nature machine intelligence

Article

https://doi.org/10.1038/s42256-022-00583-4

Phy-Q as a measure for physical reasoning intelligence

Received: 18 May 2022

Accepted: 4 November 2022

Published online: 25 January 2023

Check for updates

Cheng Xue[1,2] ✉, Vimukthini Pinto[1,2] ✉, Chathura Gamage[1,2] ✉, Ekaterina Nikonova[1], Peng Zhang[1] & Jochen Renz[1]

Humans are well versed in reasoning about the behaviours of physical objects and choosing actions accordingly to accomplish tasks, while this remains a major challenge for artificial intelligence. To facilitate research addressing this problem, we propose a new testbed that requires an agent to reason about physical scenarios and take an action appropriately. Inspired by the physical knowledge acquired in infancy and the capabilities required for robots to operate in real-world environments, we identify 15 essential physical scenarios. We create a wide variety of distinct task templates, and we ensure that all the task templates within the same scenario can be solved by using one specific strategic physical rule. By having such a design, we evaluate two distinct levels of generalization, namely local generalization and broad generalization. We conduct an extensive evaluation with human players, learning agents with various input types and architectures, and heuristic agents with different strategies. Inspired by how the human intelligence quotient is calculated, we define the physical reasoning quotient (Phy-Q score) that reflects the physical reasoning intelligence of an agent using the physical scenarios we considered. Our evaluation shows that (1) all the agents are far below human performance, and (2) learning agents, even with good local generalization ability, struggle to learn the underlying physical reasoning rules and fail to generalize broadly. We encourage the development of intelligent agents that can reach the human-level Phy-Q score.

The ability to reason about objects' properties and behaviours in physical environments lies at the core of human cognitive development[1]. A few days after birth, infants understand object solidity[2], and within the first year after birth, they understand notions such as object permanence[3], spatiotemporal continuity[4], stability[5], support[6], causality[7] and shape constancy[8]. Generalization performance on novel physical puzzles is commonly used as a measure of physical reasoning abilities for children[9,10], animals[11] and artificial intelligence (AI) agents[12–15].

Chollet's study[16] on the measure of intelligence proposes a qualitative spectrum of different forms of generalization that includes local generalization and broad generalization. Current evidence[17–19] suggests that contemporary deep learning models are local generalization systems, that is, systems that adapt to known unknowns within a single task. Broad generalization, on the other hand, can be characterized as 'adaptation to unknown unknowns across a broad category of related tasks' and is being increasingly emphasized among the AI research community[16,20]. Moreover, when solving physics puzzles, it is common that a player must use a strategy to work out a plan and use dexterity to accurately execute the strategic plan[21]. For instance, in a snooker game, a player needs to plan the path of the white cue ball, for example, where it should go and where it should stop, and then execute the strike that precisely produces the planned path. Some cognitive psychology researchers believe that humans possess inaccurate forward physics prediction models[22,23] and hence require practice to improve dexterity,

[1]School of Computing, The Australian National University, Canberra, Australian Capital Territory, Australia. [2]These authors contributed equally: Cheng Xue, Vimukthini Pinto and Chathura Gamage. ✉e-mail: cheng.xue@anu.edu.au; vimukthini.inguruwattage@anu.edu.au; chathura.gamage@anu.edu.au

Table 1 | Comparison of Phy-Q with related physics benchmarks and competitions

Test environment	Generalization to individual physical scenario/s	Categorization of tasks to physical scenarios	Procedurally generated tasks/variations	Destructible objects	Observe outcome of a desired physical action	Human player data
PHYRE[12]	✗	✗	✓	✗	✓	✗
Virtual Tools[14]	✗	✓	✗	✗	✓	✓
OGRE[13]	✗	✗	✓	✗	✓	✗
IntPhys 2019 [24]	✓	✓	✓	✗	✗	✓
CLEVERER[25]	✓	✗	✓	✗	✗	✗
CATER[31]	✓	✓	✓	✗	✗	✗
Physion[27]	✓	✓	✓	✗	✗	✓
COPHY[26]	✓	✓	✓	✗	✗	✓
CausalWorld[15]	✓	✓	✓	✗	✓	✗
RLBench[32]	✗	✗	✓	✗	✓	✗
Computational Pool[33]	✗	✗	✗	✗	✓	✗
Geometry Friends[34]	✗	✗	✓	✗	✓	✗
AIBIRDS[35]	✗	✗	✓	✓	✓	✓
Phy-Q (this study)	✓	✓	✓	✓	✓	✓

high dexterity requirements of physics tasks make it unfair to compare AI agents' physical reasoning ability with that of average humans. For example, when a human player fails a physics puzzle, it is hard to tell if this is owing to incorrect physical reasoning or the inability to make precise actions (dexterity). Despite the recent advancement in physical reasoning benchmarks and testbeds[12–15,24–27], there is a lack of a benchmark or a testbed with human-comparable strategic physics puzzles and that explicitly evaluates learning agents' local and broad generalization.

To close these gaps, we propose a new testbed (Phy-Q) and the associated Phy-Q score that measures physical reasoning intelligence using the physical scenarios we identified. Inspired by the physical knowledge acquired in infancy and the abilities required by the robots to operate in the real world, we created a wide variety of tasks with low dexterity requirements in the video game Angry Birds[28]. We believe that the contributions of this paper pave the way for the development of agents with human-level strategic physical reasoning capabilities.

Our main contributions can be summarized as follows:

- Phy-Q: A testbed for physical reasoning. We designed a variety of task templates in Angry Birds with 15 physical scenarios, where all the task templates of a scenario can be solved by following a common strategic physical rule. Then, we generated task instances from the templates using a task variation generator. This design allows us to evaluate both the local and the broad generalization ability of an agent. We also define the Phy-Q score, a quantitative measure that reflects physical reasoning intelligence using the physical scenarios we considered.
- An agent-friendly framework. To facilitate agent training in our testbed, we propose a framework that allows the training of multi-agent instances simultaneously with game play speed accelerated up to 50 fold.
- Establishing results for baseline agents. The evaluation consists of nine baseline agents: four of our best-performing learning agents, four heuristic-based agents and a random agent. For each of the baseline agents, we present the Phy-Q score, the broad generalization performance and the local generalization performance. We have collected human player data so that agent performance can be compared directly with human performance.
- Guidance for agents in the AIBIRDS competition. Angry Birds is a popular physical reasoning domain among AI researchers, with the AIBIRDS competition running since 2012 (ref. [29]). In 2016, Angry Birds was considered to be the next milestone where AI will surpass humans[30]. A time horizon of 4 years was predicted, but so far such a breakthrough seems very unlikely. In the AIBIRDS competition, heuristic methods generally perform better than their deep learning counterparts, but it remains unclear what has contributed to the gap between their performance. It has also not yet been analysed why current AI agents fall short when compared with humans. By the systematic analysis of agents from the AIBIRDS competition, we show how they need to be improved to achieve human-level performance.

Background and related work

In this section, we conduct a comparison between ten related physical reasoning benchmarks and two physics-based AI game competitions to show how the Phy-Q testbed advances upon existing work. The comparison is done with respect to six criteria:

1. Measuring broad generalization in individual physical scenario(s), that is, testing the ability of an agent to generalize to tasks that require the same physical rule to solve.
2. Categorization of tasks of the test environment into different physical scenarios, that is, agents can be evaluated for individual scenarios to recognize the scenarios that they can perform well.
3. Procedural generation of tasks or variations of the tasks, that is, the tasks/variants of the tasks in the test environment are created algorithmically, helping users to generate any amount of data.
4. Destructibility of objects in the environment, that is, if the environment contains objects that can be destroyed upon the application of forces. Having destructible objects makes the environment more realistic than an environment that only has indestructible objects since the agents need to consider the magnitude of the force that is applied to the objects. For example, when a robot moves a cup, it needs to reason that the force to exert should be large enough to grab the cup but not large enough to break it.
5. Observing the outcome of a desired physical action, that is, whether an agent can physically interact and observe the outcome of the action the agent takes.

https://doi.org/10.1038/s42256-022-00583-4

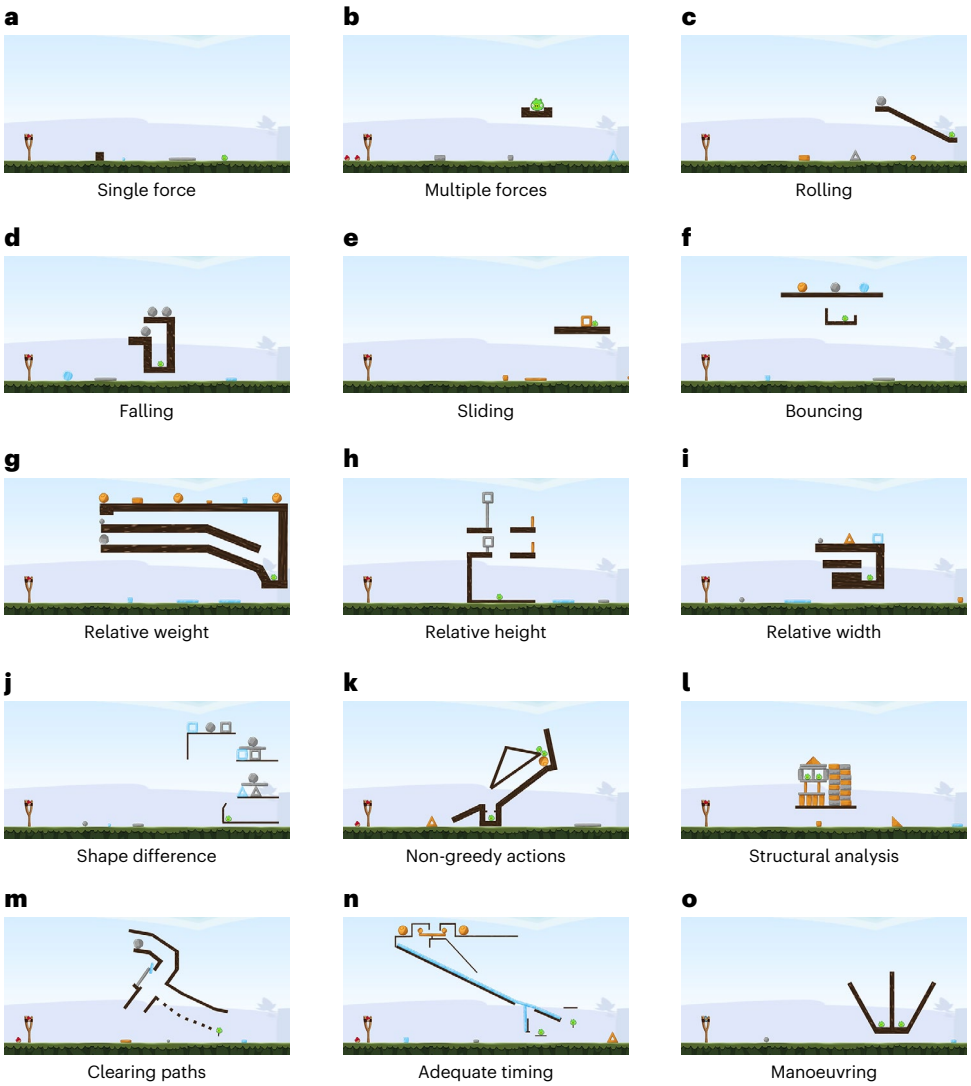


Fig. 1 | Example tasks in Phy-Q representing the 15 physical scenarios. The slingshot with birds is situated on the left of the task. The goal of the agent is to kill all the green pigs by shooting birds from the slingshot. The dark-brown objects are static platforms. The objects with other colours are dynamic and subject to the physics in the environments.

6. Inclusion of human player data, that is, if the evaluation has results of human players.

We consider PHYRE[12], the Virtual Tools game[14] and OGRE[13] as game-based benchmarks, IntPhys[24], CLEVERER[25], CATER[31] and Physion[27] as video-based benchmarks, COPHY[26] as an image-based benchmark, and CausalWorld[15] and RLBench[32] as robotic benchmarks. The AI game competitions we consider are Computational Pool[33] and Geometry Friends[34]. We also included the AIBIRDS[35] competition in the comparison to show what properties in Phy-Q facilitate the systematic evaluation of AIBIRDS competition agents. Table 1 summarizes the comparison.

The physical reasoning test environment that is most closely related to ours is PHYRE[12], which also consists of tasks to measure two levels of generalization of agents. The PHYRE benchmark tests whether agents can generalize to solve tasks within a task template (within template) and whether agents can generalize between different task templates (cross template). The cross-template evaluation in PHYRE does not guarantee that the physical rules required to solve the testing tasks exist in the training tasks. This leads to uncertainties in understanding agents' performance: inferior performance may not be an indicator of inferior physical reasoning but of a difficult training and testing split. In contrast, the broad generalization evaluation in our testbed always ensures that the physical rules required in testing tasks are covered in the training tasks, thereby guaranteeing a more systematic evaluation of the physical reasoning capabilities of AI agents. According to the task design in PHYRE, tasks must be solved by trial and error. That is, even when the physical rule is known, multiple attempts are still needed to solve the tasks. Therefore, PHYRE promotes the development of agents with physical dexterity. In contrast, we focus on strategy-based physical reasoning tasks that can be solved in a single attempt when the physical rule is understood. We promote the development of agents that can understand a physical rule rather than taking a precise action in a physical environment (that is, agents with strategic physical reasoning capabilities). Furthermore, a limited number of object shapes, motion and material properties on scene dynamics hinder the ability for a comprehensive evaluation, as performing well on these tests might not indicate a greater physical reasoning ability in more general and realistic contexts[27]. Therefore, compared with PHYRE, Phy-Q offers (1) three more object shapes (rectangles, squares and triangles) to allow more diverse physical dynamics, (2) destructible objects to make our environment more realistic and (3) objects with three different materials that have different densities, bounciness and friction to allow physical reasoning in a more realistic context.

https://doi.org/10.1038/s42256-022-00583-4

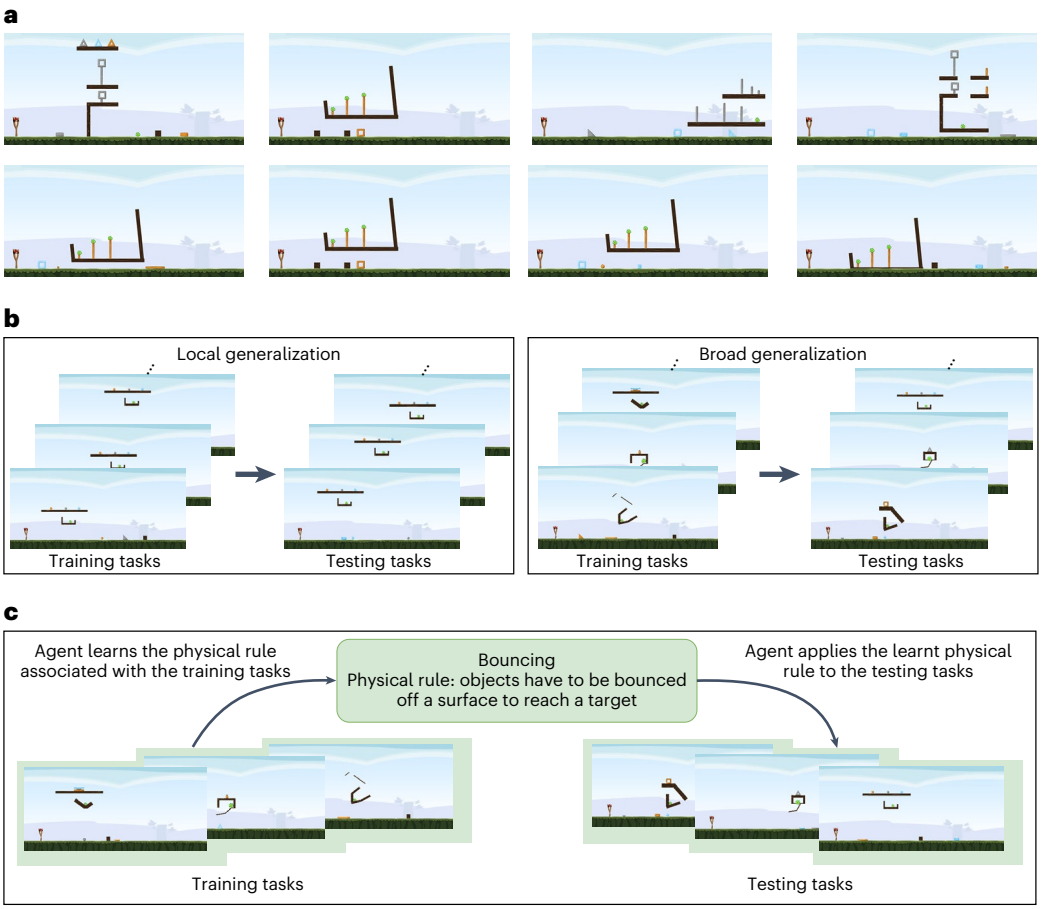


Fig. 2 | Phy-Q task templates and evaluation settings. a, The task templates of the relative height scenario (first row) and the tasks generated using the second task template in the first row (second row). **b**, The local generalization and the broad generalization evaluation settings. **c**, An illustration of how generalizing a physical rule is evaluated in the broad generalization evaluation using the bouncing scenario as an example.

As a recent visual and physical prediction benchmark, Physion[27] evaluates algorithms' physical prediction capability using videos of eight different physical scenarios. Compared with Physion, the Phy-Q testbed has a more comprehensive set of 15 physical scenarios enabling the evaluation of agents in a wider range of physical scenarios. The Phy-Q testbed requires agents to interact with the environment and select the desired action to accomplish physical tasks. Therefore, on top of predicting a physical event's outcome, agents need to apply the acquired physical knowledge to solve new situations, which is considered to be a more advanced type of task in Bloom's taxonomy[36]. In addition, a study on forward prediction for physical reasoning[37] confirms that higher forward prediction accuracy does not necessarily increase performance in domains that require selecting an action. Therefore, Physion and Phy-Q focus on different research problems.

Despite Angry Birds being a simplified and controlled physics environment as compared with the much messier real physical world, no AI system that comes close to human performance has been developed. To encourage the development of AI agents that can reason with physics as humans do, the AIBIRDS competition has been organized annually since 2012, mostly held at the International Joint Conference on Artificial Intelligence[35]. Since then, many different AI approaches have been proposed, ranging from modern deep reinforcement learning methods to more old-school heuristic methods, for example, qualitative physical reasoning methods. However, none of these approaches has reached the milestone of achieving human-level performance. One major reason is that an agent's performance in the competition does not enable an agent developer to identify the physical scenarios that the agent falls short of. This is because the tasks in the competition are generally complex with multiple physical scenarios within the same task. In this work, we show how the Phy-Q testbed can be used towards guiding the competition agents through a systematic evaluation of agents' performance.

The Phy-Q testbed

In this section, we introduce our testbed and discuss the physical scenarios we have identified.

Introduction to the Phy-Q testbed

Based on the 15 identified physical scenarios (discussed in detail in Section 3.2), we develop a physical reasoning testbed using Angry Birds. In Angry Birds, the player interacts with the game by shooting birds at 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[38]. The game environment is a deterministic two-dimensional world where objects in motion follow Newtonian physics. The game objects are of four types: birds, pigs, blocks and platforms. There are five types of birds, four of which have 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 their 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 remain at a fixed position and are not affected by forces and are indestructible. All other objects are dynamic, that is, can be moved by applying forces.

https://doi.org/10.1038/s42256-022-00583-4

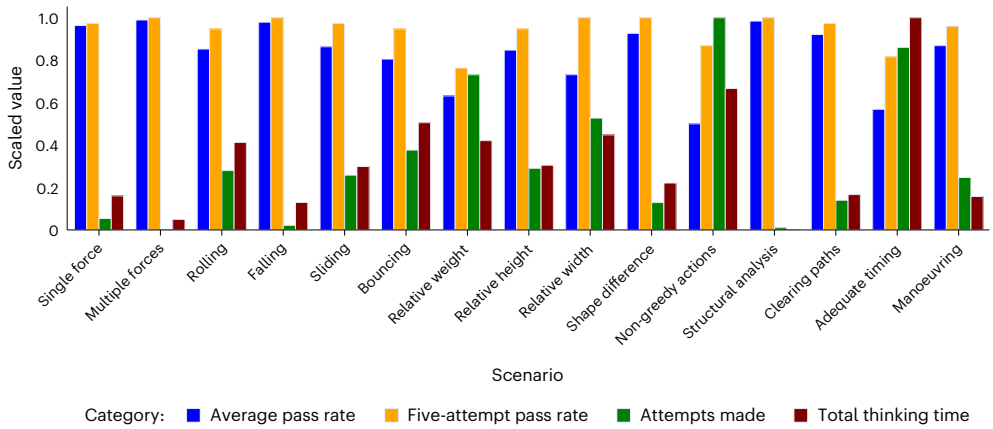


Fig. 3 | Performance of human players.

Dynamic objects have health points that are reduced upon collisions with other objects, and they get destroyed and disappear when their health points reach zero. The initial state of a game level is physically stable (that is, 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. We have included a module that aids trajectory planning to reduce the dexterity requirement. Additionally, the agent provides the tap time of the bird to activate powers (if available). The selection of the release point and the tap time makes the action space essentially continuous. When playing, an agent takes a sequence of actions, that is, shoots the birds in a predefined order. The agent passes a game level when it destroys all the 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 or their physical properties such as mass and friction 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 × 640 coloured image, and the symbolic representation is in JavaScript object notation format, containing all the 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 eight-bit quantized colours that appear in the game object with their respective percentages.

Physical scenarios in the Phy-Q testbed

In this section, we introduce the 15 physical scenarios we consider in our testbed. Firstly, we consider the basic physical scenarios associated with applying forces directly on the target objects, that is, the effect of a single force and the effect of multiple forces[39]. On top of the application of a single force, we also include scenarios associated with more complex motion including rolling, falling, sliding and bouncing, which are inspired by the physical reasoning capabilities developed in human infancy[40]. Furthermore, we define the objects' relative weight[41], the relative height[42], the relative width[43], the shape differences[44] and the stability[45] scenarios, which require physical reasoning abilities that infants acquire typically at a later stage. On the other hand, we also incorporate clearing path, adequate timing and manoeuvring[46] and taking non-greedy actions[47], which are required to overcome challenges for robots to work safely and efficiently in physical environments. Each of these scenarios tests a different aspect of the agent's skill, physical understanding and planning ability. To sum up, the physical scenarios we consider and the corresponding high-level strategic physical rules that can be used to achieve the goal of the associated tasks are mentioned below. Example task templates from those scenarios are shown in Fig. 1.

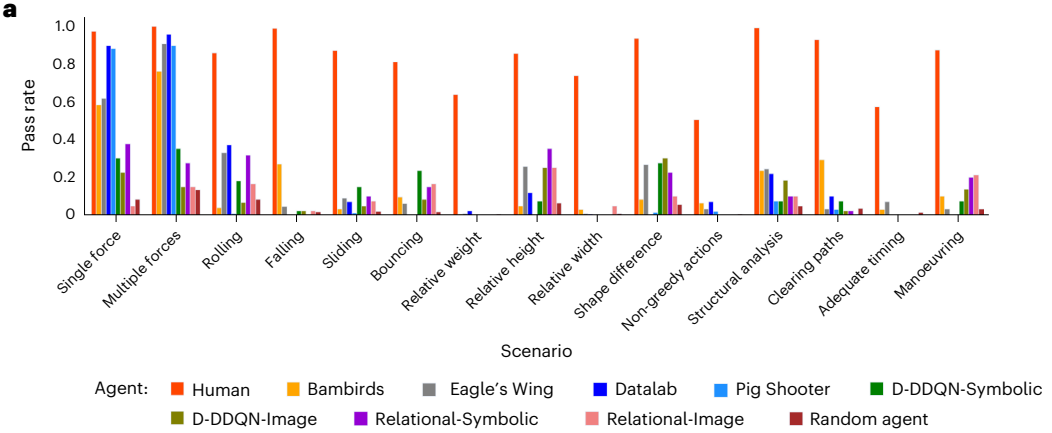
1. Single force: Target objects have to be destroyed with a single force.
2. Multiple forces: Target objects need multiple forces to be destroyed.
3. Rolling: Circular objects have to be rolled along a surface to a target.
4. Falling: Objects have to fall onto a target.
5. Sliding: Non-circular objects have to be slid along a surface to a target.
6. Bouncing: Objects have to be bounced off a surface to reach a target.
7. Relative weight: Objects with the correct weight have to be moved to reach a target.
8. Relative height: Objects with the correct height have to be moved to reach a target.
9. Relative width: Objects with the correct width or the opening with the correct width have to be selected to reach a target.
10. Shape difference: Objects with the correct shape have to be moved/destroyed to reach a target.
11. Non-greedy actions: Actions have to be selected in the correct order based on physical consequences. The immediate action may be less effective in the short term but advantageous in long term, that is, reach fewer targets in the short term to reach more targets later.
12. Structural analysis: The correct target has to be chosen to break the stability of a structure.
13. Clearing paths: A path must be created before the target can be reached.
14. Adequate timing: Correct actions have to be performed within time constraints.
15. Manoeuvring: Objects have to be carefully guided to reach a target.

Conclusion and future work

The goal of the Phy-Q testbed is to facilitate the development of physical reasoning AI methods with broad generalizing abilities similar to that of humans. As mentioned above, humans may possess inaccurate forward physics prediction models. We focus on tasks that can be solved by using a strategic physical rule and with low dexterity requirements instead of tasks that require precise forward prediction. Therefore, towards that goal, we designed 75 task templates considering 15 different physical scenarios in our testbed. The tasks that belong to the same physical scenario can be solved by a specific strategic physical rule, enabling us to measure the broad generalization of agents by allowing the agent to learn a strategic physical rule in the learning phase that can be used in the testing phase. Apart from the broad generalization

https://doi.org/10.1038/s42256-022-00583-4

a



b

	Human	BB	EW	DL	PS	D-DDQN-Sy	D-DDQN-Im	Rel-Sy	Rel-Im	Random
Local generalization	–	0.1536	0.1367	0.1481	0.1050	**0.3350**	0.2487	0.3220	0.2393	0.0338
Broad generalization	0.8288	0.1776	**0.1999**	0.1880	0.1282	0.1209	0.0997	0.1411	0.0897	0.0427

Fig. 4 | Generalization performance of the baseline agents. a, Broad generalization of the baseline agents. **b**, Comparison of agents for local generalization and broad generalization (Phy-Q score) for nine agents: Bambirds (BB), Eagle's Wing (EW), Datalab (DL), Pig Shooter (PS), D-DDQN-Symbolic (D-DDQN-Sy), D-DDQN-Image (D-DDQN-Im), Relational-Symbolic (Rel-Sy), Relational-Image (Rel-Im) and random (Random). The performance of the best-performing agent is shown in bold. Learning agents have higher local generalization values but lower values in broad generalization than the performance of heuristic agents. Human performance is way beyond agents.

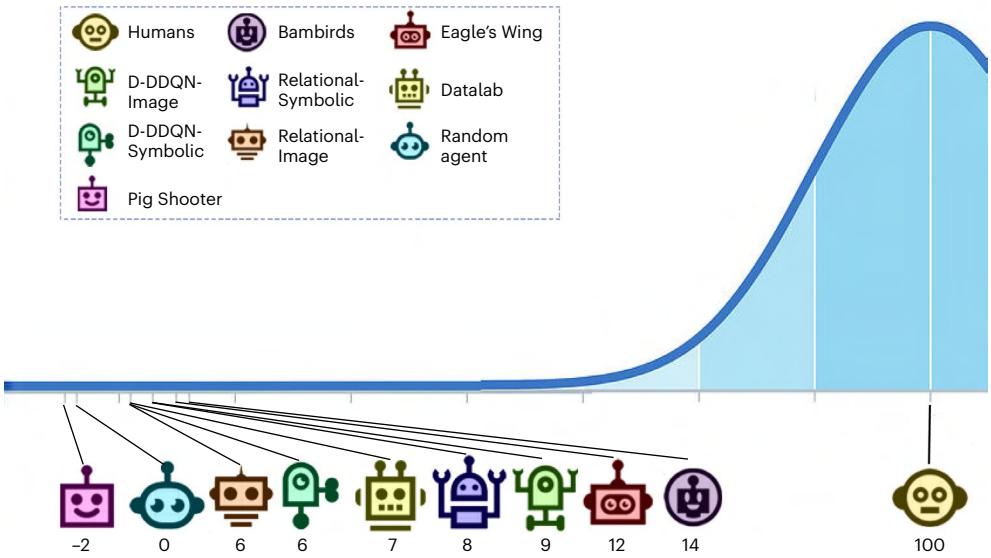


Fig. 5 | Phy-Q score of human players and the agents. The Phy-Q score of human players is at 100, and all agents are far below human players.

performance evaluation, the Phy-Q testbed also enables evaluating agents' local generalization performance. We have established baseline results from the testbed and have shown that, even though current learning agents can generalize locally, the broad generalization ability of these agents is below heuristic agents and far below human performance. Further, we have defined the Phy-Q score to reflect the physical reasoning ability of agents. In addition, we have shown how the testbed can be used for the advancement of the AIBIRDS competition agents.

Although we discourage the development of heuristic agents with hard-coded rules that apply only to Angry Birds, we believe that the superior performance of these rule-based systems, given that none of the agent developers has seen the Phy-Q tasks previously, indicates that the human-extracted strategic physical rules are highly generalizable. Therefore, we foresee several areas of improvement: (1) Agents should learn and store generalizable abstract causal knowledge[48], for example, strategic physical rules. For example, humans understand not only that shooting a bird at a pig can destroy the pig, but also that the pig is destroyed because, when the bird hits the pig, a force is exerted by the bird on the pig[49] and, if the force is large enough, an object will be destroyed. One possible way to learn this abstract causal knowledge is through explanation-based learning[50], where an agent constructs an explanation for initial exemplars and then constructs a candidate rule that depends only on the explanation. If the rule is proven true for a small number of additional exemplars, the rule is adopted. As the representation of abstract and causal knowledge allows for symbolic manipulation[48], (2) it is also worthwhile to explore the possibility of combining deep learning techniques with reasoning over knowledge systems in physical domains. Neural symbolic methods, such as Neuro-Symbolic Dynamic Reasoning[25], have shown promising results on physical reasoning.

Phy-Q can be advanced in different directions. Characteristics such as deforming can be introduced to the objects in the tasks. Further, complex scenarios can be added to the testbed by combining the existing scenarios. This will also enable the combinatorial generalization of the agents to be measured. Moreover, additional physical scenarios that are not covered in the testbed could be introduced, such as shape constancy, object permanence, spatiotemporal continuity, and causality. We hope that Phy-Q will provide a foundation for future research on the development of AI agents with human-level physical reasoning capabilities, thereby coordinating research efforts towards ever new goals.

Methods

Phy-Q testbed tasks and evaluation

In this section, we discuss the details of the designing of task templates and the generation of task instances. We also explain the evaluation settings we have used in the testbed.

Task templates and task generation. We design task templates in Angry Birds for each of the 15 physical scenarios mentioned above. A task template can be solved by a specific strategic physical rule, and all the templates belonging to the same scenario can be solved by the high-level strategic physical rules discussed above. To guarantee this, in the Phy-Q testbed, we hand-crafted the task templates because existing task generators for Angry Birds[51,52] do not generate tasks according to a strategic physical rule. Also, we ensure that, if an agent understood the strategic physical rule to solve the template, it can solve the template without requiring highly accurate shooting, for example, the template can be solved by shooting at a specific object rather than shooting a specific coordinate. This design criterion is followed to reduce the dexterity requirement when solving the tasks in our testbed. We have developed 2–8 task templates for each scenario, totalling 75 task templates. Figure 1 shows example task templates for the 15 scenarios.

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 strategic physical rule to solve. Similar to ref. [12], 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. When generating the tasks, each task template has constraints to satisfy such that the physical rule of the template is preserved. For example, the constraints can be: which game objects should be directly reachable by a bird shot from the slingshot, which game objects should be unreachable to the bird, which locations in the game level space are feasible to place the game objects, etc. These constraints are specific to each task template. They were determined by the template developers and hard coded in the task generator.

Although we provide 100 tasks for each task template, we also provide a task variation generation module to generate more tasks if needed. Figure 2a shows task templates of the relative height scenario and example tasks generated from a single task template. All 75 task templates and example task variations can be found in Supplementary Sect.. C.

Proposed evaluation settings. The spectrum of generalization proposed by Chollet[16] can be used to measure intelligence as laid out by theories of the structure of intelligence in cognitive psychology. There are three different levels in the spectrum: local generalization, broad generalization and extreme generalization. Having 15 physical scenarios, a variety of task templates for each scenario and task variations for each task template, our testbed is capable of evaluating all three different generalization levels. However, in this work, we focus on measuring the local generalization and the broad generalization of agents, as local generalization is the form of generalization that has been studied from the 1950s up to this day and there is increasing research interest in achieving broad generalization[16].

More formally, consider each scenario_i in the set of all scenarios SCENARIO, where $|\text{SCENARIO}| = 15$. We define $\text{template}_j \in \text{scenario}_i$, where $|\text{scenario}_i| = \text{NT}_i$ and NT_i is the number of templates we included for scenario_i. As we have 100 tasks for each templates, we define $\text{task}_k \in \text{template}_j$, where $|\text{template}_j| = 100$ for all templates, that is, each scenario is a set of tasks and the tasks in a scenario are partitioned into templates.

To evaluate local generalization within a particular template, we train an agent on some (80% in practice) of the tasks in a template and evaluate it on the remaining tasks of the same template. To evaluate broad generalization within a particular scenario, we train an agent on the tasks of some of the templates of that scenario and evaluate it on the tasks of the other templates of the same scenario (see Supplementary Sect. E for the division of task templates for training and testing for each scenario).

We evaluate the broad generalization performance for all 15 scenarios. We assume that, if an agent learns the strategic physical rule required to solve a set of task templates, it should be able to apply the same strategic physical rule to solve unseen tasks from other templates within the same scenario. As opposed to this, the performance on local generalization evaluation may not represent an agent's physical rule generalizing capability but memorizing a special-purpose heuristic. Figure 2b shows a diagrammatic representation of the two evaluation settings, and Fig. 2c shows an illustration of how generalizing a physical rule is evaluated in the broad generalization evaluation setting.

Our physical reasoning quotient (Phy-Q) is inspired by the deviation intelligence quotient[53] of humans. We calculate the Phy-Q of an agent by using the results of our broad generalization evaluation, since we consider that this evaluation measures the agent's ability in generalizing strategic physical rules. When calculating the Phy-Q, we exclude the first two scenarios (single force and multiple forces), as the solution for these two scenarios is directly shooting the bird to the exact location of the pig. Given that we have provided a trajectory planner for both humans and agents, solving the tasks of these two scenarios is straightforward. This is also evident from the exceptionally high results (Section 5.3.2) of the Pig Shooter agent that directly shoots at the pigs without doing any physical reasoning. We define the Phy-Q score as follows:

$$Z_{\text{agent}} = \frac{1}{|\text{SCENARIO} - \{\text{scenario}_1, \text{scenario}_2\}|} \sum_{m=3}^{|\text{SCENARIO}|} \frac{P_{\text{agent},m} - P_{\text{human},m}}{\sigma_{\text{human},m}}, \tag{1}$$

$$\text{Phy-Qscore}_{\text{agent}} = 100 + Z_{\text{agent}} \frac{100}{|Z_{\text{random}}|}, \tag{2}$$

where $P_{n,m}$ is the average pass rate of subject n in the mth scenario. $\sigma_{\text{human},m}$ is the s.d. of the human pass rate in scenario m and 'random' indicates the random agent that selects a random action (Section 5.2.1). A Phy-Q score of 100 represents an agent having an average human level performance, whereas if the score is less than 100, the agent's performance is less than the average human performance and vice versa. As the random agent does not have any physical reasoning capabilities, we bring the random agent's Phy-Q score to zero. Therefore, we set the scaling factor to $100/|Z_{\text{random}}|$, which is 13.58, as compared with 15 for the intelligence quotient. Therefore, a Phy-Q score of more than zero indicates performance better than an agent that selects a random action.

Experiments

We conduct experiments on baseline learning agents to measure how well they can generalize in two different settings: local generalization and broad generalization. We also conduct experiments using heuristic baseline agents in the two generalization settings. In addition, we

establish human performance in the 15 scenarios. Further, we conduct an additional experiment using heuristic agents in AIBIRDS competition game levels to examine whether the performance of agents in the testbed resembles the performance in the competition.

Baseline agents. We present experimental results of nine baseline agents: two DQN agents (one using screenshot input and the other using symbolic representation), two relational agents (one using screenshot input and the other using symbolic representation), four heuristic agents from the AIBIRDS competition and a random agent.

Learning agents: For the learning agents, we tested value-based and policy-based (Supplementary Sect. I) reinforcement learning algorithms and report the results of double duelling deep Q-network (DQN) agents and relational DQN agents.

- DQN: The DQN[54] agent collects state–action–reward–next state quadruplets at the training time following decaying epsilon greedy. We define the reward function as task pass status, meaning that the agent receives 1 if the task is passed and 0 otherwise. We report the performance of double duelling DQN[55,56] with two different input types: symbolic representation (D-DDQN-Image) and screenshot (D-DDQN-Symbolic).
- Relational DQN: The relational agent consists of the relational module[57] that was built on top of the deep Q-network. The aim of this agent is to generalize over the presented templates/events by using structured perception and relational reasoning. In our experiments, we wanted to test whether the relational agent would be able to learn the important relations between the objects that could be generalized to other templates or events. We have tested the agent with symbolic and image input types and refer to them as Relational-Symbolic and Relational-Image agents, respectively.

Heuristic agents: The heuristic agents are based on hard-coded strategic physical rules designed by the developers. We included four heuristic agents from the AIBIRDS competition. We compare the heuristic agents' performance on our testbed with the generalization performance of the baseline learning agents.

- Bambirds: Bambirds was the winner of the 2016 and 2019 AIBIRDS competitions. The agent chooses one of nine different strategies. The strategies include creating a domino effect, targeting blocks that support heavy objects, maximum structure penetration, prioritizing protective blocks, targeting pigs and utilizing certain bird's powers[58].
- Eagle's Wing: Eagle's Wing was the winner of the 2017 and 2018 AIBIRDS competitions. This agent selects an action based on strategies including shoot at pigs, destroy most blocks, shoot high round objects and destroy structures[59].
- Datalab: Datalab was the winner of the 2014 and 2015 AIBIRDS competitions. The agent uses the following strategies: destroy pigs, destroy physical structures and shoot at round blocks. The agent selects a strategy based on the game states, possible trajectories, bird types and the remaining birds[60].
- Pig Shooter: The strategy of the Pig Shooter is to shoot directly at the pigs. The agent shoots the bird on the slingshot by randomly selecting a pig and a trajectory to shoot the pig[61].

Random agent: For each shot, the agent selects a random release point (x, y), where x is sampled from $[-100, -10]$ and y from $[-100, 100]$ relative to the slingshot. It also provides a tapping time when the bird is between 50% and 80% of the trajectory length, where applicable.

Experimental setups. Human experiment setup: Experiments were approved by the Australian National University committee on human ethics under protocol 2021/293. Participation was voluntary with no monetary compensation. The volunteers were males and females with age in the range of 18–35 years. They were not experienced Angry Birds players. Participants provided consent to use their play data. For each of them, we provided two tasks from each physical scenario for the 15 scenarios in Phy-Q (except the manoeuvring scenario, which used four tasks representing the four types of birds with powers). We provided a trajectory visualizer of the bird to the participants to remove the need for precise shooting. If the participants solved a task or failed to solve a task in five attempts, they moved on to the next task. As humans acquire physical reasoning capabilities from their infancy[40,62], using an evaluation setting that we proposed for agents does not exactly measure the generalization ability of humans. Therefore, we measure the task performance in humans using the pass rate.

D-DDQN and relational DQN experimental setup: We conducted separate experiments on the D-DDQN and relational agents in the two settings: local generalization and broad generalization. For the local generalization evaluation, we run ten sampling agents that use the same DQN model to collect experiences. Each sampling agent runs on the randomly selected task for ten episodes. After the set of experiences is collected, the DQN model is trained for ten epochs with a batch size of 32. We train DQN until it either converges or reaches N update steps, where N is the number of training tasks per template divided by 5. Similar to ref. [12], for each batch, we sample 16 experiences in which a task is solved and 16 that failed. We train our agent on 80% of the tasks of the task template and evaluate on the rest of the tasks of the same template. We used the same training setting for all of the task templates. At the testing time, the agent runs on each of the testing tasks only once and selects the action that has the highest Q-value for a given state. For the broad generalization evaluation, we use the same training and testing setting as in the local generalization evaluation, except we train our agents on the tasks in the training templates in each scenario and test on the tasks from the testing templates.

Heuristic agents experimental setup: We conduct two experiments using the AIBIRDS heuristic agents. The first experiment is to evaluate the local and broad generalization capabilities, and the second is to evaluate the performance in the AIBIRDS competition game levels.

- Local and broad generalization setup: Due to the randomness in the heuristic agents, we allow them to have five attempts per task and calculate the task pass rate by averaging the result over these five attempts. For the local generalization setting, the agents were tested on the same 20% of the test tasks from each task template (1,500 tasks in total) as used for the D-DDQN evaluation. We report the local generalization performance by averaging the pass rates of all templates. For the broad generalization setting, the same testing templates as used for the D-DDQN evaluation were applied, and the within-scenario pass rate is calculated by averaging over all the tested templates within the scenario.
- AIBIRDS competition setup: We evaluate the AIBIRDS heuristic agents on 2021 AIBIRDS competition game levels to compare their performance in the competition game levels and the Phy-Q testbed tasks. We exclude the competition game levels with unrealistic effects as our focus in the testbed is scenarios with realistic physics. The game levels used for this evaluation are shown in Supplementary Sect. G. In the AIBIRDS competition, the agent with the highest score wins the competition. Therefore, in this experiment, we record the score and pass rate of the agents. The agents are allowed to have five attempts per game level to account for their randomness. Altogether, an agent had 40 plays.

Random agent experimental setup: The random agent was tested on the same testing tasks set from each task template. We run the random agent 50 times per task and report the average pass rate of these 50 attempts. The same as how we evaluate the heuristic agents, we further average the task performance within the same task template

and average the pass rate of all the templates to present the local generalization performance. For the broad generalization setting, the within-scenario pass rate is calculated by averaging over all the tested templates within the scenario.

Results and analysis

In this section, we first present and analyse the results obtained from our experiment with human players. Next, we present the results obtained from our experiments in measuring the local and broad generalization ability of agents and the Phy-Q score. We further analyse the results and discuss what we can derive from the experiments. We also discuss the results obtained from the heuristic agents in the 2021 AIBIRDS competition levels and the Phy-Q testbed tasks to show how the testbed can be used as a guide for the competition.

Human performance. Figure 3 presents the average pass rate, the pass rate the human players achieved within five attempts, the maximum number of attempts made and the total thinking time of human players for the 15 capabilities. The average pass rate is calculated as 100% if the player passes at the first attempt, whereas if the player passes at the fifth attempt, the pass rate is 20%. We record the thinking time of an attempt as the time between the task loading and the player making the first action. The total thinking time of a player is the sum of the thinking time of all their attempts. The number of attempts made and the total thinking time is scaled to 0–1 using min–max scaling in Fig. 3. Charts with the real values are available in Supplementary Sect. F.

Overall, human players passed almost all the tasks in each scenario within the five attempts. On average, they used 1.86 attempts per task and took 23.73 s to think per task. On average, the low number of attempts to pass the tasks shows that the dexterity required to solve the tasks when the strategy is determined is low. The average thinking time per task shows that humans have to think carefully about the strategic physical rule required to solve the task.

Humans have the longest thinking time for the tasks in the adequate timing scenario, but the average pass rate for these tasks is the second lowest. Similarly, the tasks from the non-greedy actions scenario have the lowest average pass rate with the highest number of attempts, while the thinking time is the second longest. This shows that figuring out the correct strategies for the tasks of these scenarios was difficult for humans. In the relative weight scenario, the pass rate achieved within five attempts is the lowest, but the thinking time is average for this scenario. This suggests that some humans take the action without carefully thinking about the strategy, and the strategy realized at a glance is not the correct strategy to solve the task. This also agrees with our observation that humans are overconfident in their wrong actions.

Local and broad generalization performance and Phy-Q score.
Local generalization performance: Figure 4b (first row) presents the average local generalization evaluation pass rate for all of our baseline agents. We also include the full results for the pass rate per agent per template in Supplementary Sect. D. The table shows that the four learning agents perform significantly better than their heuristic counterparts. While both the symbolic learning agents and both the image learning agents on average pass approximately 33% and 24% of the test levels, respectively, the previous champions in the AIBIRDS competition (Bambirds, Eagle's Wing and Datalab) pass around only half of the levels as compared with the learning agents, averaging 15%, 14% and 15%, respectively. This agrees with what is generally accepted that deep learning systems can perform a single narrow task much better than heuristic methods when enough densely sampled training data are available.

Broad generalization performance: Figure 4a presents the average pass rate of test templates of the broad generalization evaluation of all the baseline agents and human players. It is clear that the humans substantially outperform all the other agents, while all the agents have above-chance performance compared with the random agent. Heuristic agents achieved a better pass rate in the single force scenario (scenario 1) and multiple force scenario (scenario 2) as these two scenarios correspond to the essential ability needed to play Angry Birds, that is, shooting directly at pigs. It can be seen that the heuristic agents generally perform better if the physical scenario is covered in their built-in rules. For example, Datalab and Eagle's Wing have a built-in strategic physical rule to roll round objects, and they have the highest pass rate in scenario 3 (rolling) among all the agents. For scenario 4 (falling) and scenario 13 (clearing paths), Bambirds dominates the leaderboard of pass rate because it explicitly analyses spatial relationships between blocks and pigs and is the only heuristic agent with the 'prioritizing protective blocks' rule.

The second row in Fig. 4b shows the overall average pass rate for the broad generalization evaluation of the agents and humans. The heuristic agents' results were obtained in a similar way as applied for the local generalization evaluation, except we only consider the tasks from the testing task templates given to the learning agents. In contrast to the local generalization results, in this evaluation setup, the learning agents have worse results than all the heuristic agents. The D-DDQN-Symbolic and the D-DDQN-Image agents have an average pass rate of 12% and 10%, respectively, while Relational-Symbolic and Relational-Image have 14% and 9%, respectively. The champions in the AIBIRDS competition have almost twice the pass rate compared with the learning agents. This result further advocates the claim that deep learning agents often exploit spurious statistical patterns instead of learning in a meaningful and generalizable way as humans do[16,48,50,63,64].

Phy-Q score: As discussed in Section 5.1.2, the Phy-Q score of humans is set to 100 while that of the random agent is set to 0. A Phy-Q score above 100 indicates superhuman performance. Figure 5 shows the positions of agents and humans in the Phy-Q score distribution. Even though Eagle's Wing was the first in the broad generalization leaderboard (where Eagle's Wing scored 0.1142 while Bambirds scored 0.1022, even after removing the results of the first two scenarios), Bambirds took the lead in terms of the Phy-Q score, pushing Eagle's Wing into second place. This is because the Phy-Q score positions the agent with respect to human performance. Interestingly, the D-DDQN-Image and Relational-Symbolic agents achieved higher Phy-Q values compared with Datalab. Similar to the above reason, this is due to the positioning of the agents with respect to human performance. Moreover, the Phy-Q of the Pig Shooter is negative. This result is expected as the Pig Shooter only shoots at the pigs, thus exhibiting below-chance performance compared with the random agent. Overall, it can be seen that all the agents are far below the humans' Phy-Q score.

AIBIRDS competition performance. Extended Data Fig. 1 presents the results of the AIBIRDS heuristic agents in the AIBIRDS competition game levels. As can be seen from these results, the pass rate of the agents in competition game levels agree with the rank they achieved using the Phy-Q score. Eagle's Wing and Datalab achieved the same pass rate of 0.2. However, considering the total score, Eagle's Wing obtained 667,394 while Datalab obtained 634,435, pushing Eagle's Wing into second position. Overall, these results illustrate that the tasks in the Phy-Q testbed are representative of the tasks in the AIBIRDS competition.

On the basis of this result, we infer that the physical scenarios available in the Phy-Q tasks are also the scenarios that are commonly encountered in the AIBIRDS competition game levels. The within-scenario (broad generalization) evaluation that we have conducted can be used to identify an agent's ability in performing in those individual physical scenarios. Therefore, one can use the Phy-Q testbed to thoroughly analyse the physical reasoning capabilities of an AIBIRDS agent and determine where it falls short and improve on those capabilities. Additionally, the human performance results that we have established in the 15 scenarios facilitate the comparison of the agents' performance with

humans' performance, allowing us to set targets for agents to achieve human-level performance in those scenarios. Thus, the Phy-Q testbed and the evaluation settings we proposed in the testbed can be used to better evaluate AIBIRDS agents and guide them towards achieving human-level performance in the competition.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data collected from human players and baseline agents have been made available at https://github.com/phy-q/benchmark/tree/master/playdata.

Code availability

The testbed software and baseline agents' codes have been made available at https://github.com/phy-q/benchmark[65].

References

1. Davis, E. Physical reasoning. *New York University* https://cs.nyu.edu/~davise/papers/handbookKR.pdf (2006).
2. Valenza, E., Leo, I., Gava, L. & Simion, F. Perceptual completion in newborn human infants. *Child Dev.* **77**, 1810–1821 (2006).
3. Baillargeon, R. & DeVos, J. Object permanence in young infants: further evidence. *Child Dev.* **62**, 1227–1246 (1991).
4. Leslie, A. Spatiotemporal continuity and the perception of causality in infants. *Perception* **13**, 287–305 (1984).
5. Baillargeon, R., Needham, A. & Devos, J. The development of young infants' intuitions about support. *Early Dev. Parent.* **1**, 69–78 (1992).
6. Baillargeon, R. & Hanko-Summers, S. Is the top object adequately supported by the bottom object? Young infants' understanding of support relations. *Cogn. Dev.* **5**, 29–53 (1990).
7. Saxe, R. & Carey, S. The perception of causality in infancy. *Acta Psychol.* **123**, 144–165 (2006).
8. Day, R. H. & McKenzie, B. E. Perceptual shape constancy in early infancy. *Perception* **2**, 315–320 (1973).
9. Diezmann, C. M. & Watters, J. J. Identifying and supporting spatial intelligence in young children. *Contemp. Issues Early Child.* **1**, 299–313 (2000).
10. Cheke, L. G., Loissel, E. & Clayton, N. S. How do children solve Aesop's fable? *PLOS ONE* **7**, 1–12 (2012).
11. Emery, N. J. & Clayton, N. S. Tool use and physical cognition in birds and mammals. *Curr. Opin. Neurobiol.* **19**, 27–33 (2009).
12. Bakhtin, A., van der Maaten, L., Johnson, J., Gustafson, L. & Girshick, R. PHYRE: a new benchmark for physical reasoning. In *Proc. of Conference and Workshop on Neural Information Processing Systems* (eds Wallach H.M., et al.) (Curran Associates Inc., 2019).
13. Allen, K. R., Bakhtin, A., Smith, K., Tenenbaum, J. B. & van der Maaten, L. Ogre: an object-based generalization for reasoning environment. In *Proc. of NeurIPS Workshop on Object Representations for Learning and Reasoning* (2020).
14. Allen, K. R., Smith, K. A. & Tenenbaum, J. B. Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proc. Natl Acad. Sci. USA* **117**, 29302–29310 (2020).
15. Ahmed, O. et al. Causalworld: a robotic manipulation benchmark for causal structure and transfer learning. In *Proc. of 9th International Conference on Learning Representations* (OpenReview.net, 2021).
16. Chollet, F. On the measure of intelligence. Preprint at https://arxiv.org/abs/1911.01547 (2019).
17. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behav. Brain Sci.* **40**, e253 (2017).
18. Jo, J. & Bengio, Y. Measuring the tendency of cnns to learn surface statistical regularities. Preprint at https://arxiv.org/abs/1711.11561 (2017).
19. Marcus, G. Deep learning: a critical appraisal. Preprint at https://arxiv.org/abs/1801.00631 (2018).
20. Geirhos, R. et al. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).
21. Isaksen, A., Wallace, D., Finkelstein, A. & Nealen, A. Simulating strategy and dexterity for puzzle games. In *Proc. of 2017 IEEE Conference on Computational Intelligence and Games* 142–149, (IEEE, 2017).
22. McCloskey, M. Naive theories of motion. *Ment. Models* **14**, 299–324 (1983).
23. Smith, K. A. & Vul, E. Sources of uncertainty in intuitive physics. *Top. Cogn. Sci.* **5**, 185–199 (2013).
24. Riochet, R., et al. IntPhys 2019: A benchmark for visual intuitive physics understanding. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 5016–5025 (IEEE, 2022).
25. Yi*, K. et al. Clevrer: Collision events for video representation and reasoning. In *Proc. of 8th International Conference on Learning Representations* (OpenReview.net, 2020).
26. Baradel, F. & Neverova, N. & Mille, J. & Mori, G. & Wolf, C. Cophy: counterfactual learning of physical dynamics. In *Proc. of 8th International Conference on Learning Representations* (OpenReview.net, 2020).
27. Bear, D. M. et al. Physion: evaluating physical prediction from vision in humans and machines. In *Proc. of Conference and Workshop on Neural Information Processing Systems* (Curran Associates Inc., 2021).
28. Angry Birds game. *Rovio Enternainment* https://www.rovio.com/games/angry-birds (2022).
29. Renz, J., Ge, X., Stephenson, M. & Zhang, P. AI meets angry birds. *Nat. Mach. Intell.* https://doi.org/10.1038/s42256-019-0072-x (2019).
30. Grace, K., Salvatier, J., Dafoe, A., Zhang, B. & Evans, O. Viewpoint: when will AI exceed human performance? Evidence from AI experts. *J. Artif. Intell. Res.* **62**, 729–754 (2018).
31. Girdhar, R. & Ramanan, D. Cater: a diagnostic dataset for compositional actions and temporal reasoning. In *Proc. of 8th International Conference on Learning Representations* (2020).
32. James, S., Ma, Z., Arrojo, D. R. & Davison, A. J. Rlbench: the robot learning benchmark & learning environment. *IEEE Robot. Autom. Lett.* **5**, 3019–3026 (2020).
33. Archibald, C., Altman, A., Greenspan, M. & Shoham, Y. Computational Pool: a new challenge for game theory pragmatics. *AI Mag.* **31**, 33–41 (2010).
34. Prada, R., Lopes, P., Catarino, J., Quitério, J. & Melo, F. S. The geometry friends game AI competition. In *Proc. of 2015 IEEE Conference on Computational Intelligence and Games* 431–438, (IEEE, 2015).
35. Angry Birds AI competition. *AIBIRDS* http://aibirds.org/ (2022).
36. Krathwohl, D. R. A revision of Bloom's taxonomy: an overview. *Theory Pract.* **41**, 212–218 (2002).
37. Girdhar, R., Gustafson, L., Adcock, A. & van der Maaten, L. Forward prediction for physical reasoning. Preprint at https://arxiv.org/abs/2006.10734 (2020).
38. Ferreira, L. & Toledo, C. A search-based approach for generating angry birds levels. In *Proc. of 2014 IEEE Conference on Computational Intelligence and Games* 1–8, (IEEE, 2014).
39. Sanborn, A., Mansinghka, V. & Griffiths, T. Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychol. Rev.* **120**, 411–437 (2013).
40. Bliss, J. & Ogborn, J. Force and motion from the beginning. *Learn. Instr.* **4**, 7–25 (1994).

41. Wang, Z., Williamson, R. A. & Meltzoff, A. N. Preschool physics: using the invisible property of weight in causal reasoning tasks. *PLoS One* **13**, e0192054 (2018).
42. Baillargeon, R. & Devos, J. Object permanence in young infants: further evidence. *Child Dev.* **62**, 1227–1246 (1991).
43. Wang, S. Young infants reasoning about hidden objects: evidence from violation-of-expectation tasks with test trials only. *Cognition* **93**, 167–198 (2004).
44. Baillargeon, R., Li, J., Ng, W. & Yuan, S. in *Learning and the Infant Mind* (eds Woodward, A. & Needham, A.) 66–116 (Oxford Univ. Press, 2008).
45. Wilcox, T. & Chapa, C. Priming infants to attend to color and pattern information in an individuation task. *Cognition* **90**, 265–302 (2004).
46. Kemp, C. C., Edsinger, A. & Torres-Jara, E. Challenges for robot manipulation in human environments [grand challenges of robotics]. *IEEE Robot. Autom. Mag.* **14**, 20–29 (2007).
47. Knox, W., Glass, B., Love, B., Maddox, W. & Stone, P. How humans teach agents: a new experimental perspective. *Int. J. Soc. Robot.* **4**, 409–421 (2012).
48. Marcus, G. The next decade in AI: four steps towards robust artificial intelligence. Preprint at https://arxiv.org/abs/2002.06177 (2020).
49. Leslie, A. M. in *Mapping the Mind* (eds Hirschfeld L. A. & Gelman, S. A.) 119–148 (Cambridge Univ. Press, 1994).
50. Baillargeon, R. & DeJong, G. Explanation-based learning in infancy. *Psychon. Bull. Rev.* **24**, 1511–1526 (2017).
51. Stephenson, M. et al. The 2017 AIBIRDS level generation competition. *IEEE Trans. Games* **11**, 275–284 (2019).
52. Gamage, C., Pinto, V., Renz, J. & Stephenson, M. Deceptive level generation for angry birds. In *Proc. of 2021 IEEE Conference on Games*, (ed.) 1–8 (IEEE Press, 2021).
53. Wechsler, D. The measurement and appraisal of adult intelligence (Williams & Wilkins, 1958).
54. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
55. Wang, Z., Schaul, T., Hessel, M., Hasselt, H., Lanctot, M., & Freitas, N. Dueling network architectures for deep reinforcement learning. In *International Conference On Machine Learning* (eds Balcan M., Weinberger K.Q.) 1995–2003 (JMLR, 2016).
56. van Hasselt, H., Guez, A. & Silver, D. Deep reinforcement learning with double q-learning. In *Proc. of the Thirtieth AAAI Conference on Artificial Intelligence* (eds Schuurmans, D. & Wellman M. P.) (AAAI Press, 2016).
57. Zambaldi, V. F. et al. Relational deep reinforcement learning. Preprint at https://arxiv.org/abs/1806.01830 (2018).
58. Felix Haase, D. W. Bambird 2020. *GitHub* https://github.com/dwolter/BamBirds (2022).
59. Wang, T. J. AI Angry Birds eagle wing. *GitHub* https://github.com/heartyguy/AI-AngryBird-Eagle-Wing (2022).
60. Borovička, T., Špetlík, R. & Rymeš, K. Datalab Angry Birds AI. *AIBIRDS* http://aibirds.org/2014-papers/datalab-birds.pdf (2022).
61. Stephenson, M., Renz, J., Ge, X. & Zhang, P. The 2017 AIBIRDS competition. Preprint at https://arxiv.org/abs/1803.05156 (2018).
62. Kaiser, M., Proffitt, D. & McCloskey, M. The development of beliefs about falling objects. *Atten. Percept. Psychophys.* **38**, 533–539 (1985).
63. Bengio, Y. et al. A meta-transfer objective for learning to disentangle causal mechanisms. In *Proc. of ICLR 2020: Eighth International Conference on Learning Representations* (OpenReview.net. 2020).
64. Nie, Y. et al. Adversarial NLI: a new benchmark for natural language understanding. In *58th Annual Meeting of the Association for Computational Linguistics* (eds Jurafsky, D., Chai, J., Schluter, N., & Tetreault, J.R.) (Association for Computational Linguistics, 2020).
65. Xue, C., Pinto, V. & Gamage, C. Phy-Q – a testbed for physical reasoning – code repository. *Zenodo* https://doi.org/10.5281/zenodo.6933441 (2022).

Acknowledgements

This work is supported by the Defense Advanced Research Projects Agency and the Army Research Office and was accomplished under cooperative agreement no. W911NF-20-2-0002 (received by J.R. and P.Z.). We thank all the volunteers who played the game and the anonymous reviewers for their constructive input. We thank A. Yang for fruitful discussions.

Author contributions

C.X., V.P., C.G. and J.R. conceived this study. C.X., V.P. and C.G. wrote the manuscript. C.X. implemented learning agents. V.P. designed task templates and analysed the data. C.G. designed task templates and developed the task generator. C.X. conducted the experiments on learning agents. C.X. and E.N. wrote the section on learning agents. C.G. and V.P. conducted experiments on non-learning agents and conducted the AIBIRDS competition evaluation. C.X. and P.Z. adapted the Angry Birds framework to the testbed. J.R. provided feedback and supervision.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-022-00583-4.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-022-00583-4.

Correspondence and requests for materials should be addressed to Cheng Xue, Vimukthini Pinto or Chathura Gamage.

Peer review information *Nature Machine Intelligence* thanks Zoe Falomir Llansola and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s42256-022-00583-4

Rank from Phy-Q score	1	2	3	4
Agent name	Bambirds	Eagle's Wing	Datalab	Pig Shooter
Competition performance	0.2727	0.2000	0.2000	0.0000
Phy-Q score	14	12	7	-2

Extended Data Fig. 1 | Comparison of AIBIRDS competition performance with Phy-Q score of heuristic agents. Results of the AIBIRDS heuristic agents in the AIBIRDS competition game levels. The competition performance (pass rate) of the agents in competition game levels agree to the rank they achieved using the Phy-Q score.

nature research

Corresponding author(s): Chathura Gamage

Last updated by author(s): Oct 7, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The testbed software that includes data collection codes has been made available at https://github.com/phy-q/benchmark
Data analysis	Custom codes written in Python 3 for data visualization can be made available upon request.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data collected from human players and baseline agents have been made available at https://github.com/phy-q/benchmark/tree/master/playdata.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We conducted two experiments. The first experiment using AI agents and the other experiment using human players playing Angry Birds game levels. 1) AI agents: 9 AI agents were used and quantitative data on scorers and pass/fail status were recorded. 2) Human players: 20 human players participated and quantitative data on scorers, pass/fail, and time taken to solve were recorded.
Research sample	1) AI agents: four of our best performing learning agents, four heuristic-based agents, and a random agent. Learning agents are with varying input types and architectures, and heuristic agents are with different physical reasoning strategies. Rational for selecting these agents was that the heuristic agents represent varying heuristics and learning agents are general reinforcement agents in AI research. 2) Human players: Voluntary participants in Australia, representing both males and females and age ranging from 18-35. Rational for selecting these participants is that generally these game levels are easily solvable by humans and no special consideration is required.
Sampling strategy	1) AI agents: Data is collected from the two main types of agents, learning agents and heuristic agents. Learning agents are with varying input types and architectures, and heuristic agents are with different physical reasoning strategies. Sample size calculations was not conducted for agents as we used agents that represent different heuristics and learning architectures. 2) Human players: Convenience sampling method is used after sending out emails/ posters in social media for voluntary participation. Sample size was sufficient as the game levels could easily be solved by humans. Human players do not need special requirements to participate in this study. According to the sample size calculation conducted using our prior knowledge in AIBIRDS competitions (human vs machine challenge) and after playing these game levels among researchers, the pass rate standard deviation was around 0.1, and therefore to enable 5% standard error at 95% level of confidence 15 samples were sufficient. Therefore, we used data from the 20 players who volunteered.
Data collection	1) AI agents: Data was collected under the experimental settings presented in the paper. The game scores and pass/fail status and time taken to solve is recorded from the testbed software. 2) Human players: Data was collected from voluntary participants after sending out emails/ posters in social media. The game scores and pass/fail status and time taken to solve is recorded from the testbed software. Only the researchers were present in the room when participants were playing to solve any technical difficulties during the experiment. However, data does not record details of the participant. Therefore researchers were not aware of the participant details from the recorded data.
Timing	1) AI agents: 01/03/2021 - 31/12/2021 2) Human players: 01/03/2021 - 01/08/2021
Data exclusions	1) AI agents: Some learning agents that we obtained data from are excluded due to poor performance. Explained in the paper. 2) Human players: Data from one player was excluded as the data was not recorded correctly.
Non-participation	No participant dropped out
Randomization	Participants were not allocated into groups as it is not applicable to our study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above
Recruitment	Convenience sampling method is used after sending out emails/ posters in social media for voluntary participation. Voluntary participants were in Australia, representing both males and females and age ranging from 18-35. The participant selection will not impact results as the game levels that humans' played require simple physical reasoning capabilities. No remuneration was provided for the participants.
Ethics oversight	Informed consent was obtained from all the participants to use their palydata. Ethics approval was obtained from the Australian National University committee on human ethics under protocol 2021/293

Note that full information on the approval of the study protocol must also be provided in the manuscript.