objects.
3. Actions	New classes, attributes, or representations of external agent behavior.	The non-novel external agent, Air Turbulence, increases the magnitude of its upward force.
4. Interactions	New classes, attributes, or representations of dynamic properties of behaviors impacting multiple entities.	Existing circular wood object now has magnetic properties: repels objects of its type and attracts other object types.
5. Relations	New classes, attributes, or representations of static properties of the relationships between multiple entities.	The slingshot which is at the left side of the tasks is now at the right side of the tasks (i.e., the spatial relationship between the slingshot and other objects is changed).
6. Environments	New classes, attributes, or representations of elements independent of specific entities.	The gravity in the environment is now inverted, which affects the behaviour of the dynamic objects.
7. Goals	New classes, attributes, or representations of external agent objectives.	The non-novel external agent, Air Turbulence, changes its goal from pushing objects up to pushing objects down.
8. Events	New classes, attributes, or representations of series of state changes.	When the first bird is dead, a storm occurs that affects the motion of the objects (by applying a force to the right direction).

Table 1: SAIL-ON Open-world novelty hierarchy [31] and the representative novelties in NovPhy for each hierarchy level.



(a) Objects (b) Agents (c) Actions

(d) Interactions (e) Relations (f) Environments

(g) Goals (h) Events

Figure 2: Example tasks of the rolling scenario for the eight novelties. In each subfigure, the top figure is the normal task and the bottom figure is the corresponding task with the novelty. The arrows show the trajectories of the objects when the solution is executed.

rolling scenario we design templates that require rolling an object when: it is a new object, there is an effect from the Fan, the Air Turbulence agent changes the magnitude of the upward force, the slingshot is in the right side of the object, there is an effect from a magnetic field, the gravity is inverted, Air Turbulence pushes the object down, and there is a storm after shooting the first bird. We term a physical scenario with a novelty applied as a novelty-scenario. Since we consider eight novelties and five scenarios, we have 40 novelty-scenarios.

When designing novel task templates for NovPhy, we follow the desiderata discussed in Section 3. For each novelty-scenario, we designed and handcrafted a novel task template. A novel task template is a task template where a novelty is present. In the design, we ensure the novel tasks meet our desideratum, that it is necessary to work under the effects of the novelty to complete the task. Appendix Table C.4 contains more details on how this desideratum is satisfied when designing the tasks, considering the physical interaction phases of the solution that are affected when the novelty is introduced. Then, for each novel task template, we design a corresponding normal task template as well. This is done by removing the novelty from the novel task template and adjusting the template accordingly to make it solvable without the novelty. Considering our next desideratum, when designing the normal task templates, we also ensure that there is a solution path change from the normal task to the novel task.

In the task template design, we also ensure that all the templates of a given scenario can be solved by the associated physical rule of that scenario discussed in Section 4.2. Figure 2 shows how the eight novelties are applied

to the tasks of the rolling scenario. In each subfigure, the top figure is the normal task and the bottom figure is the corresponding task with the novelty. The arrows show the trajectories of the objects when the solution is executed (in red: bird's trajectory, in blue: other objects' trajectories when the bird is hit). The dotted arrows represent the trajectories affected by the novelty. To solve the novel task shown in each subfigure:

(a) objects: the new pink coloured object has to be rolled.

(b) agents: the force of the new Fan agent has to be used to roll the object further.

(c) actions: the magnitude of the upward force of the Air Turbulence agent is increased, which helps to roll the ball upwards in the ramp.

(d) interactions: the circular wood (brown) objects have magnetic properties, thus repels the object of the same type and attracts the stone (grey) object helping to roll the stone upwards in the ramp.

(e) relations: the slingshot is now placed on the right side of the task, instead of shooting left to right now the shooting has to be done from right to left to roll the object in the left direction.

(f) environments: the inverted gravity makes the objects attract towards the sky, hence objects can be rolled upwards in ramps.

(g) goals: the goal changed Air Turbulence agent (from the goal of pushing objects up to pushing objects down) hinders rolling on flat surfaces while helping to roll on inclined surfaces.

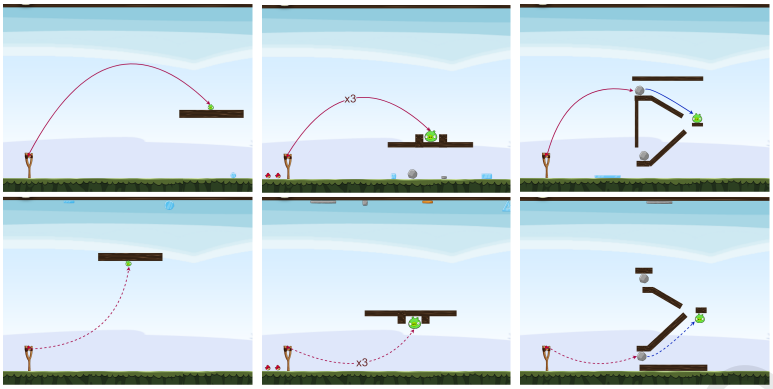
(h) events (the novel template has two birds as shown in the first bottom figure and when the first bird dies it activates the storm as shown in the second bottom figure): when the first bird is wasted, the storm occurs which applies a force to the right direction of the moving objects, hence shooting the second bird to the circular object makes the object to roll further to reach the pig.

Figure 3 shows how the inverted gravity novelty is applied across the tasks of the five physical scenarios. All 40 task templates in NovPhy can be found in Appendix B.

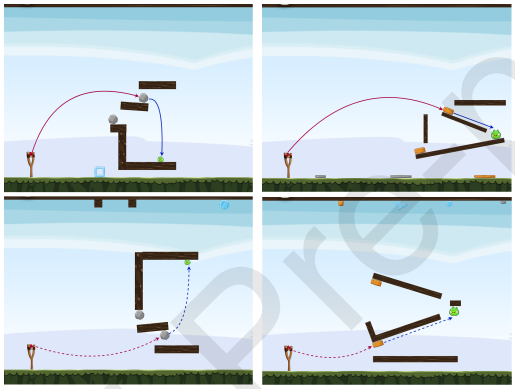
4.5. Task Generation

We developed a task generator that can generate an unlimited number of tasks from a given template. The game levels generated from a task template are termed as the tasks of that template. When generating the tasks we vary the locations of the game objects within a suitable range in the level space. Additionally, some random game objects are added as distraction objects at random positions of the game level to trick the agents. In the generation, we ensure that the task can still be solved by the solution path in the original template. To achieve this, we define template specific constraints such as, which game objects are reachable/unreachable to the bird, which objects should be excluded from the distraction objects, what are the feasible regions to place specific objects, etc. These constraints are determined by the template designers and are input to the task generator.

We provide 350 generated tasks for each task template, but we also provide the task generator in case it is necessary to generate more tasks.



(a) Single force (b) Multiple forces (c) Rolling



(d) Falling (e) Sliding

Figure 3: Example tasks of the inverted gravity novelty applied to the five physical scenarios. In each scenario, the top figure is the normal task and the bottom figure is the novel task. The arrows show the trajectories of the objects when the solution is executed. Solid arrows are the trajectories of the objects that are not affected by the novelty and the dotted arrows are the trajectories of the objects that are affected by the novelty. The inverted gravity has made all the dynamic objects attract towards the sky, hence they have been stopped from platforms to avoid flying away. The motion of the dynamic objects is also affected by the inverted gravity.

Appendix Figure B.17 shows the variations of the tasks generated from a single scenario template with a single novelty applied.

4.6. Evaluation Protocol

In NovPhy benchmark, as in a standard OWL evaluation, we evaluate the novelty detection and novelty adaptation capabilities of the agents. In the novelty detection evaluation, we measure if an agent can detect if a novelty is present in the given task. In novelty adaptation evaluation, we measure the task performance of the agent in the presence of a novelty. Both novelty detection and novelty adaptation evaluation are done by using a trial setting [39]. A trial is a sequence of tasks, which starts from normal tasks and after a random number of normal tasks switches to novel tasks. After switching to novel tasks, all the subsequent tasks until the end of the trial are novel tasks. Figure 4 shows how evaluations are done through the trial setup. A trial-set is a set of trials. A given trial-set consists of trials of the same novelty-scenario. i.e., all the trials of a given trial-set only have normal and novel tasks from the same novelty-scenario. The agent is not allowed to share knowledge in between trials, i.e., at the start of each trial of a novelty-scenario, the agent is in the same initial state, as the agent was at the beginning of the evaluation.

To evaluate an agent on a novelty-scenario, the agent is trained on the normal task distribution of that novelty-scenario and tested using a trial-set of that novelty-scenario. This evaluation resembles the local generalization evaluation of the agents, i.e., the agent trains on the tasks of a normal template and is tested on the tasks of the same template and the corresponding novel template. Even though the benchmark facilitates broad generalization evaluations (i.e., the agent trains on the tasks of a normal template of a physical scenario. Then the agent is tested on the tasks of a different

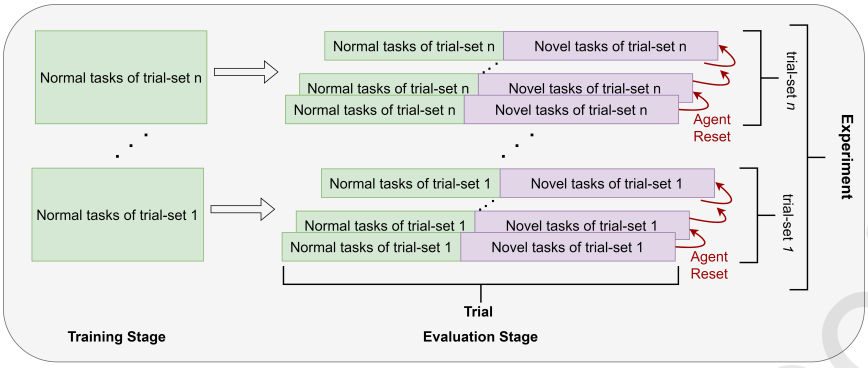


Figure 4: The trial-based evaluation protocol used in NovPhy. The evaluation stage follows the training stage. An experiment contains *trial-sets* where a trial-set contains multiple trials. A trial contains a variable number of tasks drawn first from the normal task distribution and then from the novel task distribution. In a given trial-set the agent is only evaluated on a single novelty-scenario.

template of the same scenario and its corresponding novel template), we use this local generalization based evaluation setup as it is proven that learning agents still struggle to generalize broadly in physical reasoning [5]. Moreover, it is a precondition for agents to have a good normal task performance before adapting to novel tasks.

When the agent is playing a trial, the task completion status (whether the task is passed/failed) and the score the agent achieved at the end of the task are recorded. This data is used to calculate the novelty adaptation performance of the agent. For the novelty detection performance calculation, the agent has to inform in which task of the trial it believes the novelty occurred.

In this work, we focus on evaluating agents in the below two evaluation settings.

1. Novelty Informed Evaluation: In this evaluation, an agent will be in-

formed when the novelty appears in each trial. The agent will only be evaluated on the novelty adaptation ability.

2. Novelty Uninformed Evaluation: In this evaluation, an agent will be evaluated on both novelty detection and novelty adaptation. The agent will not be informed when the novelty appears in a given trial.

4.7. Evaluation Measures

4.7.1. Novelty Detection Evaluation

For novelty detection, we use standard OWL measures used in the SAIL-ON program: the percentage of correctly detected trials (CDT) and the detection delay (DD) calculated using the average number of tasks taken to detect the novelty [39]. Consider a trial $t \in T$, where T represents a set of trials for a novelty-scenario, FP_t represents the number of normal tasks in trial t where the agent incorrectly detected as a novel task. TP_t represents the number of novel tasks in trial t where the agent correctly detected as a novel task. A correctly detected trial is a trial where the agent detected novelty only after entering the novel task sequence. CDT is defined in Equation 1. CDT varies between 0 and 1 and 1 is the best result.

$$CDT = \frac{1}{|T|} \sum_{t=1}^{|T|} \begin{cases} 1, & \text{if } FP_t = 0 \text{ and } TP_t \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

DD quantifies the delay in detection using the number of tasks required to correctly detect the novelty. DD is defined in Equation 2. The lower the DD, the better the detection performance (in terms of timeliness), and the best possible DD is 1.

$$DD = \frac{1}{N_{cdt}} \sum_{t=1}^{|T|} \begin{cases} d_t, & \text{if } FP_t{=}0 \text{ and } TP_t \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where,

$$N_{cdt} = \sum_{t=1}^{|T|} \begin{cases} 1, & \text{if } FP_t{=}0 \text{ and } TP_t \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

and d_t is the number of novel tasks taken until the agent informs a novelty detection in trial t (including the task that the agent detected novelty).

In addition to these existing evaluation measures, we propose to include a set of non-novel trials in the evaluation and statistical tests to assess the reliability of detection. In Appendix G we have discussed the limitations of the above-mentioned standard evaluation using additional experiments. A non-novel trial t is a trial where all instances i are drawn from the normal task distribution. Similar to the existing evaluation setup, an agent will attempt to solve each instance in the given order and report the probability that it believes the distribution shift has occurred.

We use the set of non-novel trials to identify undesirable patterns in the agent's behavior, such as random detections or detection bias. In random detection case, an agent that performs random detections outputs a detection probability exceeding the detection threshold at random instances, without genuinely detecting any novelty. In trials with a low number of non-novel instances, this behavior increases the likelihood of correct detections, resulting in low false positives but higher true positives. Consequently, a high percentage of trials would be marked as correctly detected, leading to an increased CDT, even though the agent's detection was merely random.

In the detection bias case, an agent reports a detection probability p_i exceeding the threshold towards the trial's end, without genuinely detecting novelty. The existing measures are unable to be used to differentiate such AI systems. For example, an agent might have an inbuilt rule to detect novelty after a certain number of instances if it does not detect anything novel within that period. This rule could be based on the agent developer's prior knowledge about the domain and the trial setting.

Therefore, using the non-novel trials, we calculate the percentage of incorrectly detected trials and the average number of instances needed for incorrect detection. We recommend evaluators compare the distribution of instances needed for detection in non-novel trials ($D_{non-novel-distribution}$) with the distribution of instances needed for detection in novel trials ($D_{novel-distribution_t}$) using a Kolmogorov–Smirnov test (KS test)[46].

We define a wrongly detected trial as follows.

$$WDT = \frac{1}{|T|} \sum_{t=1}^{|T|} \begin{cases} 1, & \text{if } FP_t \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

We define the instances needed for detection in trial t as follows.

$$INFD = n_t, \text{ if } FP_t \neq 0 \text{ or } TP_t \neq 0 \tag{5}$$

where,

$$n_t = \begin{cases} \textit{TN instances until the first FP instance,} & \text{if } FP_t \neq 0 \\ \textit{TN + FN instances until the first TP instance,} & \text{if } FP_t = 0 \textit{ and } TP_t \neq 0 \end{cases} \tag{6}$$

The number of instances needed for detection is defined irrespective of whether the trial is a correctly detected trial or whether it is a wrongly detected trial. Thus, comparing the $D_{non-novel-distribution}$ with the $D_{novel-distribution_t}$ enables evaluators to reliably comment on the detection ability of the agent. As an additional test, we recommend evaluators conduct a non-parametric Mann-Whitney test [46] to determine if the two distributions have the same median. If an agent takes the same median number of instances for detection in a novel distribution and in a non-novel distribution, it indicates that the agent may have a rule to detect after a certain number of instances without actually detecting a novelty.

4.7.2. Novelty Adaptation Evaluation

To measure novelty adaptation performance, we use the area under the pass rate curve (success curve). First, in a given novelty-scenario, to measure the performance of the agent after adapting to novelty, we use the pass rate of the asymptotic tasks. We refer to this measure as the asymptotic performance (AP). In equation 7 for AP, n represents the length of the novel task sequence and m represents the asymptotic length we consider. The asymptotic length can be adjusted based on the percentage of novel tasks in the trial.

$$AP = \frac{1}{m}\sum_{i=n-m}^{n}\left(\frac{1}{|T|}\sum_{t=1}^{|T|}\begin{cases} 1, & \text{if } i^{th} \text{ novel task in } t^{th} \text{ trial is passed} \\ 0, & \text{otherwise} \end{cases}\right) \tag{7}$$

Second, as AP does not capture the timeliness of adaptation, we compute

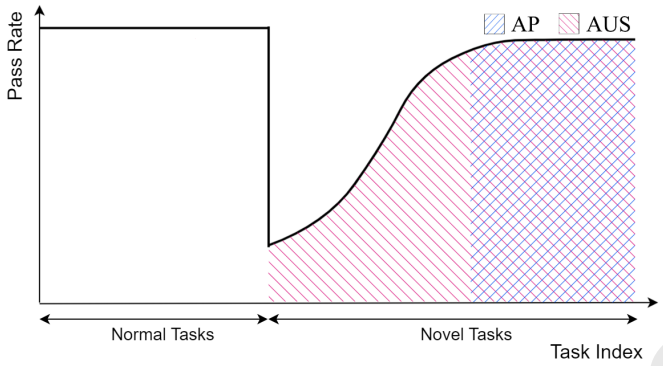


Figure 5: An example pass rate curve of an agent that played a set of trials. The area shaded in blue is considered for the asymptotic performance (AP) and the area shaded in red is considered for the area under success curve performance (AUS).

the total area under the success curve (AUS). AUS can be defined as follows.

$$AUS = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{|T|}\sum_{t=1}^{|T|}\begin{cases} 1, & \text{if } i^{th} \text{ novel task in } t^{th} \text{ trial is passed} \\ 0, & \text{otherwise} \end{cases}\right) \quad (8)$$

Both AP and AUS vary between 0 and 1 and 1 is the best achievable adaptation performance. Figure 5 graphically illustrates the areas considered in the pass rate curve of an agent when calculating AP and AUS.

Deriving from the above measures for novelty detection and novelty adaptation, we define *NPS* (Novelty Performance in Scenarios) and *SPN* (Scenario Performance under Novelties). *NPS* measures how agents perform in a single novelty, when the novelty is applied to different physical scenarios. *SPN* measures how agents perform in a single physical scenario, when different novelties are applied to the scenario. To calculate the *NPS* measures, the average performance is taken across all scenarios for a single

novelty. Formally, consider i to represent the i^{th} novelty, and j to represent the j^{th} scenario. The NPS measures, $NPS_{CDT,i}$, $NPS_{DD,i}$, $NPS_{AP,i}$, and $NPS_{AUS,i}$ are defined for a given novelty i as the average of CDT_{ij}, DD_{ij}, AP_{ij}, and AUS_{ij} respectively for all the scenarios j with the novelty i. Additionally, we conduct the KS test and Mann-Whitney test based on non-novelty trials and considering all trials in a given novelty i, denoted as $NPS_{KS,i}$ and $NPS_{MW,i}$.

Similarly, for *SPN* measures, for a given scenario j, the average performance is taken across all the novelties i for the scenario j. That is, *SPN* measures $SPN_{CDT,j}$, $SPN_{DD,j}$, $SPN_{AP,j}$, and $SPN_{AUS,j}$ are calculated from the average performance taken for CDT_{ij}, DD_{ij}, AP_{ij}, and AUS_{ij} respectively from all the novelties i in the scenario j. Similar to *NPS*, we conduct the KS test and Mann-Whitney test based on non-novelty trials and considering all trials in a given scenario j, denoted as $SPN_{KS,j}$ and $SPN_{MW,j}$.

5. Experiments

We conduct experiments on baseline agents on the 40 novelty-scenarios to measure how agents detect and adapt to novelty in each of those novelty-scenarios. In addition, we establish human performance on all the novelty-scenarios. This section describes the baseline agents we provide and the experimental setups used for each experiment.

5.1. Baseline Agents

We include experimental results of 11 baseline agents which consist of three heuristic agents, seven learning agents, and a random agent.

Heuristic Agents. The heuristic agents are based on hard-coded physical rules developed by agent developers. All the agents were participating agents from the AIBIRDS competition [47], which is an annual competition held to find the best Angry Birds game-playing AI agent. Following is the list of heuristic agents evaluated in NovPhy.

- Datalab: Datalab is a planning agent that has six strategies. The strategies include destroying pigs, destroying physical structures, and shooting at round blocks. The agent selects which action to take based on the game objects available, possible trajectories, the bird on the sling, and the birds remaining [48].
- Eagle's Wing: Eagle's Wing agent selects from a suit of five strategies based on structural analysis. The five strategies include: shooting at unprotected pigs, destroying as many blocks as possible, and shooting at objects close to round objects [49].
- Pig Shooter: Pig Shooter has only one strategy: shooting at pigs. The agent randomly selects which pig to shoot and which trajectory to use [50].

All these heuristic agents work under the uninformed evaluation setting, in which the agent is not informed when the first novel task appears in the trial.

Learning Agents. In this work, we evaluate seven learning agents/versions of agents. All seven learning agents we present here work under the informed evaluation setting in which the agent is informed when the novelty appears in the trial. Therefore, we do not evaluate the novelty detection performance of these agents. The seven agents are *DQN Offline/ Online/ Adapt*, *Relational Offline/ Online/ Adapt*, and *Naive Adapt*. Same as the deep reinforcement learning agents used in [5], we train a DQN [51, 52] agent and the Relational agent that contains a relational module [53]. Both agents are trained on the tasks generated from a normal task template and are evaluated on the trials that contain tasks from the corresponding novel task template. We evaluate the DQN and Relational agents in three versions: offline, online, and adapt. With the offline version, the two deep reinforcement learning agents DQN and Relational, always select the action with the highest q-value throughout the trial. On the other hand, online learning agents update the q-network after novelty is introduced and try to relearn the policy to solve novel tasks. We also evaluate the recently developed open-world learning component *NAPPING* [54] together with DQN and Relational agents. We call these agents DQN Adapt and Relational Adapt.

The *Naive Adapt* is built on top of the *Pig Shooter* agent in [5], which shoots only at the pigs. *Naive Adapt* uses the strategy of the *Pig Shooter* in the pre-novelty game tasks. After the agent is informed that the novelty has occurred, it searches for a combination of (objects, trajectories, and delays) that solve a game level and keeps a record of each triplet tried. Once a solution triplet (e.g., a solution triplet could be (pig_2, high trajectory, delay 5 seconds)) is found for a trial, the *Naive Adapt* will keep using the triplet

until it does not solve the tasks anymore, where the agent starts to search for another triplet.

Random Agent. The Random Agent selects a random release point (x,y) relative to the slingshot. The x is sampled from [-200, 200] and y is sampled from [-200, 200]. This agent works under the uninformed evaluation setting.

5.2. Experimental Setups

5.2.1. Human Experiment Setup

The experiments with human participants were approved by the Australian National University committee on human ethics under the protocol 2021/293. Participation was entirely voluntary, and no monetary compensation was provided. There were 47 participants with ages ranging from 20 to 35 years and there were both males and females. They were not experienced Angry Birds players. Some of the participants have never played the game and some of them knew the general game mechanics and had played the game on an occasional basis in the past, but did not have an extensive understanding of the game's strategies. Participants provided their consent to use their play-data.

For a single participant, we provided 10 trials from 10 novelty-scenarios. We had four such trial-sets to cover all 40 novelty-scenarios. In a single trial, there were 1-4 normal tasks and 4 novel tasks. On average participants spent 25-30 minutes to complete the experiment. Participants attempted to solve the tasks and at the end of each task, they indicated if they detected a novelty or not.

5.2.2. Agent Experiment Setup

We use the standard SAIL-ON evaluation setup for all agents. As mentioned previously, we have eight novelties and five physical reasoning scenarios which results in 40 novelty-scenarios. For a single novelty-scenario, we test the agent on 40 trials. A trial consists of 1-40 normal tasks and 40 novel tasks. All heuristic agents and the Random agent do not require any training. However, learning agents are trained on the normal tasks of the corresponding novelty-scenario and then the agent is evaluated on the trial-set.

For the agents that we established the detection performance, we conducted an additional experiment with non-novel trials to understand the reliability of the detection. We used 40 non-novel trials with 40 non-novel tasks for this experiment.

6. Results and Analysis

In this section, we present and discuss the results of the experiments we conducted: the human player experiment and the baseline agent experiment. For both experiments, we report the novelty detection and novelty adaptation performance.

6.1. Human Performance

Figure 6 shows CDT and DD results of the human players. Overall, the participants were able to correctly detect the novelties in almost all the trials (CDT is close to 1). The lowest CDT is for the novelty-scenario actions-single force, in which the upward force of the Air Turbulence agent

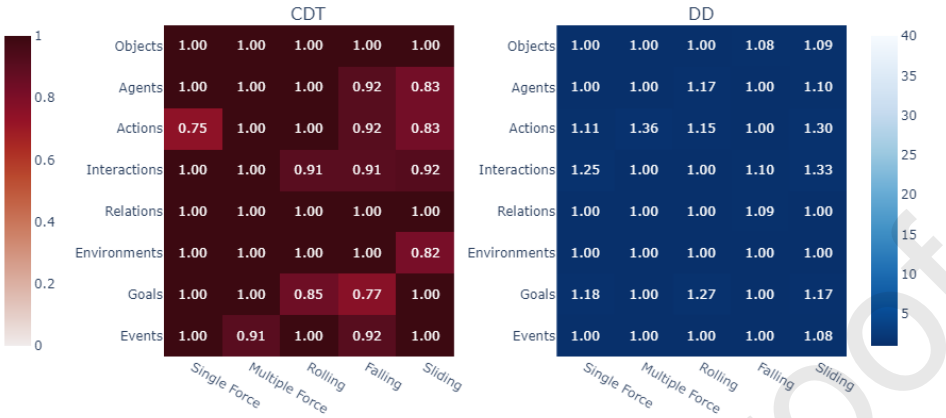


Figure 6: CDT (left) and DD (right) results of the human players. In the heat maps, the x-axis is the physical scenario and the y-axis is the novelty applied.

is increased. The reason for this is likely because this novelty is not visually detectable until interacted with. Also, when this novelty is applied to the single force scenario, the player has only to slightly adjust the shooting angle of the bird compared to the shooting angle in the normal tasks. As the impact of this novelty is subtle, it might not be perceivable to humans. This is also likely to be the case with the novelty-scenario that has the second lowest CDT: goals-falling. When we look at the DD results, in most cases it can be seen that the participants could detect the novelty in the first game level where the novelty was encountered (DD is close to 1). Generally, it can also be seen that for the novelties that are not visually detectable without interaction (actions, interactions, and goals), humans have a higher detection delay compared to the other novelties that can be visually detected before interaction.

The AP and AUS performances of the human players are shown in Figure 7. For the AP calculation, we used the asymptotic length as 2 (i.e., the

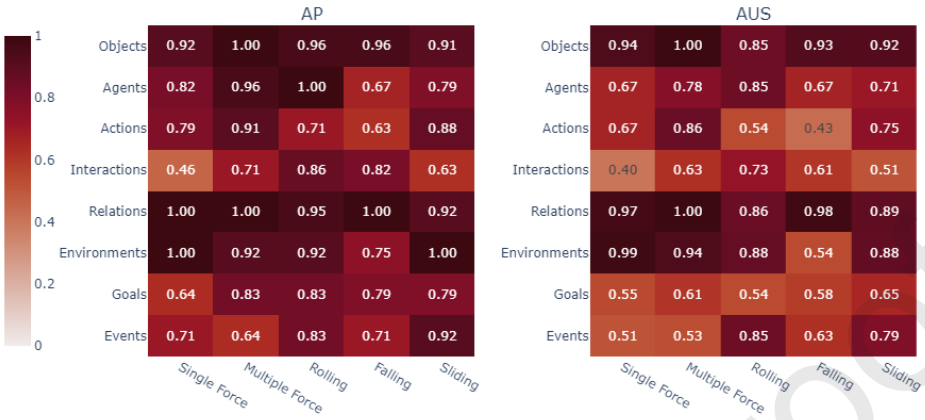


Figure 7: AP (left) and AUS (right) results of the human players. In the heat maps, the x-axis is the physical scenario and the y-axis is the novelty applied. The asymptotic length considered for the AP calculation is 2.

performance of the last two tasks in the trial). As the AP results depict, the participants obtained above 80% performance for most of the novelty-scenarios. Generally, for both the AP and AUS, participants showed lower results for the novelty-scenarios which required highly accurate actions. For example, interactions-single force, goals-single force, events-multiple forces, interactions-sliding, and actions-falling have slightly lower results for both the measures as it is required to shoot the bird with high accuracy to solve the tasks in those novelty-scenarios. The rate of adaptation is captured from the AUS results. When the accuracy requirements are higher, it takes more tasks for humans to adapt to the novelty. This is reflected in the relatively lower AUS results for the novelties that demand accurate actions: interactions, goals, and events. On the other hand, novelties that do not require accurate actions such as objects and relations have nearly perfect AUS results, as humans can adapt to such novelties straight away.

Moreover, we analyzed the relationship between the detection performance and adaptation performance of human players using the non-parametric Mann-Whitney test and Spearman's rank correlation (see Appendix F for more details). Considering the relationship between CDT and adaptation (both the AP and AUS), only the results of the actions-single force novelty-scenario shows that, the adaptation performance is independent of whether participants detected it or not. As discussed in the above paragraphs, this is because, even though humans have successfully adapted to the actions-single force, the novelty may not be perceivable to humans. The correlation between detection delay and adaptation performance (both the AP and AUS) shows that there are some novelty-scenarios such as goals-single force, goals-rolling, and interactions-sliding have moderately negative correlations implying that the longer a player takes to detect the novelty, the lower the adaptation performance. The correlation plots are shown in Appendix Figure F.44.

6.2. Baseline Agent Performance

In this section, we discuss the performance of the 11 baseline agents and humans in terms of novelty detection and novelty adaptation.

6.2.1. Baseline Agent Novelty Detection Performance

As discussed in Section 5.1 some of the agents in NovPhy work under the uninformed evaluation. We compute the detection performance measures for those agents. The detection modules in those baseline agents are based on the pass rate deviation (discussed in detail in Appendix D). The results of the detection measures per novelty are presented in Table 2 and detection

measures per scenario are presented in Table 3 for all the heuristic agents, the random agent, and humans.

As the results indicate, Datalab has the highest overall CDT while Pig Shooter has the lowest. It is expected for Pig Shooter to have the lowest CDT as the agent only directly shoots at the pigs, generally resulting in the same outcome in the tasks of a given trial, hence does not show a significant deviation in pass rates. Considering the overall DD, Random Agent has the highest DD. This is because the agent's randomness in the actions results in novelty detection at random positions of a trial causing the overall DD to fall around the mean number of the tasks in the trial. The Pig Shooter has the best (lowest) DD, this is because, for the few trials it correctly detects, it is done rapidly due to the deterministic nature of action selection. All the agents are far below the humans' novelty detection performance which is near perfect (CDT = 0.96 and DD = 1.07).

Considering the test with non-novel trials, the WDT% in non-novel trials for Random Agent is the lowest (0.35), for Datalab and Eagle's Wing, it is at 0.45, while for Pig Shooter it is at 0.75. We have conducted the KS test and Mann-Whitney test and the Tables 2 and 3 present the p-value obtained from the tests. To conduct the tests, we filtered out the trials where the total number of instances is between 40-55 as the non-novel trials contain only 40 instances. Considering a 5% level of significance, the p-values < 0.05 for the KS-test indicate that there is a significant difference between the detections in non-novel trials and other trials. Similarly, for the Mann-Whitney test, p-values < 0.05 indicate that there is a significant difference between the median instance the agent detects novelty in non-novel trials

Novelty	Measure	Human	DQN Offline	DQN Online	DQN Adapt	Rel Offline	Rel Online	Rel Adapt	Naive Adapt	Datalab	Eagle'sWing	PigShooter	Random
1. Objects	$NPS_{CDT,1}$	1.00 ± 0.00	-	-	-	-	-	-	-	0.65 ± 0.13	0.57 ± 0.08	0.25 ± 0.08	0.23 ± 0.0
	$NPS_{DD,1}$	1.03 ± 0.02	-	-	-	-	-	-	-	7.85 ± 4.04	5.46 ± 1.17	6.07 ± 2.53	12.79 ± 4.
	$NPS_{AP,1}$	0.95 ± 0.02	0.65 ± 0.17	0.66 ± 0.14	0.79 ± 0.10	0.54 ± 0.10	0.50 ± 0.10	0.80 ± 0.06	0.61 ± 0.19	0.37 ± 0.19	0.40 ± 0.21	0.02 ± 0.02	0.01 ± 0.0
	$NPS_{AUS,1}$	0.93 ± 0.02	0.66 ± 0.17	0.67 ± 0.16	0.81 ± 0.10	0.56 ± 0.11	0.55 ± 0.10	0.80 ± 0.06	0.61 ± 0.19	0.38 ± 0.19	0.41 ± 0.21	0.02 ± 0.02	0.01 ± 0.0
	$NPS_{KS,1}$	-	-	-	-	-	-	-	-	0.0008	0.0002	0.9997	0.9043
	$NPS_{MW,1}$	-	-	-	-	-	-	-	-	0.0004	0.0000	0.3190	0.4599
2. Agents	$NPS_{CDT,2}$	0.95 ± 0.03	-	-	-	-	-	-	-	0.49 ± 0.17	0.42 ± 0.12	0.17 ± 0.07	0.24 ± 0.0
	$NPS_{DD,2}$	1.05 ± 0.03	-	-	-	-	-	-	-	10.96 ± 4.29	10.01 ± 2.25	4.57 ± 0.39	19.59 ± 1.
	$NPS_{AP,2}$	0.85 ± 0.02	0.06 ± 0.03	0.21 ± 0.07	0.73 ± 0.13	0.07 ± 0.04	0.11 ± 0.04	0.67 ± 0.09	0.11 ± 0.06	0.05 ± 0.03	0.05 ± 0.03	0.01 ± 0.01	0.01 ± 0.0
	$NPS_{AUS,2}$	0.73 ± 0.03	0.06 ± 0.03	0.18 ± 0.06	0.62 ± 0.10	0.07 ± 0.04	0.10 ± 0.04	0.56 ± 0.09	0.09 ± 0.05	0.05 ± 0.03	0.05 ± 0.03	0.00 ± 0.00	0.01 ± 0.0
	$NPS_{KS,2}$	-	-	-	-	-	-	-	-	0.2047	0.4745	0.9999	0.3502
	$NPS_{MW,2}$	-	-	-	-	-	-	-	-	0.0395	0.0783	0.0748	0.0765
3. Actions	$NPS_{CDT,3}$	0.90 ± 0.05	-	-	-	-	-	-	-	0.44 ± 0.15	0.49 ± 0.19	0.24 ± 0.02	0.17 ± 0.0
	$NPS_{DD,3}$	1.19 ± 0.07	-	-	-	-	-	-	-	10.11 ± 1.69	8.12 ± 2.20	5.67 ± 0.57	19.27 ± 1.
	$NPS_{AP,3}$	0.78 ± 0.05	0.07 ± 0.06	0.13 ± 0.08	0.56 ± 0.17	0.05 ± 0.03	0.12 ± 0.07	0.63 ± 0.16	0.10 ± 0.07	0.03 ± 0.02	0.08 ± 0.04	0.00 ± 0.00	0.01 ± 0.0
	$NPS_{AUS,3}$	0.65 ± 0.08	0.08 ± 0.06	0.13 ± 0.08	0.53 ± 0.15	0.05 ± 0.03	0.10 ± 0.06	0.55 ± 0.14	0.08 ± 0.05	0.03 ± 0.02	0.08 ± 0.04	0.00 ± 0.00	0.01 ± 0.0
	$NPS_{KS,3}$	-	-	-	-	-	-	-	-	0.7698	0.0329	0.9998	0.7320
	$NPS_{MW,3}$	-	-	-	-	-	-	-	-	0.5983	0.0013	0.0211	0.3267
4. Interactions	$NPS_{CDT,4}$	0.95 ± 0.02	-	-	-	-	-	-	-	0.42 ± 0.15	0.46 ± 0.18	0.18 ± 0.05	0.55 ± 0.
	$NPS_{DD,4}$	1.14 ± 0.07	-	-	-	-	-	-	-	5.87 ± 1.32	16.31 ± 7.38	5.25 ± 0.35	18.40 ± 5.
	$NPS_{AP,4}$	0.69 ± 0.07	0.10 ± 0.04	0.24 ± 0.07	0.75 ± 0.16	0.08 ± 0.03	0.19 ± 0.05	0.72 ± 0.15	0.33 ± 0.14	0.10 ± 0.06	0.3 ± 0.17	0.03 ± 0.02	0.06 ± 0.0
	$NPS_{AUS,4}$	0.58 ± 0.05	0.10 ± 0.04	0.25 ± 0.08	0.64 ± 0.14	0.08 ± 0.03	0.19 ± 0.04	0.63 ± 0.13	0.27 ± 0.11	0.18 ± 0.07	0.35 ± 0.18	0.02 ± 0.02	0.06 ± 0.0
	$NPS_{KS,4}$	-	-	-	-	-	-	-	-	0.0717	0.0077	0.9998	0.5790
	$NPS_{MW,4}$	-	-	-	-	-	-	-	-	0.0761	0.0121	0.0610	0.2097
5. Relations	$NPS_{CDT,5}$	1.00 ± 0.00	-	-	-	-	-	-	-	0.27 ± 0.18	0.28 ± 0.17	0.08 ± 0.07	0.01 ± 0.0
	$NPS_{DD,5}$	1.02 ± 0.02	-	-	-	-	-	-	-	3.47 ± 0.40	7.49 ± 0.02	3.86 ± 0.54	4.00 ± 0.0
	$NPS_{AP,5}$	0.97 ± 0.02	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.0
	$NPS_{AUS,5}$	0.94 ± 0.03	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.0
	$NPS_{KS,5}$	-	-	-	-	-	-	-	-	0.0000	0.0358	0.9999	0.0863
	$NPS_{MW,5}$	-	-	-	-	-	-	-	-	0.0000	0.0104	0.2962	0.0809
6. Environments	$NPS_{CDT,6}$	0.96 ± 0.04	-	-	-	-	-	-	-	0.44 ± 0.15	0.35 ± 0.13	0.01 ± 0.01	0.40 ± 0.
	$NPS_{DD,6}$	1.00 ± 0.00	-	-	-	-	-	-	-	4.75 ± 0.19	5.50 ± 1.03	2.50 ± 0.95	14.64 ± 0.
	$NPS_{AP,6}$	0.92 ± 0.05	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.03 ± 0.0
	$NPS_{AUS,6}$	0.85 ± 0.08	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.03 ± 0.0
	$NPS_{KS,6}$	-	-	-	-	-	-	-	-	0.0295	0.0120	0.9999	0.4373
	$NPS_{MW,6}$	-	-	-	-	-	-	-	-	0.0035	0.0004	0.7050	0.2473
7. Goals	$NPS_{CDT,7}$	0.92 ± 0.05	-	-	-	-	-	-	-	0.43 ± 0.13	0.55 ± 0.10	0.19 ± 0.05	0.39 ± 0.0
	$NPS_{DD,7}$	1.12 ± 0.05	-	-	-	-	-	-	-	11.62 ± 1.48	12.28 ± 2.99	5.73 ± 0.22	15.06 ± 2.
	$NPS_{AP,7}$	0.78 ± 0.04	0.07 ± 0.02	0.09 ± 0.03	0.45 ± 0.15	0.04 ± 0.02	0.14 ± 0.08	0.54 ± 0.18	0.29 ± 0.12	0.08 ± 0.04	0.09 ± 0.03	0.01 ± 0.01	0.02 ± 0.0
	$NPS_{AUS,7}$	0.59 ± 0.02	0.08 ± 0.03	0.10 ± 0.03	0.42 ± 0.13	0.04 ± 0.02	0.12 ± 0.05	0.49 ± 0.16	0.29 ± 0.12	0.08 ± 0.04	0.09 ± 0.04	0.01 ± 0.01	0.02 ± 0.0
	$NPS_{KS,7}$	-	-	-	-	-	-	-	-	0.7128	0.3289	0.9998	0.1746
	$NPS_{MW,7}$	-	-	-	-	-	-	-	-	0.4532	0.2492	0.0409	0.1011
8. Events	$NPS_{CDT,8}$	0.97 ± 0.02	-	-	-	-	-	-	-	0.46 ± 0.13	0.46 ± 0.12	0.07 ± 0.03	0.26 ± 0.
	$NPS_{DD,8}$	1.02 ± 0.02	-	-	-	-	-	-	-	7.36 ± 1.73	11.00 ± 2.36	4.80 ± 1.35	18.14 ± 2.
	$NPS_{AP,8}$	0.76 ± 0.05	0.27 ± 0.12	0.21 ± 0.10	0.21 ± 0.1	0.29 ± 0.11	0.2 ± 0.09	0.2 ± 0.09	0.00 ± 0.00	0.13 ± 0.05	0.12 ± 0.04	0.01 ± 0.00	0.02 ± 0.0
	$NPS_{AUS,8}$	0.66 ± 0.07	0.27 ± 0.12	0.24 ± 0.11	0.23 ± 0.1	0.29 ± 0.11	0.22 ± 0.10	0.21 ± 0.09	0.00 ± 0.00	0.13 ± 0.05	0.12 ± 0.04	0.01 ± 0.00	0.02 ± 0.0
	$NPS_{KS,8}$	-	-	-	-	-	-	-	-	0.6271	0.8557	0.9999	0.7366
	$NPS_{MW,8}$	-	-	-	-	-	-	-	-	0.5114	0.9053	0.9999	0.3860

Table 2: Results of the NPS measures of the agents and humans for the eight novelties.

Scenario	Measure	Human	DQN Offline	DQN Online	DQN Adapt	Rel Offline	Rel Online	Rel Adapt	Naive Adapt	Datalab	Eagle'sWing	PigShooter	Random
1. Single Force	$SPN_{CDT,1}$	0.97 ± 0.03	-	-	-	-	-	-	-	0.59 ± 0.07	0.67 ± 0.05	0.08 ± 0.04	0.31 ± 0.
	$SPN_{DD,1}$	1.07 ± 0.04	-	-	-	-	-	-	-	6.61 ± 0.88	7.42 ± 1.46	4.82 ± 1.02	14.21 ± 2.
	$SPN_{AP,1}$	0.79 ± 0.07	0.11 ± 0.05	0.13 ± 0.05	0.37 ± 0.14	0.12 ± 0.06	0.13 ± 0.05	0.36 ± 0.13	0.13 ± 0.11	0.08 ± 0.04	0.16 ± 0.11	0.01 ± 0.00	0.03 ± 0.
	$SPN_{AUS,1}$	0.71 ± 0.08	0.11 ± 0.05	0.13 ± 0.05	0.33 ± 0.12	0.12 ± 0.06	0.13 ± 0.06	0.33 ± 0.11	0.13 ± 0.12	0.08 ± 0.05	0.17 ± 0.12	0.01 ± 0.00	0.03 ± 0.
	$SPN_{KS,1}$	-	-	-	-	-	-	-	-	0.1168	0.0440	0.9999	0.8943
	$SPN_{MW,1}$	-	-	-	-	-	-	-	-	0.0210	0.0024	0.1349	0.6008
2. Multiple Force	$SPN_{CDT,2}$	0.99 ± 0.01	-	-	-	-	-	-	-	0.55 ± 0.14	0.55 ± 0.10	0.14 ± 0.03	0.14 ± 0.
	$SPN_{DD,2}$	1.05 ± 0.05	-	-	-	-	-	-	-	4.29 ± 1.28	4.88 ± 0.47	5.16 ± 0.38	17.48 ± 1.
	$SPN_{AP,2}$	0.87 ± 0.05	0.07 ± 0.03	0.15 ± 0.06	0.11 ± 0.05	0.05 ± 0.03	0.06 ± 0.03	0.16 ± 0.08	0.17 ± 0.12	0.01 ± 0	0.08 ± 0.05	0.00 ± 0.00	0.01 ± 0.
	$SPN_{AUS,2}$	0.79 ± 0.07	0.08 ± 0.03	0.14 ± 0.06	0.12 ± 0.06	0.04 ± 0.02	0.07 ± 0.03	0.15 ± 0.07	0.17 ± 0.12	0.05 ± 0.05	0.10 ± 0.07	0.00 ± 0.00	0.01 ± 0.
	$SPN_{KS,2}$	-	-	-	-	-	-	-	-	0.0065	0.0050	0.9998	0.0908
	$SPN_{MW,2}$	-	-	-	-	-	-	-	-	0.0013	0.0001	0.0203	0.0160
3. Rolling	$SPN_{CDT,3}$	0.97 ± 0.02	-	-	-	-	-	-	-	0.26 ± 0.07	0.24 ± 0.08	0.20 ± 0.04	0.34 ± 0.
	$SPN_{DD,3}$	1.07 ± 0.04	-	-	-	-	-	-	-	10.49 ± 3.11	12.02 ± 2.28	4.45 ± 0.42	16.13 ± 1.
	$SPN_{AP,3}$	0.88 ± 0.03	0.24 ± 0.13	0.27 ± 0.12	0.59 ± 0.14	0.24 ± 0.11	0.22 ± 0.10	0.60 ± 0.14	0.18 ± 0.10	0.07 ± 0.03	0.07 ± 0.03	0.00 ± 0.00	0.02 ± 0.
	$SPN_{AUS,3}$	0.76 ± 0.05	0.25 ± 0.13	0.28 ± 0.12	0.57 ± 0.13	0.25 ± 0.11	0.23 ± 0.11	0.57 ± 0.13	0.15 ± 0.08	0.07 ± 0.02	0.07 ± 0.03	0.00 ± 0.00	0.02 ± 0.
	$SPN_{KS,3}$	-	-	-	-	-	-	-	-	0.1732	0.3269	0.9998	0.4444
	$SPN_{MW,3}$	-	-	-	-	-	-	-	-	0.1116	0.1336	0.1068	0.1322
4. Falling	$SPN_{CDT,4}$	0.93 ± 0.03	-	-	-	-	-	-	-	0.48 ± 0.11	0.44 ± 0.11	0.19 ± 0.05	0.36 ± 0.
	$SPN_{DD,4}$	1.03 ± 0.02	-	-	-	-	-	-	-	10.04 ± 1.75	12.51 ± 2.37	4.89 ± 0.48	13.35 ± 1.
	$SPN_{AP,4}$	0.79 ± 0.05	0.17 ± 0.11	0.20 ± 0.1	0.57 ± 0.14	0.14 ± 0.09	0.2 ± 0.08	0.55 ± 0.14	0.21 ± 0.09	0.19 ± 0.10	0.19 ± 0.10	0.02 ± 0.01	0.02 ± 0.
	$SPN_{AUS,4}$	0.67 ± 0.07	0.18 ± 0.11	0.22 ± 0.10	0.51 ± 0.13	0.14 ± 0.09	0.19 ± 0.08	0.50 ± 0.13	0.19 ± 0.09	0.20 ± 0.10	0.19 ± 0.10	0.02 ± 0.01	0.02 ± 0.
	$SPN_{KS,4}$	-	-	-	-	-	-	-	-	0.2704	0.3204	0.9997	0.8495
	$SPN_{MW,4}$	-	-	-	-	-	-	-	-	0.1202	0.2305	0.0150	0.3387
5. Sliding	$SPN_{CDT,5}$	0.93 ± 0.03	-	-	-	-	-	-	-	0.37 ± 0.13	0.33 ± 0.13	0.12 ± 0.06	0.25 ± 0.
	$SPN_{DD,5}$	1.13 ± 0.04	-	-	-	-	-	-	-	9.10 ± 1.72	12.97 ± 4.41	6.99 ± 2.25	22.12 ± 2.
	$SPN_{AP,5}$	0.85 ± 0.04	0.17 ± 0.11	0.21 ± 0.12	0.55 ± 0.15	0.13 ± 0.08	0.18 ± 0.07	0.55 ± 0.14	0.21 ± 0.10	0.13 ± 0.10	0.16 ± 0.11	0.02 ± 0.01	0.01 ± 0.
	$SPN_{AUS,5}$	0.76 ± 0.05	0.18 ± 0.12	0.21 ± 0.12	0.49 ± 0.13	0.13 ± 0.08	0.18 ± 0.07	0.48 ± 0.12	0.19 ± 0.10	0.14 ± 0.10	0.16 ± 0.12	0.01 ± 0.01	0.01 ± 0.
	$SPN_{KS,5}$	-	-	-	-	-	-	-	-	0.1795	0.0218	0.9764	0.3878
	$SPN_{MW,5}$	-	-	-	-	-	-	-	-	0.2176	0.0170	0.1847	0.1255

Table 3: Results of the SPN measures of the agents and humans for the five scenarios.

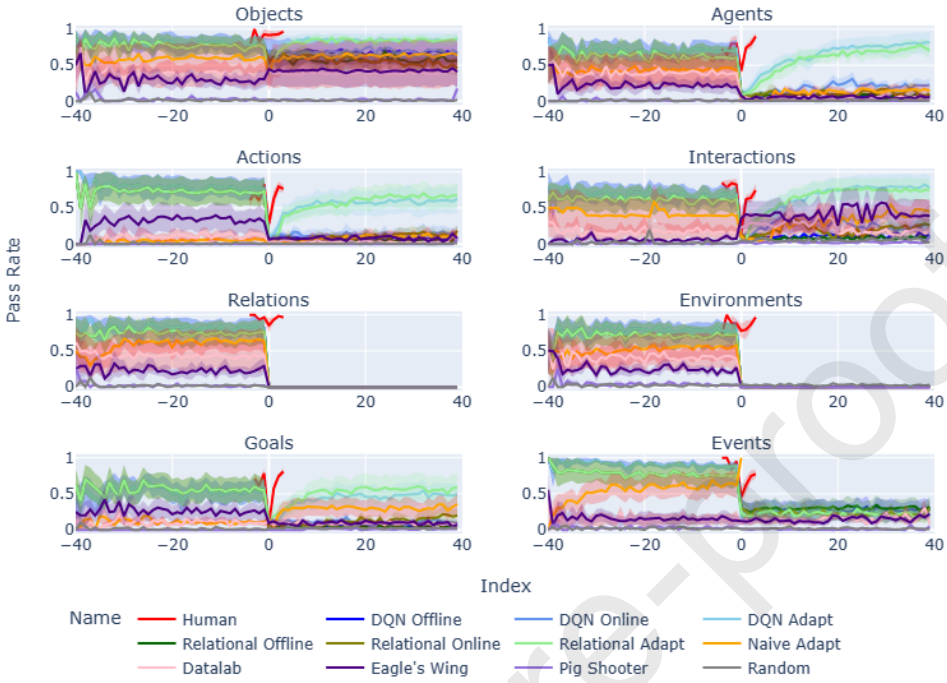


Figure 8: The pass rate of the agents per novelty. The x-axis represents the index of the task in trials. Indexes -40 to -1 represent normal tasks and 0 to 40 represent novel tasks. The y-axis shows the pass rate averaged across all trials relevant to the respective novelty.

and other trials. Notably, there's no significant difference in distributions (and medians) between Pig Shooter and the Random Agent across most novelties and scenarios. This result is expected since Pig Shooter is designed to target pigs, while the Random Agent randomly shoots regardless of the trial they operate and thus makes their detections unreliable.

6.2.2. Baseline Agent Novelty Adaptation Performance

The adaptation plots per novelty are shown in Figure 8 and the adaptation plots per scenario are shown in Figure 9. The novelty adaptation measures derived from the adaptation curves, AP and AUS results per sce-

Figure 9: The pass rate of the agents per scenario. x-axis represents the index of the task in trials. Indexes -40 to -1 represent normal tasks and 0 to 40 represent novel tasks. The y-axis shows the pass rate averaged across all trials relevant to the respective scenario.

nario are presented in Table 2 and per novelty results are presented in Table 3. The AP results are based on the last 50% of the novel tasks (m=20). Ideally, an agent would have a higher pass rate in normal tasks and when novel tasks begin (at index 0 in Figures 8 and 9), the performance would drop and recover its performance to reach the normal task performance within a few tasks. Therefore, an ideal agent would have AP and AUS results close to 1.

In Figure 9, the agents DQN Adapt and Relational Adapt show a better adaptation behaviour compared to other agents in four scenarios out of five, the only exception is multiple forces. In those four scenarios, out of the two agents, DQN Adapt shows a slightly better performance (at 10% level of significance) than Relational Adapt when we look at the $SPN_{AP,j}$, and $SPN_{AUS,j}$. Similarly as shown in Figure 8, the two agents adapt in five novelties except the novelties Relations, Environments, and Events. This

is because to adapt to the Relations and Environments novelties, the agent needs to adjust the pre-defined trajectory planner. For example, in the Environments novelty, the trajectory is upside down. However, the -Adapt agents currently use the provided trajectory planner. Similarly, the Naive Adapt agent also shows the adaptation behaviour in all scenarios except for single force. This agent shows an adaptation behaviour in all the novelties except Relations, Environments, and Events. However, as the Naive Adapt agent searches for just one solution tuple through a trial, it can not adapt to novelty trials that require different solution tuples under different situations. As a result, the Naive Adapt agent does not reach the performance level of the two agents, DQN Adapt and Relational Adapt, who learn efficiently how to handle novelties in different scenarios through each trial. Overall, the agents show the best adaptation performance in the Objects novelty. This is because the Objects novelty tested here is a change of the colour of an existing object, which has not impacted the agents' actions drastically. It is interesting to note that none of the agents has reached the humans' pass rate in any scenario or in any novelty, except for interactions novelty where Relational Adapt and DQN Adapt exceed the humans' $NPS_{AP,j}$, and $NPS_{AUS,j}$. However, within the same number of tasks that were given to humans, those two agents have not reached the performance the humans could achieve, showing that there is room for improvement in terms of adaptation efficiency.

7. Conclusion and Future Work

The objective of NovPhy is to facilitate the development of AI systems that can perform physical reasoning tasks in the presence of novelties, which is the condition that a system in an open-world physical environment would encounter. Towards this objective, NovPhy was designed to evaluate the abilities of an agent to detect novelties and adapt to perform under those novelties in a physical environment. We designed task templates for five commonly encountered real-world physical scenarios. Then, we designed novel tasks by introducing a diverse set of novelties to those task templates. This design enables to measure novelty detection and adaptation of agents in two directions: 1) how an agent performs in a novelty when the novelty is applied to different physical scenarios, and 2) how an agent performs in a physical scenario when different novelties are applied to it. To measure the true novelty adaptation performance of the agents, when designing the tasks we ensure that the agent has to work under the influence of the novelty rather than bypassing the novelties to solve the tasks. We evaluated the agents using a trial setting, in which the agent has to play a sequence of tasks of a scenario without novelties followed by a sequence of tasks of that scenario with novelties.

We have established the baseline results of the benchmark using human players, learning agents, and heuristic agents. The results show that novelties affect the agents' performance severely and some agents can recover as they play more and more tasks. However, agents' solving rate and efficiency in adaptation are subpar compared to humans' performance. Although our

results show that DQN Adapt and Relational Adapt agents are able to adapt to most novelties, there are still some novelties that the agents fail to adapt to. The main reason is that the agents still use the provided trajectory planner to plan for the shot (the release point). Future work on agents may focus on easing the need of using a trajectory planner to allow the agent to adapt to a wider range of novelties.

We foresee different directions of improvement for NovPhy. NovPhy can be advanced by introducing more novelties representing the levels of the novelty hierarchy. Further, more complex physical reasoning scenarios such as relative height, relative weight, and clearing paths can be introduced to the benchmark after agents show efficient novelty detection and novelty adaptation in the existing scenarios. Moreover, the benchmark can be extended to assess the novelty characterization ability of agents (i.e., to evaluate whether an agent correctly detects 'what is novel' in a task). Additionally, the concept in NovPhy, evaluating the physical reasoning capabilities under the influence of novelties, can be extended to other physical reasoning domains. For example, novelties can be introduced to physics-based robotic benchmarks such as CausalWorld [55] and RLBench [56], which will facilitate evaluating agents on physical scenarios such as pulling, picking and placing, and stacking, which are not seen in the Angry Birds domain. We believe that NovPhy builds the foundation for future research on developing agents that can efficiently detect and adapt to novelty in the physical world as humans do.

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

[1] E. Davis, Physical reasoning (2006).
URL `https://cs.nyu.edu/~davise/papers/handbookKR.pdf`

[2] R. Baillargeon, Physical reasoning in infancy, The cognitive neurosciences (1995) 181–204.

[3] R. Baillargeon, J. Li, W. Ng, S. Yuan, An account of infants' physical reasoning, Learning and the infant mind 66 (2009) 116.

[4] F. Chollet, On the measure of intelligence (2019). arXiv:1911.01547.

[5] C. Xue*, V. Pinto*, C. Gamage*, E. Nikonova, P. Zhang, J. Renz, Phy-q as a measure for physical reasoning intelligence., Nature Machine Intelligence 5 (2023) 83–93, *equal contribution.

[6] K. R. Allen, K. A. Smith, J. B. Tenenbaum, Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning, Proceedings of the National Academy of Sciences 117 (47) (2020) 29302–29310. doi:10.1073/pnas.1912341117.
URL https://www.pnas.org/content/117/47/29302.full.pdf

[7] D. Bear, E. Wang, D. Mrowca, F. J. Binder, H.-Y. Tung, R. Pramod, C. Holdaway, S. Tao, K. A. Smith, F.-Y. Sun, et al., Physion: Evaluating physical prediction from vision in humans and machines, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1), 2021.

[8] Z. Zeng, E. Davis, Physical reasoning in an open world, Advances in Cognitive Systems (ACS) Conference (2022).

[9] Wikimedia Foundation, Self-checkout, [Accessed: January. 04, 2023] [cited 04.01.2022].
URL https://en.wikipedia.org/wiki/Self-checkout

[10] Wikimedia Foundation, Multiple-vehicle collision, [Accessed: January. 04, 2023] [cited 04.01.2022].
URL https://en.wikipedia.org/wiki/Multiple-vehicle_collision

[11] B. Goertzel, C. Pennachin, Artificial general intelligence, Vol. 2, Springer, 2007.

[12] T. Senator, Science of artificial intelligence and learning for open-world novelty (SAIL-ON) (2019) [cited 10.11.2022].

URL https://www.darpa.mil/program/science-of-artificial-intelligence-and-learning-for-open-world-novelty

[13] P. Langley, Open-world learning for radically autonomous agents, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 13539–13543.

[14] M. Kejriwal, S. Thomas, A multi-agent simulator for generating novelty in monopoly, Simulation Modelling Practice and Theory 112 (2021) 102364. doi:https://doi.org/10.1016/j.simpat.2021.102364.
URL https://www.sciencedirect.com/science/article/pii/S1569190X21000770

[15] P. Feeney, S. Schneider, P. Lymperopoulos, L. Liu, M. Scheutz, M. C. Hughes, NovelCraft: A dataset for novelty detection and discovery in open worlds, arXiv preprint arXiv:2206.11736 (2022).

[16] CartPole, CartPole 3D domain (2022).
URL https://github.com/holderlb/WSU-SAILON-NG/tree/master/domains/cartpole

[17] J. Renz, X. Ge, S. Gould, P. Zhang, The angry birds AI competition, AI Magazine 36 (2015) 85–87. doi:10.1609/aimag.v36i2.2588.

[18] J. Renz, X. Ge, M. Stephenson, P. Zhang, AI meets angry birds, Nature Machine Intelligence 1 (7) (2019) 328–328.

[19] C. Xue, V. Pinto, P. Zhang, C. Gamage, E. Nikonova, J. Renz, Science birds novelty: An open-world learning test-bed for physics domains,

Proceedings of the AAAI Conference on Artificial Intelligence, Designing Artificial Intelligence for Open Worlds (2022).

[20] C. Gamage, V. Pinto, C. Xue, M. Stephenson, P. Zhang, J. Renz, Novelty Generation Framework for AI Agents in Angry Birds Style Physics Games, in: 2021 IEEE Conference of Games, COG 2021, 2021.

[21] V. Pinto, C. Xue, C. N. Gamage, J. Renz, The difficulty of novelty detection in open-world physical domains: An application to angry birds, arXiv preprint arXiv:2106.08670 (2021).

[22] R. Riochet, M. Y. Castro, M. Bernard, A. Lerer, R. Fergus, V. Izard, E. Dupoux, Intphys 2019: A benchmark for visual intuitive physics understanding, ArXiv abs/1803.07616 (2020).

[23] K. Yi*, C. Gan*, Y. Li, P. Kohli, J. Wu, A. Torralba, J. B. Tenenbaum, Clevrer: Collision events for video representation and reasoning, in: International Conference on Learning Representations, 2020.
URL https://openreview.net/forum?id=HkxYzANYDB

[24] F. Baradel, N. Neverova, J. Mille, G. Mori, C. Wolf, Cophy: Counterfactual learning of physical dynamics, in: ICLR, 2020. doi:10.21227/ps5q-8m55.
URL https://dx.doi.org/10.21227/ps5q-8m55

[25] A. Bakhtin, L. van der Maaten, J. Johnson, L. Gustafson, R. Girshick, Phyre: A new benchmark for physical reasoning, in: NeurIPS, 2019.

[26] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang,

A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al., Mastering the game of go without human knowledge, nature 550 (7676) (2017) 354–359.

[27] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al., A general reinforcement learning algorithm that masters chess, shogi, and go through self-play, Science 362 (6419) (2018) 1140–1144.

[28] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al., Grandmaster level in starcraft ii using multi-agent reinforcement learning, Nature 575 (7782) (2019) 350–354.

[29] SAIL-ON-BBA, Broad agency announcement, science of artificial intelligence and learning for open-world novelty (sail-on), [Accessed: Jan. 17, 2023] (2019).
URL `https://sam.gov/opp/88fdca99de93ddbb74cd8fb51916ceaa/view`

[30] T. Boult, P. A. Grabowicz, D. Prijatelj, R. Stern, L. Holder, J. Alspector, M. Jafarzadeh, T. Ahmad, A. R. Dhamija, Cli, S. Cruz, A. Shrivastava, C. Vondrick, W. Scheirer, Towards a unifying framework for formal theories of novelty, Proceedings of the AAAI Conference on Artificial Intelligence 35 (2021) 15047–15052.

[31] K. Doctor, C. Task, E. Kildebeck, M. Kejriwal, L. Holder, R. Leong, Toward defining a domain complexity measure across domains, Pro-

ceedings of the AAAI Conference on Artificial Intelligence, Designing Artificial Intelligence for Open Worlds (2022).

[32] M. Molineaux, D. Dannenhauer, An environment transformation-based framework for comparison of open-world learning agents, Proceedings of the AAAI Conference on Artificial Intelligence, Designing Artificial Intelligence for Open Worlds (2022).

[33] J. Balloch, Z. Lin, M. Hussain, A. Srinivas, R. Wright, X. Peng, J. Kim, M. Riedl, Novgrid: A flexible grid world for evaluating agent response to novelty, arXiv preprint arXiv:2203.12117 (2022).

[34] S. Goel, G. Tatiya, M. Scheutz, J. Sinapov, NovelGridworlds: A benchmark environment for detecting and adapting to novelties in open worlds, International Foundation for Autonomous Agents and Multiagent Systems, AAMAS (2021).

[35] M. Chevalier-Boisvert, L. Willems, S. Pal, Minimalistic gridworld environment for gymnasium (2018).
URL `https://github.com/Farama-Foundation/Minigrid`

[36] Minecraft, Minecraft official game (2022).
URL `https://www.minecraft.net/en-us`

[37] R. A. Smaldone, C. M. Thompson, M. Evans, W. Voit, Teaching science through video games, Nature chemistry 9 (2) (2017) 97–102.

[38] Rovio Enternainment, Angry birds game, [Accessed: December. 07,

2022] [cited 07.12.2022].

URL https://www.rovio.com/games/angry-birds

[39] V. Pinto, J. Renz, C. Xue, P. Zhang, K. Doctor, D. W. Aha, Measuring the performance of open-world AI systems, Proceedings of the AAAI Conference on Artificial Intelligence, Designing Artificial Intelligence for Open Worlds (2022).

[40] M. Jafarzadeh, A. R. Dhamija, S. Cruz, C. Li, T. Ahmad, T. E. Boult, A review of open-world learning and steps toward open-world learning without labels, arXiv e-prints (2020) arXiv–2011.

[41] X. Peng, J. C. Balloch, M. O. Riedl, Detecting and adapting to novelty in games, arXiv preprint arXiv:2106.02204 (2021).

[42] F. Muhammad, V. Sarathy, G. Tatiya, S. Goel, S. Gyawali, M. Guaman, J. Sinapov, M. Scheutz, A novelty-centric agent architecture for changing worlds, in: Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, 2021, pp. 925–933.

[43] L. Ferreira, C. Toledo, A search-based approach for generating angry birds levels, in: 2014 IEEE Conference on Computational Intelligence and Games, 2014, pp. 1–8.

[44] A. Sanborn, V. Mansinghka, T. Griffiths, Reconciling intuitive physics and newtonian mechanics for colliding objects, Psychological review 120 (03 2013). doi:10.1037/a0031912.

[45] J. Bliss, J. Ogborn, Force and motion from the beginning, Learning

and Instruction 4 (1) (1994) 7–25. doi:https://doi.org/10.1016/0959-4752(94)90016-7.
URL https://www.sciencedirect.com/science/article/pii/0959475294900167

[46] G. Z, A brief survey of nonparametric statistics, Communications in Statistics - Theory and Methods 5 (5) (1976) 429–453. arXiv:https://doi.org/10.1080/03610927608827365, doi:10.1080/03610927608827365.
URL https://doi.org/10.1080/03610927608827365

[47] AIBIRDS, Angry birds AI competition [cited 22.12.2022].
URL http://aibirds.org/

[48] T. Borovička, R. Špetlík, K. Rymeš, Datalab angry birds ai [cited 22.12.2022].
URL http://aibirds.org/2014-papers/datalab-birds.pdf

[49] T. J. Wang, Ai angry birds eagle wing [cited 22.12.2022].
URL https://github.com/heartyguy/AI-AngryBird-Eagle-Wing

[50] M. Stephenson, J. Renz, X. Ge, P. Zhang, The 2017 AIBIRDS competition, ArXiv abs/1803.05156 (2018).

[51] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, N. Freitas, Dueling network architectures for deep reinforcement learning, in: International conference on machine learning, PMLR, 2016, pp. 1995–2003.

[52] H. Van Hasselt, A. Guez, D. Silver, Deep reinforcement learning with

double q-learning, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 30, 2016.

[53] V. Zambaldi, D. Raposo, A. Santoro, V. Bapst, Y. Li, I. Babuschkin, K. Tuyls, D. Reichert, T. Lillicrap, E. Lockhart, et al., Relational deep reinforcement learning, arXiv preprint arXiv:1806.01830 (2018).

[54] C. Xue, E. Nikonova, P. Zhang, J. Renz, Rapid open-world adaptation by adaptation principles learning, Submitted to Artificial Intelligence Journal (2023).

[55] O. Ahmed*, F. Träuble*, A. Goyal, A. Neitz, Y. Bengio, B. Schölkopf, M. Wüthrich, S. Bauer, Causalworld: A robotic manipulation benchmark for causal structure and transfer learning, in: 9th International Conference on Learning Representations (ICLR), 2021, *equal contribution.
URL https://openreview.net/pdf?id=SK7A5pdrgov

[56] S. James, Z. Ma, D. R. Arrojo, A. J. Davison, Rlbench: The robot learning benchmark & learning environment, IEEE Robotics and Automation Letters 5 (2) (2020) 3019–3026.

[57] D. J. Musliner, M. J. Pelican, M. McLure, S. Johnston, R. G. Freedman, C. Knutson, Openmind: Planning and adapting in domains with novelty, Proceedings of the Ninth Annual Conference on Advances in Cognitive Systems (2021).

[58] AIBirds-NoveltyTrack, Angry birds AI competition, novelty track,

http://aibirds.org/angry-birds-ai-competition/novelty-track.html, [Accessed: Jan. 19, 2023] (2021).

[59] M. Klenk, W. Piotrowski, R. Stern, S. Mohan, J. de Kleer, Model-based novelty adaptation for open-world AI, in: International Workshop on Principles of Diagnosis (DX), 2020.

[60] D. D. Jensen, Improving causal inference by increasing model expressiveness, Proceedings of the AAAI Conference on Artificial Intelligence 35 (17) (2021) 15053–15057. doi:10.1609/aaai.v35i17.17767.
URL https://ojs.aaai.org/index.php/AAAI/article/view/17767

[61] C. R. Starke, T. Miemietz, T. R. Hennig, L. Thies, L. Schweizer, SimbaDD: Simulation based agent dresden (2019).
URL http://www.aibirds.org/2019/SimbaDD.pdf