Hi-Phy: A Benchmark for Hierarchical Physical Reasoning

Cheng Xue∗  Vimukthini Pinto∗  Chathura Gamage∗
Peng Zhang  Jochen Renz
School of Computing
The Australian National University
Canberra, Australia
{cheng.xue, vimukthini.inguruwattage, chathura.gamage}@anu.edu.au
{p.zhang, jochen.renz}@anu.edu.au

Abstract

Reasoning about the behaviour of physical objects is a key capability of agents operating in physical worlds. Humans are very experienced in physical reasoning while it remains a major challenge for AI. To facilitate research addressing this problem, several benchmarks have been proposed recently. However, these benchmarks do not enable us to measure an agent’s granular physical reasoning capabilities when solving a complex reasoning task. In this paper, we propose a new benchmark for physical reasoning that allows us to test individual physical reasoning capabilities. Inspired by how humans acquire these capabilities, we propose a general hierarchy of physical reasoning capabilities with increasing complexity. Our benchmark tests capabilities according to this hierarchy through generated physical reasoning tasks in the video game Angry Birds. This benchmark enables us to conduct a comprehensive agent evaluation by measuring the agent’s granular physical reasoning capabilities. We conduct an evaluation with human players, learning agents, and heuristic agents and determine their capabilities. Our evaluation shows that learning agents, with good local generalization ability, still struggle to learn the underlying physical reasoning capabilities and perform worse than current state-of-the-art heuristic agents and humans. We believe that this benchmark will encourage researchers to develop intelligent agents with advanced, human-like physical reasoning capabilities.

1 Introduction

The ability to reason about object behaviours in physical environments lies at the core of human cognitive development [1][2]. Humans, also non-human animals such as apes [3][4] and crows [5][6] have shown abilities of understanding and reasoning abstract physical concepts including gravity, mass, and friction [7]. Few days after birth, infants can understand object solidity [8] and within the first year of birth, they understand notions such as object permanence [9], spatiotemporal continuity [10], stability [11], support [12], causality [13], and shape constancy [14]. However, it still remains a major challenge to develop AI systems that have the above physical reasoning capabilities to successfully operate in the physical environment.

As to facilitate research in developing AI systems with physical reasoning capabilities, environments such as PHYRE [7], Virtual Tools [15], OGRE [16], CLEVRER [1], and IntPhys 2019 [17] were introduced recently. Despite the recent advancement in physical reasoning benchmarks, there is a

∗First three authors have equal contribution
lack of a benchmark to evaluate agent’s performance in granular physical reasoning capabilities (i.e.,
to understand if agents can learn to perform a specific type of physical reasoning tasks such as rolling,
falling, etc.). Understanding the agent’s strengths in learning granular physical reasoning capabilities
is important to recognize the tasks that the agent can perform well. On the other hand, identifying
agent’s weaknesses in granular capabilities enables to specifically focus and improve on those weak
capabilities.

To accommodate limitations in current physical reasoning benchmarks, we propose 1) a general
hierarchy of physical reasoning capabilities with increasing complexity, based on the hierarchy, 2) a
benchmark that evaluates the physical reasoning capabilities in our hierarchy using generated tasks
in the popular physics puzzle game Angry Birds. We can summarize the main contributions of the
paper as follows.

- **A hierarchy for physical reasoning**: We propose a general hierarchy of physical reasoning
capabilities. Agents require the capabilities gained in the lower levels of the hierarchy to
successfully perform in the tasks that require higher level capabilities. The hierarchy enables
a comprehensive evaluation of an agents’ physical reasoning capabilities.

- **Hi-Phy, a Hierarchical PHYSical reasoning benchmark**: According to the general hier-
archy, we design task templates for the Angry Birds game to measure the granular physical
reasoning capabilities of the agents. To facilitate agent development, we developed an Angry
Birds framework that allows training multi agent instances at the same time with accelerated
game-play up to 50 times of the normal game speed. We also created task generators that
are able to generate game tasks that require particular reasoning capability to solve.

- **A comprehensive evaluation of baseline agents**: We present a quantifiable comparison
between humans and AI agents in our benchmark. The evaluation consists of five baseline
agents: our best performing learning agent, three Angry Birds heuristic agents, and a random
agent.

## 2 Background and Related Work

### 2.1 Related Physics Benchmarks and Competitions

We conduct a comparison between nine other physical reasoning benchmarks and three physics
based AI games competitions with respect to six criteria. 1) Evaluating individual physical reasoning
capabilities of an agent, i.e., we check if the benchmark has tasks that are designed to evaluate
if agents can perform on a specific physical reasoning capability. 2) Hierarchical categorization
of tasks according to the complexity of capabilities required to accomplish them, i.e., we check
if the benchmark tests agents on tasks that are designed with different levels of complexity of the
capabilities needed to achieve the tasks. 3) Procedural generation of tasks or variations of the tasks,
i.e., the tasks/variants of the tasks in the benchmark are created algorithmically facilitating the users to
generate any amount of data. 4) Destructibility of objects in the environment, i.e., if the environment
contains objects that can be destroyed upon the application of forces. Destructible objects make
the environment more realistic than an environment that has indestructible objects since the agents
need to consider the magnitude of the force that is applied to the objects. 5) Observing the outcome
of a desired physical action, i.e., if agents can physically interact and observe the outcome of the
action the agent takes. 6) Inclusion of human player data, i.e., if the benchmark or the competition is
supported by the results of human players.

The benchmarks we consider are PHYRE [7], Virtual Tools game [15], and OGRE [16], which are
game based benchmarks, IntPhys [17], CLEVERER [1], and CATER [18], which are video based
benchmarks, COPHY [19] which is an image based benchmark, CausalWorld [20] and RLBen
[21] which are robotic benchmarks. The AI game competitions we consider are Computational Pool
[22], Geometry Friends [23], and AIBirds [24]. Table 1 summarises the comparison.

The most related physical reasoning benchmark to ours is PHYRE [7], which consists of tasks to
measure two levels of generalization of agents. The benchmark tests if agents can generalize to
solve tasks within a task template (i.e., a set of tasks with different locations of the objects in the
task template) and if agents can generalize between different task templates. In PHYRE, a task may
require one or more physical reasoning capabilities to solve the task. Therefore, PHYRE does not
enable to measure individual physical reasoning capabilities of the agents. However, the capabilities
Table 1: Comparison of Hi-Phy with related physics benchmarks and competitions

<table>
<thead>
<tr>
<th>Benchmark/Competition</th>
<th>individual capability evaluation</th>
<th>hierarchical categorization of capabilities</th>
<th>procedurally generated tasks/variants</th>
<th>destructible objects</th>
<th>observe desired outcomes of a physical action</th>
<th>human player data</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHYRE [7]</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Virtual Tools [15]</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>OGRE [16]</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>IntPhys 2019 [17]</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>CLEVERER [1]</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>CATER [18]</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>COPHY [19]</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>CausalWorld [20]</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>RLBench [21]</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Computational Pool [22]</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Geometry Friends [23]</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>AIBirds [24]</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Hi-Phy (ours)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

needed to solve a task of our benchmark can be decomposed to lower level capabilities of the proposed hierarchy, which enables us to perform evaluating an agent’s ability in learning individual capability. Moreover, the tasks in PHYRE have multiple actions that may require different physical reasoning capabilities to achieve the goal. Whereas our benchmark is designed with tasks that can only be solved by understanding a specific physical reasoning capability which supports the above mentioned individual physical reasoning capability evaluation. Additionally, the cross template evaluation in PHYRE does not guarantee that the capabilities in the testing templates would exist in the training templates, whereas our within capability evaluation always ensures the agent is tested and is trained on the specified capability. Furthermore, compared to PHYRE, the physical environment in Hi-Phy is more complex in terms of having more object shapes, destructible objects, and objects with different densities, bounciness, and friction. AIBirds [24] competition is also closely related to our benchmark as the competition is based on Angry Birds. However, similar to PHYRE, the tasks used in the competition are not specifically designed to measure individual physical reasoning capabilities. We discuss more details about the AIBirds competition in Section 2.2.

Hi-Phy improves upon previous benchmarks by categorizing the tasks in the benchmark according to a hierarchy of physical reasoning capabilities needed to solve them. Through careful designing of task templates, we ensure that we evaluate agents on individual physical reasoning capabilities. This allows a comprehensive analysis of strengths and weaknesses of an agent’s individual physical reasoning capabilities which facilitates better development of agents that can work in physical environments.

2.2 Angry Birds

Angry Birds is a commercial game developed by Rovio Entertainment. With physical interactions similar to the real world, Angry Birds seems simple and easy to play for human players, whereas it is challenging for AI agents [25]. A successful AI agent not only needs to learn the physical properties of game objects to correctly predict the outcome of an action but also needs to choose the desired action from the action space. The game is popular among the AI community with the long-running AIBirds competition [24] as a part of the IJCAI conference [26]. The competition encourages the development of AI agents that can successfully and efficiently solve new Angry Birds game levels. Since 2013, many different strategies have been proposed to solve this challenge and none of the approaches has outperformed human performance, as has been verified by the Human vs Machine Challenge after every AI competition. Hi-Phy differs from the AIBirds competition since the benchmark specifically focuses on measuring agents’ performance in learning individual physical reasoning capabilities using tasks tailored for the purpose. The AIBirds competition uses game levels that are randomly generated or hand-crafted for the Angry Birds game. The competition evaluates an agents’ ability to play Angry Birds rather than evaluating the agent’s physical reasoning capabilities. The objective of the Hi-Phy is not to measure an agent’s Angry Birds playing capability, but to facilitate a comprehensive evaluation of the physical reasoning capabilities of agents with specifically designed tasks and evaluations.
Table 2: The proposed hierarchy. The capabilities in the upper levels of the hierarchy are composed of capabilities in the lower levels of the hierarchy.

<table>
<thead>
<tr>
<th>Level in the hierarchy</th>
<th>Capability</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 - Understanding the instant effect of the first force applied to objects in an environment as a result of an agent action.</td>
<td>1.1 Understanding the instant effect of objects in an environment when an agent applies a single force. 1.2 Understanding the instant effect of objects in an environment when an agent applies multiple forces.</td>
</tr>
<tr>
<td>L2 - Understanding objects movement in the environment after a force is applied.</td>
<td>2.1 Understanding that objects in the environment may roll. 2.2 Understanding that objects in the environment may fall. 2.3 Understanding that objects in the environment may slide. 2.4 Understanding that objects in the environment may bounce.</td>
</tr>
<tr>
<td>L3 - Understanding complex tasks</td>
<td>3.1 Understanding relative weight of objects. 3.2 Understanding relative height of objects. 3.3 Understanding relative width of objects. 3.4 Understanding shape difference of objects. 3.5 Understanding how to perform non-greedy actions. 3.6 Understanding structural weak points/stability. 3.7 Understanding how to clear a path towards the goal. 3.8 Understanding how to perform actions with adequate timing. 3.9 Understanding how to use tools.</td>
</tr>
</tbody>
</table>

3 Hierarchy for Physical Reasoning

Generalising to novel physical reasoning tasks is a hallmark of human intelligence \[27\], and remains a major obstacle for modern machine learning methods. One reason that hinders the development of physical reasoning AI agents is that it is unclear how to differentiate if the agent has really learnt to do physical reasoning, or simply found short-cuts that solve the tasks. As a step towards addressing this issue, we propose a hierarchy for physical reasoning, which enables us to systematically evaluate agents’ physical reasoning capabilities, i.e. if an agent has learnt physical reasoning capabilities.

The hierarchy contains three levels of capabilities starting from the understanding of direct consequences of an agent’s action, followed by understanding the object movements due to the force of the agent’s action, and finally to understand more complex physical concepts. Therefore, the capabilities in the higher levels of the hierarchy include lower level capabilities. When evaluating the physical reasoning capabilities of agents, we claim that if the agents can successfully perform in the higher levels of the hierarchy, but not on lower-level capabilities that are required for the higher levels, are worse in generalisation than agents that can solve tasks in both of the hierarchy levels.

In level one of the hierarchy, we consider the instant effect of the first force applied to an object in the environment as a result of an agent’s action. In this level, the agent needs to understand if different objects in the environment can be moved, not be moved, or be destroyed by applying a force. More formally, assuming we have \( n \) different objects in an environment \( O = \{o_i\}_{i=1}^{n} \) and for each of the object \( o_i \), the agent learns that after applying a force it can be moved, not be moved, or be destroyed. This means agent learns a function \( L_1 : O \times F_a \rightarrow C_1 \) where \( C_1 = \{moved, not moved, destroyed\} \) and \( F_a \) is the force exerted on the object resulting from an action. Also, the agent learns a task can be
accomplished by at least one pair of object and consequence. This means the agent learns that at least one pair of \((o, c) \in O \times C_1\) completes a task.

In the level two of the hierarchy, the agent understands how an object is moved in addition to the level one understandings. The objects can be rolled, fallen, slid, or bounced. That is, on top of \(L_1\), the agent learns a function \(L_2 : O_{\text{moved}} \times F \rightarrow C_2\), where \(C_2 = \{\text{rolling, falling, sliding, bouncing}\}\), \(O_{\text{moved}} = \{o|L_1(o, F) = \text{moved}\}\) and \(F\) is the force object \(o \in O_{\text{moved}}\) receives. Note that we relax the constraint of a force being a result of instant consequence of an agent’s action in level one.

In the third level of the hierarchy, we consider understanding complex tasks that require capabilities obtained from the second level. In this hierarchy level, we included capabilities inspired by early stage human physical reasoning development, challenges in robotics to develop agents that work alongside people in homes and workplaces, and limitations of current reinforcement learning algorithms. For physical reasoning capabilities developed in human infancy and childhood, we included understanding relative weight \([28]\), relative height \([29][30]\), relative width \([31]\), shape difference \([32][33]\), and object stability \([34]\). Moreover, we also incorporate clearing path, adequate timing, and tool usage capabilities, which are required to overcome challenges for agents to work safely and efficiently in physical environments \([35]\). Reinforcement learning algorithms, for example, policy gradients are widely understood to converge only to a local optimum \([36]\). Therefore, to encourage development of agents that do not get stuck in sub-optima, we include non-greedy actions in level three of the hierarchy.

To sum up, each level in the hierarchy contains a set of capabilities that are distinct from one another (i.e., each capability can be learnt independently from other capabilities within the same level). If an agent can do physical reasoning and successfully perform in upper levels of the hierarchy, they should be able to perform on lower levels as the capabilities in upper levels are composed of capabilities in the lower levels. Table 2 presents the hierarchy together with the capabilities. After research progresses, more levels and more capabilities to the hierarchy can be added by means such as combining existing capabilities.

## 4 Hi-Phy Physical Reasoning Benchmark

In this section, we introduce our benchmark, discuss the application of the proposed hierarchy when designing the tasks in the benchmark, and discuss the evaluation settings we use in the benchmark.

### 4.1 Introduction to the benchmark

Based on the general hierarchy proposed in Section 3, we develop a physical reasoning benchmark using Angry Birds. In Angry Birds, the player interacts with the game by shooting birds at the pigs from a slingshot. The goal of the player is to destroy all the pigs using the provided set of birds. As the original game by Rovio Entertainment is not open-sourced, we use a research clone of the game developed in Unity [37]. The game environment is a deterministic 2D world with Newtonian physics.

The game objects are of four types: birds, pigs, blocks, and platforms. There are four types of birds (red, blue, yellow, and white) in which some of them have special powers that can be activated once tapped in their flight. There are three types of pigs varying in size, the health points of the pigs increase with the increase in size. Blocks in the game are made of three materials (wood, ice, and stone) and each of them has 12 variations in shape. Platforms are static objects that remains at a fixed position and they are not affected by forces and are indestructible. All other objects are dynamic, i.e., they can be moved by applying forces. Dynamic objects have health points that get reduced upon collisions with other objects and they get destroyed and disappear when health points reach zero.

The initial state of a game level is physically stable (i.e., none of the objects is in motion) and the goal is not achieved. The action of an agent is to shoot the bird on the slingshot by providing the release coordinates relative to the slingshot and the tap time of the bird to activate special powers (if available). This means the action space is essentially continuous. When playing, an agent takes a sequence of actions, i.e., shoots the birds in a predefined order. The agent passes a game level when it destroys all pigs with the provided set of birds, and fails otherwise.

We do not provide the full world state that includes the exact location of objects in the simulator, their physical properties such as mass, friction, and bounciness to the agents as these properties are not directly observable in the real world. Instead, an agent can request screenshots and/or a symbolic
representation of the game level at any time while playing. A game screenshot is a 480 x 640 coloured image and the symbolic representation is in JSON format containing all objects in the screenshot represented as a polygon of its vertices (provided in order) and its respective colour map. The colour map provides the list of 8-bit quantized colours that appear in the game object with their respective percentages.

4.2 Hierarchy and Task Templates

According to the hierarchy of physical reasoning capabilities presented in Section 3, we design task templates in Angry Birds for each capability in the hierarchy. The task templates are handcrafted and ensure that the agent needs the specified capability to solve the tasks of each task template. We generate 100 game levels from each template and we refer to these game levels as tasks of the task template. All tasks of the same template share the same physical reasoning capabilities. Similar to [7], the tasks are generated by varying the location of the game objects in the task template within a suitable range. Furthermore, various game objects are added at random positions in the task as distractions, ensuring that they do not alter the solution of the task. Although we provide 100 tasks for each task template, we also provide a task variation generation module to generate more tasks if needed. We have developed 2-8 task templates for each capability in the hierarchy totalling 65 task templates. Figure 1 shows task templates for three different capabilities.

4.3 Evaluation Settings

Having hierarchy levels, capabilities within a hierarchy level, task templates for each capability, and 100 tasks for each task template, the benchmark allows us to measure an agent’s granular physical reasoning capabilities. In Hi-Phy, the evaluation setting we propose is the within capability evaluation. This evaluation measures an agent’s performance in tasks that require the same reasoning capabilities. Agents train on tasks from a subset of task templates of a specific capability and is tested on tasks generated from the rest of the task templates of the same capability. In order to evaluate the reasoning capability, we assume that if an agent learns the required reasoning capability to solve the task templates, it should be able to apply the same reasoning capability to solve unseen tasks from other templates within the same capability.

In our benchmark, we do not focus on agent’s within template performance as the performance on individual task templates may not represent an agent’s reasoning capability but the characteristics of the task template itself. For example, agents may have higher performance in the training set because when the training set and the test set are similar, then it is easy to find a solution through
exploration, and the solution is simple, etc. Therefore, we use the within capability evaluation setting to specifically evaluate physical reasoning capabilities of agents.

5 Experiments

We conduct experiments 1) to show that the complexity of tasks belonging to the capabilities in the hierarchy increases when moving upwards, and 2) to evaluate physical reasoning capabilities of agents using Hi-Phy. More specifically, to experimentally test our hierarchy, we evaluate both human players and deep reinforcement learning agents. We benchmark AIBirds competition heuristic agents and a learning agent (our best performing DQN agent) using Hi-Phy. We compare the results of the heuristic agents and the learning agent with the results of a random agent and human players.

5.1 Baseline Agents

We present experimental results of five baseline agents: a DQN agent, three heuristic agents from AIBirds competition, and a random agent.

Deep Q-network (DQN): The DQN [38] agent collects state-action-reward-next state quadruplets at training time by decaying epsilon greedy. The quadruplets are then stored in a replay buffer as experiences. This means the agent collects experience \( e = (s_t, a_t, r_t, s_{t+1}) \) and store it in the experience buffer \( E \) s.t. \( e \in E \). We define the reward function as task pass status, i.e., the agent receives 1 if the task is passed and 0 otherwise. The agent uses a discretized action space of 180 actions, each corresponds to a release degree from the slingshot with maximum stretch. At the end of each step, the agent trains a deep Q network on the sampled experiences to predict the Q value of actions for a given state. The Q-network comprises a state encoder that transforms an input state into a hidden representation through convolutional filters, and then to a state-value stream and to an advantage stream [39].

We tested three improvements upon a vanilla DQN: Double DQN [40], Dueling [39], and Prioritized Experience Replay [41]. We also tested using an image as input and using symbolic representation as input to train DQN. For symbolic representation, we map each game state to a \( h \times w \times o_t \) tensor, where \( o_t \) is the number of object types. In our experiments, we set \( h = 120, w = 160 \), and \( o_t = 12 \). For the state encoder, we use a \( 1 \times 1 \) convolutional filter to squeeze the channels to 1 and then use \( h \times w \) filters to extract features from the representation that preserves crucial spatial information.

For agents with image state representation, same as [7], we use ResNet-18 [42] to extract features from input image of size \( 224 \times 224 \). However, we found that the performance of the agents using images are consistently worse than the agents using symbolic representation. We hypothesize that spatial information is omitted along ResNet-18’s filters and pooling layers, especially the last \( 7 \times 7 \) average pooling layer. As we aim to compare performance with human players and other heuristic agents, we only present the performance of our best baseline learning agent, which includes all the improvements over DQN and uses symbolic representation (see Figure 2). We trained the network using an Adam optimizer [43] with a learning rate of 0.001.

Heuristic agents: The heuristics agents have fixed hard-coded strategies and do not learn through the test.
• Eagle’s Wing: Eagle’s Wing is the winner of 2017 and 2018 AIBirds competitions. This agent selects action based on strategies including shoot at pigs, destroy most blocks, shoot high round objects, and destroy structures [44].

• Datalab: Datalab is the winner of 2014 and 2015 AIBirds competitions. The agent uses strategies: destroy pigs, destroy physical structures, and shoot at round blocks. The agent selects the strategy based on the game level, possible trajectories, the current bird, and the remaining birds [45] to find a release point.

• Pig Shooter: The strategy of the Pig Shooter 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 [46].

Random agent: This agent does not perform any training. The agent selects a random release point $x$ from $(-100, -10)$ and $y$ from $(-100, 100)$ relative to the slingshot.

5.2 Experimental Setups

Human experiment setup: Experiments were approved by the Australian National University committee on human ethics under protocol 2021/293. We recruited 20 volunteer participants for the experiment. For each of the participant, we provided two tasks from each capability for 15 capabilities in the hierarchy. If the participants solved the task or failed to solve the task in five attempts, they moved on to the next task. We recorded the number of attempts used for each task, the thinking time to solve the task, and the pass/fail status of the task.

DQN experimental setup: We conduct experiments on our DQN agent in two settings: within capability and within template. For within capability evaluation, we run 50 instances of our DQN agents that collect $(s_t, a_t, r_t, s_{t+1})$ quadruplets for 45,000 task attempts in 18 update steps. In an update step, each agent randomly samples 10 tasks from the training task set and runs on each task 5 times. At the end of each step, DQN is trained with 1024 batches, each batch consists of 32 sampled experiences from the replay buffer. We train our agent on the first half of task templates in each capability and evaluate on the other half. As to discourage hyper-parameters tuning, we used the same training setting for all of the capabilities. At testing time, the agent runs on each of the testing tasks only once. The agent selects the action that has the highest $Q$-value for a given state.

As to show an agent that achieves good performance on within template evaluation still does not have physical reasoning capabilities, we also run DQN in the within template evaluation setting. We use the same training and testing setting as within capability evaluation, except we decrease the number of task attempts to 30,000 as the training tasks are from only one template.

Heuristic agents experimental setup: The three heuristic agents were tested on 20 tasks from each task template (1300 tasks in total). Agents were given five attempts to play a task.

Random agent experimental setup: The random agent was tested on 20 tasks from each task template (1300 tasks in total). The agent was given 50 attempts to play a task.

5.3 Results and Analysis

5.3.1 Complexity of the hierarchy levels

To show that the complexity of the tasks belonging to the capabilities in the proposed hierarchy increases when moving upwards, we use learning curves of the within template evaluation of the DQN agent. Figure 3a shows the average learning rates of the DQN agent across the three hierarchy levels calculated by averaging the learning curves of the agent for the task templates belonging to all the capabilities in the hierarchy level. As the results depict, the agent learns faster in the task templates in level one of the hierarchy and shows a medium learning rate in level two and the slowest in level three.

Further, we use the percentage of failing tasks, number of attempts taken to solve the tasks, and thinking time taken to solve the tasks of the human participant’s evaluation. For a given task, we measure the thinking time of the participant starting from the time the participant sees the task to the participant’s first action (first bird shoot) in all the attempts until the participant solves the
(a) The average learning rate of the DQN agents across the three hierarchy levels

(b) The fail rate, number of attempts taken to pass, and the time taken in each level of the hierarchy by human players

Figure 3: Data analysis for the three hierarchy levels. From (a) it can be seen that agents learn faster in level 1 tasks of the hierarchy while it is the slowest in level 3. (b) represents normalized failing rates, number of attempts taken to pass, and the time taken by human players across the tasks in the hierarchy levels. All normalized values range from 0-1 and 0 is at the origin. Tasks in level 3 have the highest values while they are the lowest in level 1. These charts show the increasing complexity of tasks across the hierarchy.

We make the assumption that the participant thinks of the solution for the task within this 312 time. Figure 3b shows the normalized failing percentages, number of attempts, and thinking time of 314 human participants for each hierarchy level. As shown in the star plot, all three values increase as the 315 hierarchy level increases.

Moreover, the average passing rates of the random agent for the three hierarchy levels were 5.24%, 317 4.85%, and 2.66% respectively. Therefore, as illustrated from the results of the above three experi- 318 ments, the tasks belonging to the capabilities of the proposed physical reasoning hierarchy increases 319 in complexity when moving upwards in the hierarchy.

5.3.2 Establishing baseline results

We establish the baseline results for the best-performing DQN agent in Hi-Phy. Further, using Hi-Phy, 321 we analyze the physical reasoning capabilities of the AIBirds heuristic agents. Then, we make a 322 comparison of the results of humans, DQN agent, heuristic agents, and the random agent. We use 323 the within capability evaluation setup of the DQN agent discussed in Section 5.2, since the within 324 template evaluation does not measure the physical reasoning capabilities. We discuss the results of 325 within template evaluation in Appendix C.

The results of the human participants and the five agents are shown in Figure 4 for the 15 physical 327 reasoning capabilities in Hi-Phy. The DQN agent’s results shown are the offline agent’s results on 328 the test templates task sets of each capability. From the results of the DQN agent, it can be seen 329 that the agent performs better on the tasks of level one capabilities compared to its results on the 330 other two levels. Also, the DQN agent performs better than the random agent across the tasks of 331 all the capabilities. The three heuristic agents results are better than the DQN agent’s results, but 332 they perform relatively worse on higher level capabilities of the hierarchy. Even though they perform 333 better than the DQN agent, their strategies are hard-coded according to the capabilities needed to 334 solve Angry Birds game levels. From the human results it is clear that the human performance is far 335 beyond the performance of all the other agents performance in Hi-Phy. This shows that there is a 336 substantial gap in human capabilities and AI agents in solving physical reasoning tasks.

6 Discussion and Future Work

In this work, inspired by how humans learn, we presented a general hierarchy of physical reasoning 339 capabilities. The hierarchy starts with simple capabilities and increases in complexity when moving 340 higher. There are 15 different capabilities in the 3 hierarchy levels, and we designed, in total, 65
templates for these capabilities. Unlimited number of tasks, which require a particular reasoning
capability to solve, can be generated from the templates. Based on this hierarchy, we presented
Hi-Phy, a benchmark for measuring individual physical reasoning capabilities of AI agents. The
baseline methods evaluated in this work are far from reaching the human performance. The evaluation
shows heuristic agents can perform better than the experimented DQN agent. But, as heuristics are
human hard-coded strategies to solve physical reasoning tasks, they cannot perform on any task
outside their fixed scope. Therefore, with Hi-Phy we encourage developing intelligent algorithms
that has the ability to learn physical reasoning capabilities.

The hierarchy of physical reasoning capabilities we proposed in this work is general and can be
used to design tasks for other physical domains including 3D environments. We propose different
directions to advance the hierarchy as future work. Firstly, object deformations can be added as a
level one capability and subsequent capabilities can be introduced to the higher levels accordingly.
Moreover, a higher level to the hierarchy can be added by combining the capabilities in the existing
levels. So in this level, we would test the agent’s combinatorial generalisation. Further, the hierarchy
can be expanded by adding physical reasoning capabilities such as understanding shape constancy,
object permanence, spatiotemporal continuity, and causality. Those capabilities can be tested by
introducing novel objects to the tasks in the benchmark.

We hope that our general physical reasoning hierarchy and Hi-Phy can provide a foundation for
future research on developing AI agents with human-level physical reasoning capabilities, thereby
coordinating research efforts towards ever new goals.

References

Tenenbaum. Clever: Collision events for video representation and reasoning. In International Conference


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Explained in the conclusion
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] The paper conforms to the ethical guidelines

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Provided in the URL to the benchmark
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Provided in Appendix C and D
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Provided in Appendix C and E
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Provided in Appendix F

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] The research clone of Angry Birds is cited.
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Provided in the benchmark URL
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] Provided in Appendix E
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]