Evaluating Object Permanence in Embodied Agents using the Animal-AI Environment

Konstantinos Voudouris[1,2], Niall Donnelly[3], Danaja Rutar[1], Ryan Burnell[1], John Burden[1], José Hernández-Orallo[1,4] and Lucy G. Cheke[1,2]

[1]Leverhulme Centre for the Future of Intelligence, Cambridge, UK
[2]Department of Psychology, University of Cambridge, UK
[3]The College of Engineering, Mathematics, and Physical Sciences, University of Exeter, UK
[4]VRAIN, Universitat Politècnica de València, Spain

Abstract

Object permanence, the understanding and belief that objects continue to exist even when they are not directly observable, is important for any agent interacting with the world. Psychologists have been studying object permanence in animals for at least 50 years, and in humans for almost 50 more. In this paper, we apply the methodologies from psychology and cognitive science to present a novel testbed for evaluating whether artificial agents have object permanence. Built in the Animal-AI environment, *Object-Permanence In Animal-Ai: GEneralisable Test Suites* (O-PIAAGETS) improves on other benchmarks for assessing object permanence in terms of both size and validity. We discuss the layout of O-PIAAGETS and how it can be used to robustly evaluate OP in embodied agents.

Keywords

Object Permanence, AI Evaluation, Embodied Agents, Animal-AI Environment, Developmental Psychology, Comparative Cognition

1. Introduction

Object Permanence (OP) is the understanding and belief that objects continue to exist even when they are not directly observable. In behavioural terms, an agent has OP when they behave as though objects continue to exist when they cannot see them. Human adults use OP to reason about how objects behave and interact in the external world. Credited as the first to empirically investigate this capability, Jean Piaget observed how infants develop the tendency to search for objects that became occluded [1]. Piaget's insights have been extended considerably by developmental and comparative psychologists, usually in the visual modality [2, 3, 4], although OP is an amodal phenomenon [5].

Humans and some animals appear to understand that objects continue to exist independently of them, with the same properties. However, when an object reappears, what makes us reidentify this as the same object as before? Object reidentification has been studied in visual cognition research with adults [6, 7, 8] and primates [9], and in developmental psychology with infants [10, 11, 12]. The relation between object reidentification and OP is manifest: when an object passes out of view, we believe that it continues to exist. When it passes back into view, we use knowledge about objects to determine whether this is the same object we saw previously. Here, we use OP to mean both classical OP and object reidentification.

OP has proven difficult to build into AI systems. Deep Reinforcement Learning systems perform significantly worse than human children when solving problems involving OP [13]. Tracking objects under partial occlusion appears to be difficult for modern computer vision methods [14]. The need for AI agents with robust OP is important for creating trustworthy embodied AI such as self-driving cars. Furthermore, robust object tracking under occlusion would have many applications in the field of robotics. However, the methods for evaluating whether an agent has OP suffer from a lack of precision, reliability, and validity. Developmental and comparative psychologists have been investigating OP in biological agents for around a century, developing many experimental paradigms along the way. Until now, AI research has not applied these methods to AI evaluation [15]. In this paper, we outline a new test battery, built in the Animal-AI Environment [16] for evaluating whether embodied artificial agents have OP: the *Object-Permanence in Animal-Ai: GEneralisable Test Suites. O-PIAAGETS* is a novel attempt to use experiments designed for investigating whether biological agents have OP for AI evaluation. First, we examine why OP is a challenge for AI research. Second, we critically review existing OP testbeds. Third,

EBeM'22: Workshop on AI Evaluation Beyond Metrics, July 25, 2022, Vienna, Austria
kv301@cam.ac.uk (K. Voudouris); mail.niall.donnelly@gmail.com (N. Donnelly); dr571@cam.ac.uk (D. Rutar); rb967@cam.ac.uk (R. Burnell); jjb205@cam.ac.uk (J. Burden); jorallo@upv.es (J. Hernández-Orallo); lgc23@cam.ac.uk (L. G. Cheke)
0000-0001-8453-3557 (K. Voudouris)

we outline the structure of the test battery and how it can be used to robustly investigate whether agents have OP. Finally, we discuss how O-PIAAGETS can be used for evaluation and how it improves on existing testbeds in the field.

2. Background and Motivations

2.1. The Logical Problem of OP

OP may appear to be a trivial capacity for an agent to have. The agent must simply understand that objects continue to exist when they are not directly observable. Indeed, Renée Baillargeon and colleagues [17] hypothesise that children are born with a *Principle of Persistence*, which states exactly this [18, 19]. Why, then, can't we endow AI systems with such a principle, bias, or heuristic? Can't we simply tell an agent that objects continue existing when they are occluded? Fields [20, 19] has discussed how the notion of a Principle of Persistence is untenable, due to the *Frame Problem* (FP).

The FP implies that endowing an agent, biological or artificial, with a principle of persistence is not trivial. It cannot be overcome with a representation as simple as *objects continue to exist even when they aren't observable.* In its raw form, the FP demonstrates that when logically describing the effects of particular actions on objects in a domain, we must also describe *ad nauseam* all the non-effects of those actions on those objects. As Fields [19] says, it amounts to having to describe everything that *doesn't* change in the universe as a result of turning off the fridge (p. 443). In a domain where objects have certain properties that can change over time, as in all real-world scenarios, the FP implies that we can't simply say that the *objects* stay the same over time, without describing which properties remain unchanged and when [21].

When an agent can observe everything in a domain, and re-update what has and has not changed at every timestep, the FP rarely raises any issues. However, when objects become occluded, it becomes important to track which properties of those objects do and do not change and when, in order to identify other objects as identical or different. For example, imagine a lion watching a small antelope pass behind some bushes and then seeing a large antelope emerge at the other side. It becomes useful to know that antelope don't change size over such time periods, and therefore the smaller antelope continues to exist because of the persistence of its size (and other) properties. It also becomes useful to know that the antelope doesn't change when the lion changes their perspective, or occludes the antelope through its own actions, an analogue of the Simultaneous Location and Mapping (SLAM) problem in robotics [22]. Overcoming the FP either requires sophisticated deductive techniques [21], or robust inductive and abductive learning heuristics and biases [20, 19, 8]. It is therefore not as simple as imputing a Principle of Persistence to build AI systems with OP.

2.2. Existing Evaluation Methods for OP in AI

AI researchers, particularly those working on computer vision, embodied agents, and robotics, are interested in building AI systems capable of robustly reasoning about visual scenes, in a similar way to how humans and animals do. Researchers have built several evaluation frameworks for assessing whether embodied artificial agents and computer vision systems have OP.

Lampinen *et al.* [23] built OP tasks in a 3D Unity environment. Here, the agent was fixed as it watched three boxes. Periodically, objects would leap out of the three boxes, simultaneously or sequentially with or without a refractory time lag. The agent would then be turned away from the boxes, released, and asked to go to the box with a particular object. If it chose the correct box, it was rewarded, similar to tasks used with human infants [24] and non-human primates [4]. Crosby *et al.* [16] developed a series of 90 OP tests as part of the *Animal-AI Testbed and Olympics*, inspired and directly developed by developmental and comparative psychology. Some work has been done comparing embodied deep reinforcement learning agents to humans on these tasks. Children aged 6-10, with limited training, significantly outperformed Deep Reinforcement Learning systems on the OP tasks in the Animal-AI Testbed [13], indicating there is room to improve these systems until they reach human-level performance. Leibo *et al.* [25] developed *Psychlab* for probing psychophysical phenomena in Deep Reinforcement Learning systems using cognitive science methods and qualitatively comparing performance with human participants, but they did not investigate OP.

Having OP is not only applicable to embodied agents, but also to passive computer vision systems engaged in object tracking. The Localisation Annotations Compositional Actions and TEmporal Reasoning (LA-CATER) dataset [26] is prominent in computer vision research. LA-CATER contains 14000 video scenes where objects can move in three dimensions, contain, and carry each other. Several tasks in this dataset happen to behave similarly to OP experiments used in psychology. For example, one task involves an object being occluded by one of three identical 'cups'; once occluded, the cups are moved relative to each other. This bears resemblance to the cup-tasks used in the Primate Cognition Test Battery [4] (see Figure 3) or in the Užgiris and Hunt [24] test battery for infants. Other benchmark datasets include ParallelDomain (PD) and KITTI [27]. PD is a synthetic dataset designed to test occlusions in driving scenarios.

It contains 210 photo-realistic driving scenarios in city environments, from 3 camera angles, creating a dataset of 630. KITTI [28] has 21 labelled videos of real-world city scenes, in which cars, pedestrians, and other objects pass behind each other and become partially or fully occluded, a small fraction of the total KITTI dataset [27].

Piloto *et al.* [29] directly applied a measurement framework innovated in developmental psychology to probe physics knowledge in artificial systems, including OP. Violation of Expectation has been used by the neo-Piagetian school of developmental psychology [3], investigating infants' knowledge about the world by determining when they are surprised to see something, violating their expectations. For example, infants at about 4.5 months tend to show surprise (by looking more) if an object appears to change size whilst occluded [10, 11]. Piloto *et al.* procedurally generated 28 3-second videos that emulated a small subset of these studies, and used Kullback-Leibler divergence as the AI equivalent of looking time. They demonstrated the utility of this technique for probing physical knowledge in computer vision systems.

In both computer vision and embodied AI, several methods for detecting when agents have OP have been proposed. However, with the exception of the work of Piloto *et al.* [29] and Lampinen *et al.* [23] at DeepMind and the Animal-AI Testbed and Olympics, little attention has been paid to systematically applying the methodologies of psychology to try to understand and evaluate OP in artificial agents.

2.3. Problems with Current Evaluation Frameworks for OP

Two main problems exist with the methods for evaluating whether AI has OP. The first problem is that most of these benchmarks and testbeds use independent-and-identically-distributed (i.i.d.) test data, meaning testing data is drawn from the same distribution as training data. This especially applies to LA-CATER, PD, and KITTI. The second problem is a lack of internal validity. Sufficient controls to eliminate alternative explanations for certain behaviours are often lacking.

The problem with i.i.d. testing data is that it is in principle impossible to distinguish between an agent that has OP and one using problem-irrelevant shortcuts to maximise reward, appearing *as if* they have OP. This means that even if we had an agent that genuinely had OP, our evaluation methods limit how certain we can be of that. Geirhos *et al.* [30] argue that an effective measure against this is to test AIs on *out-of-distribution* (o.o.d.) test data, where training data and test data are drawn from different (but meaningfully related) distributions. This is related to the notion of *transfer tasks* in developmental and comparative psychology. The move from i.i.d to o.o.d. testing is still not mainstream, but is gaining prominence [31, 32, 33]. LA-CATER and the procedurally generated test sets mentioned earlier were generated according to a series of rules, with training, validation, and test sets divided arbitrarily. The PD and KITTI datasets were generated and collected non-procedurally, but again, the distinction between training and test sets is often arbitrary [27].

Moving from i.i.d. to o.o.d. test data promotes robustness in AI systems. Developing a testbed for OP in which training and test data are kept distinct means that we can be more certain that AI systems have OP if they perform successfully, rather than overfitting to the data distribution. This means we can evaluate whether an AI has an *ability* corresponding to OP, rather than a propensity for solving some distribution of tasks that require it[1] [35, 36].

O.o.d. testing enables researchers to have grounds to say they are testing for the presence of abilities. However, selecting a test distribution must be guided by some principle that tells us why the training and test distributions are *meaningfully related.* This takes us to the second problem for OP evaluation in AI: that testing lacks internal validity. Developmental and comparative psychologists have developed numerous experimental designs to test for the presence of cognitive abilities in biological agents, introducing numerous controls to eliminate alternative explanations. As a point of reference, let's take the classic A-not-B paradigm for testing OP. Participants are presented with an object of interest that is hidden for several trials at location A. In the AI context, this amounts to a training distribution around location A (with variance corresponding to minor differences between trials). To test the participant to find an object of interest at A, true OP understanding as an explanation is conflated with other explanations in terms of memorising spatial location or returning to a previously rewarding location as infants under 9 months, and many animals, do [1, 37]. To eliminate (some of) these explanations, in the test condition, participants are faced with an object hidden at location B. The testing distribution now includes objects hidden at B, and the relation between the two is meaningful in the context of OP, because an agent needs OP to solve the task. The logic is that one would only perform well on training (A-only) and testing (B-only) if one had OP. Of course, there are further alternative explanations for correct search at locations A and B, such as simply searching where the experimenter's hand has just been [24]. So internal validity tends to increase the

[1] For example, an anonymous reviewer pointed out that DeepMind's FTW agent [34] arguably has object permanence, since it can successfully fight players who duck for cover in a 3D capture-the-flag game. While this is certainly evidence for OP in an artificial agent, it remains speculative for now, since FTW has not yet been tested on an internally valid, out-of-distribution test set like O-PIAAGETS - although O-PIAAGETS itself is not yet developed for testing such a multi-agent system.

more diversity in training and test data there is, as they become mutually controlling.

Psychologically-inspired testbeds for evaluating OP in AI systems, such as Piloto *et al.* [29], Lampinen *et al.* [23] and Crosby *et al.* [16], remain small and so internal validity remains relatively low. The confluence of low internal validity in some testbeds and the lack of o.o.d. testing means that even if an AI system genuinely has OP, our evaluation frameworks and metrics are not internally valid enough to show this. In this paper, we propose a novel large testbed for conducting o.o.d. testing with high internal validity.

3. Introducing O-PIAAGETS

In the previous section, we established three things:

1. OP poses a challenging logical problem. It is not trivial for an agent to have OP.
2. Computer vision and embodied agent results suggest that trained computer architectures solve tasks involving OP at a level significantly lower than that of humans.
3. Current benchmarks and testbeds for evaluating whether AI systems possess OP have limitations such that even if an AI system had OP, we might not be able to tell with reasonable certainty.

Our novel testbed, *Object-Permanence in Animal-Ai: GEneralisable Test Suites* (O-PIAAGETS), overcomes the limitations of other testbeds by applying an out-of-distribution testing framework on a large internally valid set of tasks adapted from comparative and developmental psychology and visual cognition research.

O-PIAAGETS uses the Animal-AI Environment to generate individual tasks for training and testing, based on theoretical and empirical findings in the psychology literature. The testbed has an internal structure in which certain tasks are designed to test certain aspects of OP understanding. There is also a tailored training curriculum to ensure out-of-distribution testing, and more direct comparison between biological and artificial machines. This work complements and extend the work of Piloto *et al.* [29], Lampinen *et al.* [23], Crosby *et al.* [16], and Voudouris *et al.* [13].

3.1. The Animal-AI Environment

The Animal-AI Environment [16] is a 3D world with Euclidean geometry and Newtonian physics built in Unity [38]. The environment contains several objects, a single agent, and a finite number of actions it can perform (move and rotate in x-z plane). The agent is situated in a square arena. The arena can be populated with appetitive (green and yellow spheres) and aversive stimuli (red spheres and red *lava* zones), pink ramps, and transparent and opaque[2] blocks and tunnels (see Figure 1). These objects can be any size, constrained only by the dimensions of the arena and the fact that two objects can't occupy the same location (apart from lava zones). The lights can also be switched on or off for preset periods of time, removing all visual information (see Figure 7 for an example).

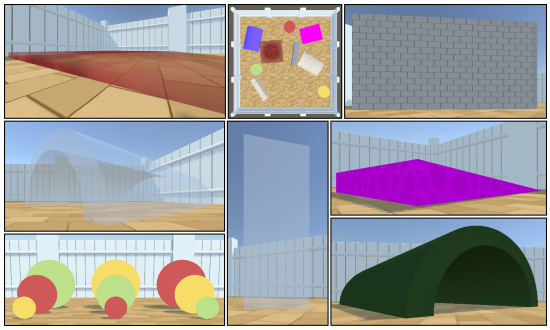

Figure 1: The Animal-AI Environment. A bird's eye view of the arena is given top centre. The various objects that can populate it are shown and described in the text.

Points are gained and lost through contact with rewards of differing size and significance, and punishments of differing severity. Obtaining a yellow sphere increases points. Obtaining a green sphere also increases points and is episode-ending. Obtaining a red sphere decreases points and is episode ending, as does touching red *lava* zones. All spheres can be stationary or in motion through all three dimensions. Points start at 0 and decrease linearly with each timestep over an episode, creating time pressure and therefore motivation for fast and decisive action.

3.2. Structure of O-PIAAGETS

O-PIAAGETS adapts some tasks from the open-source Animal-AI Testbed, but mostly includes new ones. It currently contains 5000 tasks, divided into four **suites**, although it continues to expand as new features are released for the Animal-AI Environment. There are three suites which test different aspects of OP and one suite which contains controls for non-OP based explanations. The three suites were motivated *a priori* by Brian Scholl's [7] exposition of OP research. Here, Scholl reviews work on OP from across research in psychology, neuroscience, and philosophy, arguing that OP appears to be underpinned by three key cognitive strategies. Humans appear to reason about objects under occlusion as (a) existing on continuous spatiotemporal trajectories, (b) maintaining certain properties, such as size, but not necessarily oth-

[2] Of any RGB colour combination

ers, such as colour, and (c) existing as unified cohesive wholes. O-PIAAGETS therefore contains a **Spatiotemporal Continuity suite**, a **Persistence Through Property Change suite**, and a **Cohesion suite**. Each suite is subdivided based on the psychology and AI research into **sub-suites** testing different aspects of the suites. Those sub-suites are subdivided into **experimental paradigms** from the psychology literature. To maintain high internal validity, each sub-suite has at least 3 experimental paradigms. These are further divided into **tasks** which are specific instantiations of an experimental paradigm as used in specific experiments. These tasks are composed of **instances**, that are procedurally generated variations of the global structure of the task, such as right and left versions or versions with goals of different sizes or in different positions. Finally, these instances are composed of **variants**, which are procedurally generated variations of the local structure of instances, with changes to the colours of walls and the starting orientation of the agent. In every test below, the objective is simple: maximise reward. This involves obtaining yellow and green rewards while avoiding red rewards and 'lava', as quickly as possible.

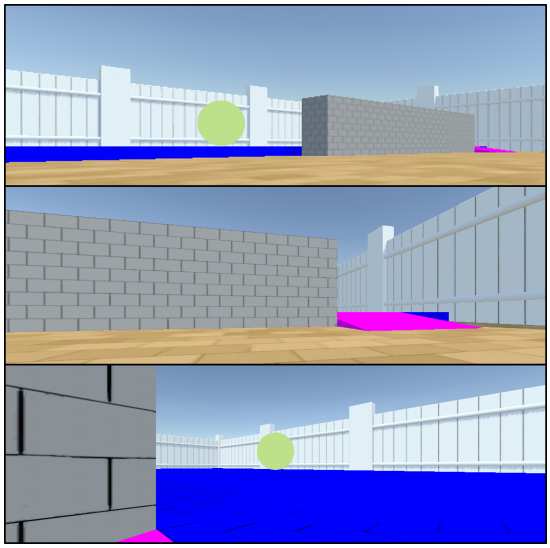

Figure 2: An example detour task. To get to the reward, the agent must navigate around the wall and up the ramp. This means that the goal will go out of view through the movement of the agent.

3.2.1. Spatiotemporal Continuity

The Spatiotemporal Continuity suite examines how participants reason about objects as persisting in the same spatiotemporal region, given initial starting velocities and other interacting objects. This suite is divided into two sub-suites: **egocentric OP** and **allocentric OP**.

Egocentric OP pertains to reasoning about objects persisting when they pass out of view *through the actions of the agent*. This allows us to evaluate how well an agent can learn about the identity and location of objects in a region while also moving around that region, a variant of the SLAM problem in robotics. An example of an egocentric OP task is a detour task where a goal is observable but inaccessible behind an obstacle. The way to obtain it is to detour around the obstacle such that the goal is temporarily left out of sight. The logic here is that one would only execute the detouring behaviour if one believed that the goal would still exist when one has finished detouring (see Figure 2).

Allocentric OP pertains to reasoning about objects that pass out of view not because of the actions of the agent, but because they become occluded by another object. The *Cup Task* in Figure 3 is an example [4]. A goal is hidden inside a 'cup' for some time. To succeed, the agent would need to search in the correct 'cup'.

The Tunnel Effect paradigm is a second example. An object passes behind an occluder, and another emerges some time later. If the second object appears as a human would expect it to, given the first object's trajectory, we perceive it as though the first object has gone through a tunnel and come out of the other side (Burke, 1952). However, if the second object appears later than expected or on a different trajectory, we do not identify it as the same object [6] (see Figure 4). The Tunnel Effect tasks enable us to probe where OP 'breaks' in the agent in question, and how it compares to human performance. In the Tunnel Effect tasks here and below, the agent is frozen until they have observed the whole scene, so they don't miss the important occlusion events we are probing, eliminating a potential explanation for why an agent failed on these tasks.

In line with developments of the Animal-AI Environment, we will introduce allocentric OP tasks involving containment in stationary and moving containers, as done in the LA-CATER and Lampinen *et al.* [23] testbeds discussed earlier.

3.2.2. Persistence Through Property Change

The second suite of tests extends the Tunnel Effect tasks, investigating which properties of an object must change under occlusion for the post-occlusion object and pre-occlusion to be classified as different. Scholl [7] reports that the Tunnel Effect is not disrupted by colour or shape change, only size changes, and the spatiotemporal changes in the previous sub-suite [39, 9, 6, 40]. Wilcox and Baillargeon [10] present evidence that the Tunnel Effect is disrupted by colour, shape, and texture changes. O-PIAAGETS permits more control over the timing and

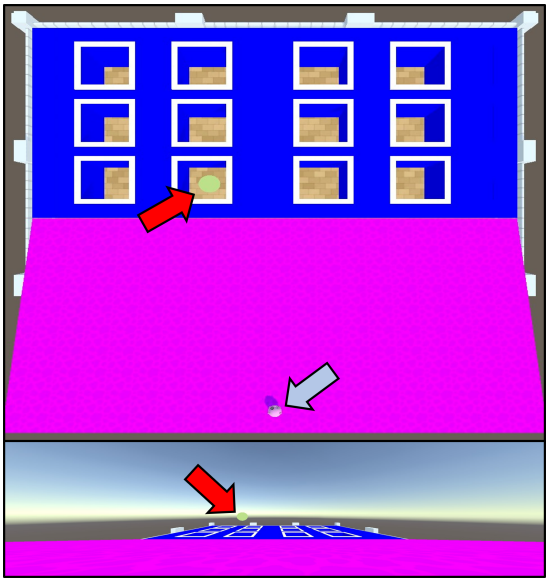

Figure 3: An example allocentric task inspired by the Primate Cognition Test Battery [4]. Red arrows indicate goals, pale arrow indicates agent.

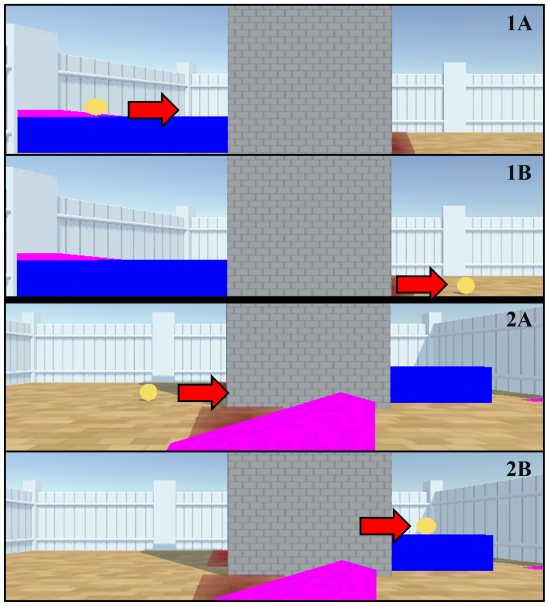

Figure 4: A Tunnel Effect task. Humans would perceive the object in 1A as the same as the object in 1B, but the object in 2A as different to the object in 2B, because of the impossible trajectory.

nature of changes, so can be used for empirical study with humans to investigate these inconsistent results, as well as to analyse under what conditions OP breaks in AI agents.

Currently, this suite only contains one sub-suite, testing the Tunnel Effect with apparent size change under occlusion. However, in line with developments in the Animal-AI Environment, we are building sub-suites for apparent shape, colour, and pattern change. An example of a task in which size appears to change is provided in Figure 5. The post-occlusion object is smaller than the pre-occlusion object, so the agent must search for two distinct objects, not just the visible one.

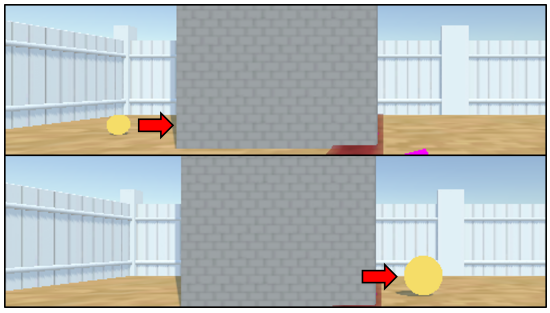

Figure 5: A tunnel effect task manipulating the property of size.

3.2.3. Object Cohesion

Scholl [7] argues that OP is not disrupted in human adults when the contours are partially or completely removed from a visual object representation, so long as size does not appear to change. Humans, and many animals, assume that an object is of constant size [41], even when contour information is partially occluded or completely removed and replaced with point lights [42, 43]. Currently, this suite contains only one sub-suite, examining size constancy under partial occlusion. An example of this is the *aperture task* in Figure 6, innovated for O-PIAAGETS based on discussion in Scholl [7]. Agents watch a large green goal roll behind a wall with a small hole in it. It is then released and given the choice to turn left and seek out the large goal behind the wall, or turn right and seek out the entirely visible smaller goal. The smaller goal is larger than the hole in the wall, so agents that compare the number of green pixels visible at one time without understanding that size remains constant under (brief) occlusion will make the wrong choice.

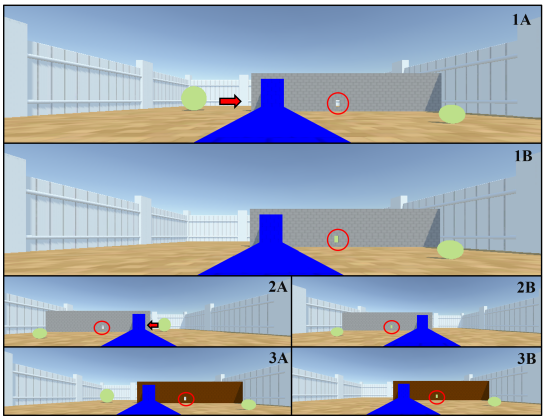

Figure 6: The Aperture Task. A and B are before and after partial occlusion. Parts 1 and 3 are variants of the same instance, with differing wall colours. Part 2 is a different instance of the aperture task, the mirror image of Part 1.

3.3. Increasing Internal Validity

3.3.1. Control Suite

The fourth suite is a set of control tests that serve to determine whether agents can solve tests not measuring OP. There are two sub-suites here. The first is an introduction to the environment, introducing basic controls and the objects present in the environment. These tasks allow an agent, human or artificial, to learn which objects increase reward and which objects decrease reward, and which objects are inert. Agents that fail some or all the tasks in the above three suites might not be failing because they lack OP, but because they do not, for example, navigate towards green rewards or away from red lava, or understand the utility of ramps for movement in the up-down plane. The second sub-suite contains further control tests for the OP tasks in the previous three suites. These are tests that do not require OP to be solved, but introduce the kind of landscapes and choices an agent might have to make. This means we can determine whether poor performance on the OP tasks was a result of a lack of OP, or a lack of understanding of the landscapes those tests took place in. Since every task in the test battery will require other abilities distinct from OP, these controls allow developers to check whether errors are a result of a lack of OP or a lack of some other ability. These tasks can either be used in training or for further testing. An example would be Figure 2 but without a grey wall and with a pink ramp the length of the blue platform. This increases internal validity, because if agents performs well on the control task, but not well on the equivalent OP task, then we have reason to believe that they lack OP. If they perform well on both, we have reason to believe that they possess OP. If they perform poorly on both, then there is some issue with understanding the environment or how to interact with it. If they perform well on the OP tasks but not the controls, then we have counter-intuitive evidence that OP can be decoupled from other abilities required to solve tasks in the environment.

3.3.2. Paradigms, Instances, Variants

Within the test suites themselves, two measures have been taken to increase internal validity. First, each task has several instances and variants. We have procedurally generated many versions of the same task that are mirror images of each other (left/right versions), have rewards and goals in different positions, or use different kinds of occluders. This counterbalanced design allows us to detect when agents are solving tasks through problem-irrelevant shortcuts. For example, in the aperture task in Figure 6.1, an agent with a bias to turning left might appear to succeed, but would not succeed at the instance which is a mirror image of this task, as in 6.2. These instances have many variants, changing the colour (often randomly) and initial orientation of the agent, as seen in 6.3. This allows us to control for policies such as *search behind the grey obstacle*, which may be successful in some tasks but does not indicate OP.

Second, the inclusion of several experimental paradigms in each sub-suite means they are mutually controlling. The philosophy of science tells us that no single experiment would be able to diagnose the presence or absence of OP [44, 45], because there are always alternative explanations that could be appealed to. Using several distinct experimental paradigms means that they can control for each other and help eliminate these alternative explanations. The cup task in Figure 3 could be solved by a policy of navigating to where the reward was last seen [46], which is not necessarily the same as understanding that the object continues to exist even though the agent can't see it. An adaptation of Chiandetti and Vallortigara's [47] paradigm controls for this (see Figure 7). Here, the agent watches a reward roll away from across lava. Then the lights go out, removing visual information for a short period. When the lights go back on, the goal is not visible. However, there is only one place it can be. Going to where the reward was last seen would end in failure, by touching lava, and the position of the goal before the lights out provides no cue as to whether the agent should go right or left. The use of several experimental paradigms in each sub-suite has the effect of reducing the likelihood of confounds that we have not foreseen.

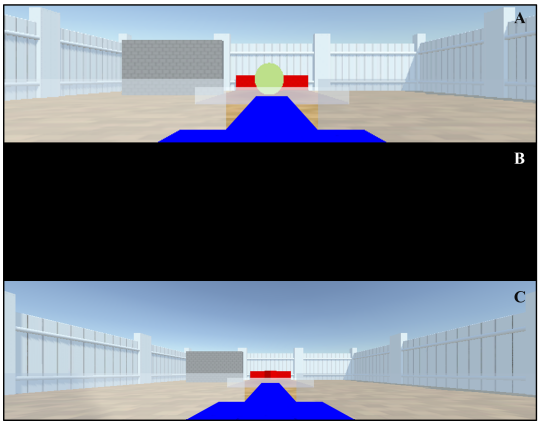

Figure 7: A task inspired by Chiandetti and Vallortigara's [47] study with day-old chicks.

4. Evaluating OP using O-PIAAGETS

4.1. Out-of-Distribution Testing

O-PIAAGETS facilitates out-of-distribution testing by providing a tailored training set using the control suite, and a separate test set using the three test suites. The control suite contains tasks where the positions and orientations of objects is specified and tasks where those positions are randomly generated, providing in principle a very large amount of training data that is on a different distribution to the test data.

4.2. Measurement Layouts

Each variant in O-PIAAGETS is tagged with its position in the test battery (i.e., what suite, sub-suite, experimental paradigm, etc., it is a member of) as well as features such as goal sizes, the abilities an agent might require in addition to OP to solve it, and the other variants it controls for. This leads to an incredibly rich dataset for evaluating agents beyond merely aggregating their score or success across the test suites. For example, developers can explore how relevant and irrelevant features of the tests, such as goal size, occluder colour, or right/left variants, correlate with performance [48], and use this to evaluate whether an agent has OP or is using other policies to solve OP tasks. For example, assuming any agent interacting with O-PIAAGETS will make errors, including humans [13], it is important to evaluate how those errors are distributed. By hypothesis, an agent with OP will produce random error, uncorrelated with experimental paradigms, goal sizes, or the colours of occluders.

5. Future Directions and Conclusions

Using O-PIAAGETS, developers can robustly evaluate whether artificial embodied agents have OP using the methodologies of cognitive science. It improves on other benchmarks and testbeds in the field both in terms of its size, internal validity, and ability to detect the presence of robust and generalisable OP in artificial systems. O-PIAAGETS is going through the final stages of development for general release of Version 1.0, including around 5000 tasks using the current Animal-AI Version 3.0.1. After validation with human participants and the development of baseline agents to characterise state-of-the-art performance in O-PIAAGETS, it will be expanded to include containment tasks, point lights, and shape, colour, and pattern changes. In its final form, O-PIAAGETS will provide a comprehensive and robust evaluation framework for assessing OP in artificial agents.

Acknowledgments

We thank the anonymous reviewers for their comments. This work was funded by the Future of Life Institute, FLI, under grant RFP2-152, EU's Horizon 2020 research and innovation programme under grant agreement No. 952215 (TAILOR), US DARPA HR00112120007 (RECoG-AI), and an ESRC DTP scholarship to KV, ES/P000738/1.

References

[1] J. Piaget, The Origins of Intelligence In The Child, Routledge & Kegan Paul, Ltd., 1923.

[2] R. Baillargeon, E. S. Spelke, S. Wasserman, Object permanence in five-month-old infants, Cognition 20 (1985) 191–208. URL: https://linkinghub.elsevier.com/retrieve/pii/0010027785900083. doi:10.1016/0010-0277(85)90008-3.

[3] R. Baillargeon, J. Li, Y. Gertner, D. Wu, How Do Infants Reason about Physical Events?, in: U. Goswami (Ed.), The Wiley-Blackwell Handbook of Childhood Cognitive Development, 2010, pp. 11–48. Publisher: Wiley Online Library.

[4] E. Herrmann, J. Call, M. V. Hernàndez-Lloreda, B. Hare, M. Tomasello, Humans Have Evolved Specialized Skills of Social Cognition: The Cultural Intelligence Hypothesis, Science 317 (2007) 1360–1366. URL: https://www.science.org/doi/10.1126/science.1146282. doi:10.1126/science.1146282.

[5] J. G. Bremner, A. M. Slater, S. P. Johnson, Perception of Object Persistence: The Origins of Object Permanence in Infancy, Child Development Perspectives 9 (2015) 7–13. URL: https://onlinelibrary.

wiley.com/doi/abs/10.1111/cdep.12098. doi:10.1111/cdep.12098, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cdep.12098.

[6] J. I. Flombaum, B. J. Scholl, A temporal same-object advantage in the tunnel effect: Facilitated change detection for persisting objects., Journal of Experimental Psychology: Human Perception and Performance 32 (2006) 840–853. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/0096-1523.32.4.840. doi:10.1037/0096-1523.32.4.840.

[7] B. J. Scholl, Object Persistence in Philosophy and Psychology, Mind & Language 22 (2007) 563–591. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0017.2007.00321.x. doi:10.1111/j.1468-0017.2007.00321.x, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0017.2007.00321.x.

[8] J. I. Flombaum, B. J. Scholl, L. R. Santos, Spatiotemporal priority as a fundamental principle of object persistence, The origins of object knowledge (2009) 135–164. Publisher: Citeseer.

[9] J. I. Flombaum, S. M. Kundey, L. R. Santos, B. J. Scholl, Dynamic Object Individuation in Rhesus Macaques: A Study of the Tunnel Effect, Psychological Science 15 (2004) 795–800. URL: https://doi.org/10.1111/j.0956-7976.2004.00758.x. doi:10.1111/j.0956-7976.2004.00758.x, publisher: SAGE Publications Inc.

[10] T. Wilcox, R. Baillargeon, Object individuation in infancy: The use of featural information in reasoning about occlusion events, Cognitive psychology 37 (1998) 97–155. Publisher: Elsevier.

[11] T. Wilcox, Object individuation: infants' use of shape, size, pattern, and color, Cognition 72 (1999) 125–166. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010027799000359. doi:10.1016/S0010-0277(99)00035-9.

[12] T. Wilcox, C. Chapa, Priming infants to attend to color and pattern information in an individuation task, Cognition 90 (2004) 265–302. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010027703001471. doi:10.1016/S0010-0277(03)00147-1.

[13] K. Voudouris, M. Crosby, B. Beyret, J. Hernández-Orallo, M. Shanahan, M. Halina, L. Cheke, Direct Human-AI Comparison in the Animal-AI Environment, Technical Report, PsyArXiv, 2021. URL: https://psyarxiv.com/me3xy/. doi:10.31234/osf.io/me3xy, type: article.

[14] R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri, D. Tran, Detect-and-Track: Efficient Pose Estimation in Videos, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Salt Lake City, UT, 2018, pp. 350–359. URL: https://ieeexplore.ieee.org/document/8578142/. doi:10.1109/CVPR.2018.00044.

[15] D. Gunning, Machine Common Sense Concept Paper (2018) 18.

[16] M. Crosby, B. Beyret, M. Shanahan, J. Hernández-Orallo, L. Cheke, M. Halina, The animal-AI testbed and competition, in: NeurIPS 2019 competition and demonstration track, PMLR, 2020, pp. 164–176.

[17] R. Baillargeon, Innate Ideas Revisited: For a Principle of Persistence in Infants' Physical Reasoning, Perspectives on Psychological Science 3 (2008) 2–13. URL: https://doi.org/10.1111/j.1745-6916.2008.00056.x. doi:10.1111/j.1745-6916.2008.00056.x, publisher: SAGE Publications Inc.

[18] Z. Pylyshyn, Perception, representation, and the world: The FINST that binds, in: D. Dedrick, L. Trick (Eds.), Computation, Cognition, and Pylyshyn, MIT Press, 2009, pp. 3–48. Google-Books-ID: zZ3HGOug3k8C.

[19] C. Fields, How humans solve the frame problem, Journal of Experimental & Theoretical Artificial Intelligence 25 (2013) 441–456. URL: http://www.tandfonline.com/doi/abs/10.1080/0952813X.2012.741624. doi:10.1080/0952813X.2012.741624.

[20] C. A. Fields, The Principle of Persistence, Leibniz's Law, and the Computational Task of Object Re-Identification, Human Development 56 (2013) 147–166. URL: https://www.jstor.org/stable/26764659, publisher: S. Karger AG.

[21] M. Shanahan, Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia, MIT Press, 1997. Google-Books-ID: z8zR3Ds7xKQC.

[22] R. Muñoz-Salinas, M. J. Marín-Jimenez, R. Medina-Carnicer, SPM-SLAM: Simultaneous localization and mapping with squared planar markers, Pattern Recognition 86 (2019) 156–171. URL: https://www.sciencedirect.com/science/article/pii/S0031320318303224. doi:10.1016/j.patcog.2018.09.003.

[23] A. Lampinen, S. Chan, A. Banino, F. Hill, Towards mental time travel: a hierarchical memory for reinforcement learning agents, in: Advances in Neural Information Processing Systems, volume 34, Curran Associates, Inc., 2021, pp. 28182–28195. URL: https://proceedings.neurips.cc/paper/2021/hash/ed519dacc89b2bead3f453b0b05a4a8b-Abstract.html.

[24] I. C. Uzgiris, J. M. Hunt, Assessment in infancy: Ordinal scales of psychological development, Assessment in infancy: Ordinal scales of psychological development, University of Illinois Press, Champaign, IL, US, 1975. Pages: xi, 263.

[25] J. Z. Leibo, C. d. M. d'Autume, D. Zoran, D. Amos, C. Beattie, K. Anderson, A. G. Castañeda,

M. Sanchez, S. Green, A. Gruslys, S. Legg, D. Hassabis, M. M. Botvinick, Psychlab: A Psychology Laboratory for Deep Reinforcement Learning Agents, 2018. URL: http://arxiv.org/abs/1801.08116, number: arXiv:1801.08116 arXiv:1801.08116 [cs, q-bio].

[26] A. Shamsian, O. Kleinfeld, A. Globerson, G. Chechik, Learning Object Permanence from Video, arXiv:2003.10469 [cs] (2020). URL: http://arxiv.org/abs/2003.10469, arXiv: 2003.10469.

[27] P. Tokmakov, A. Jabri, J. Li, A. Gaidon, Object Permanence Emerges in a Random Walk along Memory, arXiv:2204.01784 [cs] (2022). URL: http://arxiv.org/abs/2204.01784, arXiv: 2204.01784.

[28] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The KITTI vision benchmark suite, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3354–3361. doi:10.1109/CVPR.2012.6248074, iSSN: 1063-6919.

[29] L. Piloto, A. Weinstein, D. TB, A. Ahuja, M. Mirza, G. Wayne, D. Amos, C.-c. Hung, M. Botvinick, Probing Physics Knowledge Using Tools from Developmental Psychology, Technical Report arXiv:1804.01128, arXiv, 2018. URL: http://arxiv.org/abs/1804.01128, arXiv:1804.01128 [cs] type: article.

[30] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, F. A. Wichmann, Shortcut learning in deep neural networks, Nature Machine Intelligence 2 (2020) 665–673. URL: https://www.nature.com/articles/s42256-020-00257-z. doi:10.1038/s42256-020-00257-z, number: 11 Publisher: Nature Publishing Group.

[31] A. Agrawal, D. Batra, D. Parikh, A. Kembhavi, Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Salt Lake City, UT, 2018, pp. 4971–4980. URL: https://ieeexplore.ieee.org/document/8578620/. doi:10.1109/CVPR.2018.00522.

[32] M. Crosby, Building Thinking Machines by Solving Animal Cognition Tasks, Minds and Machines 30 (2020) 589–615. URL: https://doi.org/10.1007/s11023-020-09535-6. doi:10.1007/s11023-020-09535-6.

[33] D. Teney, E. Abbasnejad, K. Kafle, R. Shrestha, C. Kanan, A. van den Hengel, On the Value of Out-of-Distribution Testing: An Example of Goodhart' s Law, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 407–417. URL: https://proceedings.neurips.cc/paper/2020/hash/045117b0e0a11a242b9765e79cbf113f-Abstract.html.

[34] M. Jaderberg, W. M. Czarnecki, I. Dunning, L. Marris, G. Lever, A. G. Castañeda, C. Beattie, N. C. Rabinowitz, A. S. Morcos, A. Ruderman, N. Sonnerat, T. Green, L. Deason, J. Z. Leibo, D. Silver, D. Hassabis, K. Kavukcuoglu, T. Graepel, Human-level performance in 3D multiplayer games with population-based reinforcement learning, Science 364 (2019) 859–865. URL: https://www.science.org/doi/full/10.1126/science.aau6249. doi:10.1126/science.aau6249, publisher: American Association for the Advancement of Science.

[35] J. Hernández-Orallo, Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement, Artificial Intelligence Review 48 (2017) 397–447. URL: https://doi.org/10.1007/s10462-016-9505-7. doi:10.1007/s10462-016-9505-7.

[36] J. Hernández-Orallo, The Measure of All Minds: Evaluating Natural and Artificial Intelligence, Cambridge University Press, 2017.

[37] E. Triana, R. Pasnak, Object permanence in cats and dogs, Animal Learning & Behavior 9 (1981) 135–139. URL: http://link.springer.com/10.3758/BF03212035. doi:10.3758/BF03212035.

[38] A. Juliani, V.-P. Berges, E. Teng, A. Cohen, J. Harper, C. Elion, C. Goy, Y. Gao, H. Henry, M. Mattar, D. Lange, Unity: A General Platform for Intelligent Agents, Technical Report arXiv:1809.02627, arXiv, 2020. URL: http://arxiv.org/abs/1809.02627, arXiv:1809.02627 [cs, stat] type: article.

[39] L. Burke, On the Tunnel Effect, Quarterly Journal of Experimental Psychology 4 (1952) 121–138. URL: http://journals.sagepub.com/doi/10.1080/17470215208416611. doi:10.1080/17470215208416611.

[40] A. Michotte, G. Thines, G. Crabbé, Les complements amodaux des structures perceptives (Amodal completion of perceptual structures), Studia Psychologica. Publications Universitaires de Louvain.[GV] (1964).

[41] C. Fields, Trajectory Recognition as the Basis for Object Individuation: A Functional Model of Object File Instantiation and Object-Token Encoding, Frontiers in Psychology 2 (2011). URL: https://www.frontiersin.org/article/10.3389/fpsyg.2011.00049.

[42] G. Johansson, Configurations in Event Perception (Uppsala, Sweden, Almqvist and Wiksell. Johansson, G.(1973). Visual perception of biological motion and a model for its analysis. Perception and Psychophysics 14 (1950) 201–211.

[43] G. Johansson, Rigidity, Stability, and Motion in Perceptual Space, Nordisk Psykologi 10 (1958) 191–202. URL: https://doi.org/10.1080/00291463.1958.10780387. doi:10.1080/00291463.1958.10780387, publisher: Routledge _eprint: https://doi.org/10.1080/00291463.1958.10780387.

[44] C. Buckner, Understanding associative and cognitive explanations in comparative psychology, The Routledge handbook of philosophy of animal minds (2017) 409–419. Publisher: Routledge.

[45] M. Dacey, Evidence in Default: Rejecting default models of animal minds, The British Journal for the Philosophy of Science (2021) 714799. URL: https://www.journals.uchicago.edu/doi/10.1086/714799. doi:10.1086/714799.

[46] I. M. Pepperberg, M. R. Willner, L. B. Gravitz, Development of Piagetian Object Permanence in a Grey Parrot (Psittacus erithacus), Journal of Comparative Psychology 111 (1997) 22.

[47] C. Chiandetti, G. Vallortigara, Intuitive physical reasoning about occluded objects by inexperienced chicks, Proceedings of the Royal Society B: Biological Sciences 278 (2011) 2621–2627. URL: https://royalsocietypublishing.org/doi/full/10.1098/rspb.2010.2381. doi:10.1098/rspb.2010.2381, publisher: Royal Society.

[48] R. Burnell, J. Burden, D. Rutar, K. Voudouris, L. Cheke, J. Hernandez-Orallo, Not a Number: Identifying Instance Features for Capability-Oriented Evaluation, forthcoming, p. 9. URL: https://ryanburnell.com/wp-content/uploads/Burnell-et-al-2022-Not-a-Number.pdf.