Minds and Machines
https://doi.org/10.1007/s11023-020-09535-6

GENERAL ARTICLE

Building Thinking Machines by Solving Animal Cognition Tasks

Matthew Crosby[1]

Received: 30 April 2020 / Accepted: 5 August 2020

Abstract
In 'Computing Machinery and Intelligence', Turing, sceptical of the question 'Can machines think?', quickly replaces it with an experimentally verifiable test: the imitation game. I suggest that for such a move to be successful the test needs to be *relevant, expansive, solvable by exemplars, unpredictable*, and lead to *actionable* research. The Imitation Game is only partially successful in this regard and its reliance on language, whilst insightful for partially solving the problem, has put AI progress on the wrong foot, prescribing a top-down approach for building thinking machines. I argue that to fix shortcomings with modern AI systems a nonverbal operationalisation is required. This is provided by the recent Animal-AI Testbed, which translates animal cognition tests for AI and provides a bottom-up research pathway for building thinking machines that create predictive models of their environment from sensory input.

Keywords Turing test · AI · Comparative cognition · Deep reinforcement learning

1 Introduction

In his 1950 paper 'Computing Machinery and Intelligence', Turing posed the question 'Can machines think?', then promptly discarded it in favour of an experimentally verifiable question: 'Can machines perform well at the imitation game?', an open-ended verbal test of conversational ability (Turing 1950). Turing's move to replace the question with a verbal test provided a valuable empirical foothold for tackling the original ill-defined question. Unfortunately, the reliance on purely written questions and responses meant that physical components of intelligence were overlooked (Harnad 1991). Turing has had lasting impact on the field and this has led some research areas, such as purely symbolic AI, to attempt to build intelligence from the top-down. Here, higher cognitive functions are modelled in

✉ Matthew Crosby
m.crosby@imperial.ac.uk

[1] Leverhulme Centre for the Future of Intelligence, Imperial College London, London, UK

Published online: 13 August 2020

isolation from the low-level sensory pathways, predictive models, and interactive abilities they are scaffolded upon, at least conditionally, in our only current exemplars of thinking machines: biological organisms. Frameworks such as active inference suggest that these systems are instead built up from a drive to build predictive models and act in such a way as to best explain their sensory inputs (Friston et al. 2011). If we want to learn from the existing exemplars of the quality then a bottom-up approach is warranted.

Following arguments given by Dennett (1984) I consider the move that Turing made to be one of *weak operationalisation*. In other words, Turing replaces the original property (thinking in machines) with an empirical measure that can be taken as a non-perfect marker for presence of that property. The success of such a move depends on a number of criteria that I introduce in this paper. Namely, the test must be *relevant, expansive, solvable by exemplars*, and, especially when the test will become a target and not just, say, a physical measure, must be *unpredictable* and lead to *actionable research*. Under these metrics I show that the Imitation Game, whilst clever, is not a particularly good operationalisation of the original question. The *actionable research* metric is particularly important in this case because the starting point 'can machines think' is intended as a future hypothetical. The main pragmatic benefit of a *weak operationalisation* is not to apply the measure on existing systems, but to be able to design systems that pass the measure.

In this paper I introduce a different way to operationalise 'can machines think?' which ignores the verbal components of thought and instead focuses on physical intelligence. This bottom-up operationalisation is currently implemented as the Animal-AI testbed (Crosby et al. 2020), in a way which, I will argue, provides our best research path towards building thinking machines. Further information about the testbed and links to the fully open source code can be found at http://animalaiolympics.com.

One reason that Turing focused only on the 'higher types of thinking' in his 1950 paper is that, at the time, computer vision was expected to be an incredibly complex problem that would require perhaps unforeseeable advances in the power of machines. As Turing himself stated when arguing for the feasibility of machines that pass the imitation game, "only a very small fraction [of the storage capacity of the brain] is used for the higher types of thinking. Most of it is probably used for the retention of visual impressions." (Turing 1950). In the intervening 70 years there has been significant progress in building machines. With deep reinforcement learning it is now possible to use pixel inputs, albeit of much lesser dimensions than the number of photoreceptors in a human or animal eye, to train systems capable of playing Atari games at a human level (Mnih et al. 2015). This means that the bottom-up approach is finally at least conceivable, even if only just beginning to show promise in more sophisticated 3D environments, whether simulated (Crosby et al. 2020; Juliani et al. 2019; Guss et al. 2019), or as constrained problems in the real world (Akkaya et al. 2019). Given that seventy years later the top-down approach has failed to produce thinking machines, and that now the technology is finally in place to begin, it is now time to give the bottom-up approach a try (Lake et al. 2017).

Simultaneous to the advancements in AI, the non-AI world has also seen the development of sophisticated experiments to understand how animals interpret the physical world they inhabit, and how far they are able to react appropriately in novel environments (Shettleworth 2009). These experiments are far from perfect (Farrar and Ostojić 2019), but the translation to AI can solve most of the problems that are particular to working with animals such as small sample sizes both in terms of number of experiments and number of participants, and trophy hunting for outlier displays of cognitive ability (Allen 2014). By putting together an evolving testbed of experiments, each designed to test for a particular aspect of physical intelligence, it is possible to build up an *expansive* testbed, which is *relevant* under frameworks that believe thought is built up from low-level cognitive abilities. As the tests are solvable by most animals and certainly by humans, they are *solvable by exemplars*, indeed more so than the imitation game which is restricted to language-using entities. The Animal-AI Testbed was first presented as a competition with hidden tests which have now been released publicly. The addition of further hidden tests based on the same original ideas ensures that they remain *unpredictable*, whilst the release of the original tests, and its breakdown into categories, provides an *actionable research* path.

The tests involved in Animal-AI's *relevance* for 'thinking' depend in part in on the separation between low-level abilities of embedded systems and high-level cognitive concepts. A more continuous outlook such as that suggested by representation-friendly versions of embodied cognition or dynamical systems theory (Shapiro 2019) will naturally find the presented tests relevant. A framework that separates high-level cognitive tasks as the pure realm of thought will have to be satisfied that the testbed contains many tasks from the more exceptional purported abilities of animals. These include delayed gratification, which tests for the refusal of an immediate smaller reward in return for a greater future reward, numerosity, which tests the ability to choose a larger number of rewards even in cases that require addition of unseen rewards, and tool use, which requires manipulating environment objects to make accessible new areas of the environment.

The first goal of this paper is to set out the qualities a *weak operationalisation* must have in order to be successful. I then survey the current state of testing in comparative cognition (Sect. 3) and show how an AI environment can be set up to translate many key tests (Sect. 4). In the final part of the paper I present the Animal-AI testbed (Sect. 5) and show how it better fits our requirements than the imitation game. The testbed is designed to have the equivalent properties of a Turing Test for nonverbal thought, and contains many different types of experiment to cover different aspects of physical intelligence and common sense reasoning. The testbed was recently used as part of a competition (Crosby et al. 2020), which showed that both that we are a long way from solving all of the tests and that the testbed is robust to simple solutions and ready to be tackled in long-term research. I finish the paper (Sect. 7) by discussing some possible objections to the testbed as a route to thinking machines and consider ways in which it can be improved. I ultimately conclude that solving Animal-AI should be the next step towards building machines that think.

2 Operationalising 'Can Machines Think?'

In the 70 years since Turing discarded the question 'Can Machines Think?' as "too meaningless to deserve discussion" the term 'machine' has become ubiquitous in AI, with one of the major current research strands, 'Machine Learning', using the term liberally to apply to all kinds of AI algorithms or agents. Even attributing *thinking* and related psychological properties to machines, and potential future machines, has become more common. Prominent examples include Lake et al. (2017) who set forth a research agenda for building machines that think like people, Racanière et al. (2017) who present a deep reinforcement learning architecture where agents 'imagine' future states based on rollouts of their policy in order to find the best course of action, and Rabinowitz et al. (2018) who present a method for 'Machine Theory of Mind', ToMNet, by which agents can learn to make predictions about other agents' mental states (to use terminology from the paper).

However, there are dangers to using such psychologically loaded terms to apply to Machine Learning agents. Shevlin and Halina caution against the use of 'rich psychological concepts' (those that have complex histories in cognitive science), and in particular suggest that ToMNet would be better presented as performing behaviour reading and not as inferring mental states (Shevlin and Halina 2019). They argue that artificial behaviour reading is still impressive and useful research, and this characterisation would avoid unnecessarily posing questions about machine mental states. Turing also avoids explicitly evaluating machine 'thinking', instead positing what he claims is a closely-related empirical question: can machines defeat humans in the imitation game? I follow in these footsteps. The goal is not to provide a rigorous marker of thought, like with the related concept of cognition this may not be possible (Akagi 2018; Buckner 2015), but a pragmatic marker that can guide research.

In the imitation game, a human judge must determine which of two players (one male, one female) is male, solely by asking questions and analysing their answers. In the machine version, the male player is replaced by a machine. To avoid confounding factors such as appearance being taken into account, the questions are asked and answered via typed text only. There has been much discussion about the particular version of the test proposed by Turing, how it has evolved and been reinterpreted over the years, and what exactly it, and related tests, measure (Epstein et al. 2009; Proudfoot 2011; Block 1981; Hernández-Orallo 2000).

My reading of the original test is similar to that given by Dennett (1984), that it is intended as a *weak operationalisation* of the original question (Can machines think?). To operationalise a property (e.g. thinking in machines) is to give it an empirical measure (e.g. performance in the imitation game) that can be taken as a marker for attribution of this property. A *weak* operationalisation provides an empirical measure that is not intended as a foolproof identifier immune from false positives (or even false negatives). Even a weak operationalisation can have many benefits, especially in an engineering discipline such as computer science. In this case it lays out a research path: how do you build a machine that passes the test? The successful conclusion of this research will either be a machine that has the property in question, or, at the very least, will provide a much better vantage point from which to (re)tackle the original question.

Unfortunately, what has now become known as the Turing Test no longer operates under normal testing parameters. For one thing, it has become a target, with researchers aiming to pass by any means necessary. Goodhart's law (Goodhart 1984), often restated as

> When a measure becomes a target, it ceases to be a good measure.

applies especially in cases of weak operationalism. Consider the weak operationalisation of the question 'is a painting important?' to 'has it appeared on the wall of an art gallery that attracts at least *n* visitors a year?'. This is not a great definition, it is purposefully *weak*. Nevertheless, it is not hard to imagine it serving some informal purpose. However, if this test were considered the de facto measure, anyone could become the artist of an important painting with the right connections, or perhaps some blu-tac and a little mischief. The process of creating important art would become removed from the qualities of the piece itself. In a similar way, trying to win the Loebner prize (a long-running Turing Test inspired competition) is a very different goal from trying to build a thinking machine (Hernández-Orallo 2000).

I have suggested that Turing's operationalisation must, due to the nature of the question, be *weak*. This is not meant in a purely negative sense; a weak operationalisation may be the only sensible option when dealing with ill-defined terms like 'thinking' and can have many benefits. The Turing Test can be empirically analysed; it opens up a concrete research path; and it sidesteps ambiguous questions about the nature of thought to which there may be no definitive answer. The imitation game is not unique as a possible measure and there are many possible candidate measures that could perform a similar function. In the rest of this section I will analyse the qualities of that a good *weak* operationalisation should have and also how well the imitation game itself fares. As the term Turing Test has become so ubiquitous, and because I will later introduce a new Testbed, from now I will use the term 'test' as a stand-in for a measure or set of measures.

The first quality, which I will call *relevance*, and which must hold for any type of operationalisation, is that the introduced test must be solvable only by utilising or referencing abilities relevant to the original property in question. If it can be solved without such relevant abilities, then the test does not identify the correct attributes. If it requires too many irrelevant abilities, then these become uncontrolled confounds. Using text as a mediator in the imitation game and allowing any type of question is a way to ensure some relevance. Answering (the right kind of) questions involves something akin to thinking, and restricting the test to written words removes any concerns that the machine needs to look human, or even to look like it is thinking. Looking like you are thinking is a bad operationalisation of thinking because it fails on *relevance*. With the Turing Test *relevance* is in the hands of the judge. If the judge does not ask questions that require any thought to answer and merely exchanges pleasantries, then the test is no longer any good, and it is also fairly easy to build a machine that would pass. By limiting to textual answers, aspects of physical intelligence that might be relevant for thought are discounted, though some could be assessed with clever hypothetical questions.

The second quality is that the test must be *expansive*. By this I mean that it should test for as many aspects of the original, ambiguous property as possible. This helps to ensure that the property is captured to its fullest extent. A common tactic in analysing ill-defined terms is to break them down into more manageable components. If only a single component is used as part of a test, then this can tell only a partial story. For example, a bad operationalisation of prime-ness in very large integers would be that they are odd. While this is a relevant property, the test is not *expansive* enough and returns too many false positives. The imitation game is potentially *expansive*, including the whole set of writable questions, but in reality modern versions of the Turing Test are only as expansive as the judges' questions. If the judges do not cover a broad range of topics and employ different questioning styles then this property is not met and the test cannot be considered to be a good indicator of thinking.

The third quality, which I will call the *exemplar* quality, is that the test should be passed by as many known exemplars of the original property in question as possible, and as few from outside that set as possible. This increases the likelihood that solving it requires only relevant abilities. The imitation game is defined in terms of average human performance, which is aiming too high as it cuts out roughly half of the population of human exemplars. It might have been tempting to set the bar of a test as high as possible to reduce false positives, but this makes it too exclusive. In the case of 'thinking', this could correspond to setting the test to beat the strongest human players at the ancient strategy board game Go. These humans are clearly prime exemplars of the ability to think. However, we now have AI systems that can outperform all humans at Go (Silver et al. 2017), and they are not generally considered to be good examples of thinking machines. Defeating humans at Go was a milestone in AI progress, but the problem was too specialised to relate directly back to the property of thinking. The *exemplar* quality instead suggests that the bar should be set low, another reason to consider animal tasks as the starting point.

The first three qualities are all ways to ensure that the operationalisation does not stray too far from the original question it was trying to answer and apply equally to *strong* and *weak operationalisations*. For *weak operationalisations*, especially those providing markers for future hypothetical systems, There are two further properties that are required to ensure that the test, when used in practice, is a pragmatic marker for the initial property.

The first practical constraint for engineering tests is that the test must contain *unpredictable* elements. If the test is too predictable, then it will be subject to Goodhart's law, and will also be solvable by 'shortcut' methods that do not engage with the original property. This is especially true for tests of AI systems, where deterministic problems can be solved by a database of answers or action sequences (Searle 1980). The imitation game neatly adds unpredictability by allowing the judges to ask any questions they deem fit. The questions are not known before the test begins. Participants could be asked to interpret recent political events, to speculate about the future, or to write poetry. However, again, the quality of the test is determined by the predictability of the judges. If the questions can be predicted easily, then answers to them can be hardcoded.

In our case, the question ‘can machines think?’ is, still seventy years later, a question about possible future machines. The operationalisation of this question is not intended as a definition, and would be useless if it could never be met. As there are currently no thinking machines, the only way to potentially meet it is to make progress on the measure that it sets out. This means that an important final quality of an operationalisation of this type of question is that it leads to an *actionable* research plan. We do not want to ask if machines can think just to find out a theoretical answer, but to help us work out how to build them should that answer be ‘yes’. Even if we are only interested in the former, then progress towards the latter will still be helpful in this regard as new insights will certainly come through the process of trying to solve the test. This final quality does not apply to operationalisations in non-engineering contexts or for non-hypothetical questions, but in this context it might be the most important. Again the Turing Test has mixed results here. The last seventy years have shown the Turing Test to be unsuccessful for leading to creating thinking machines, but successful at inspiring progress in AI.

I have argued that the imitation game is just one possible weak operationalisation of the question ‘can machines think?’. It partially succeeds on most accounts, depending on the quality of the judges, but sets the bar too high to be solvable by most *exemplars*. Its *relevance* depends on whether it is possible for thinking machines to be disembodied and working in discrete action spaces, a discussion that I do not have space to elaborate fully here. For this paper I will assume that modern representation-friendly views in the tradition of embodied cognition, such as active inference (Williams 2018; Clark 2015), are correct in giving a strong emphasis to the roles of action and interaction and predictive models, in all known thinking machines. Under these views, a verbal-only test misses a necessary component of all known exemplars.

I will now turn to main goal of this paper, introducing a new nonverbal operationalisation that scores well across all five qualities introduced in this section. This nonverbal test assumes that thinking machines can be built through increasingly sophisticated abilities to interact with and model properties of their environment. To be *expansive*, a wide range of tests will be needed. For this, we can draw from numerous comparative cognition studies of aspects of ‘thinking’ in nonverbal biological entities, including tests performed on non-human animals and studies in child development.

3 Comparative Cognition

One method for creating a non-language based operationalisation could be to attempt to define a testing environment that is as open-ended as the imitation game in terms of the type of questions that can be posed inside of it. A way to do this that does not move to far from Turing would be to use open-ended formal languages in place of natural language. This has be suggested in terms of the ability to write efficient sequence descriptions in the C-Test described in Hernández-Orallo (2000).

A more general version would be problem classes contained by the AIXI model described in Hutter (2000). These sets of tasks are still language-based, just using formal languages instead of human language, but most problematically under our framework, they contain many tasks not solvable by *exemplars* of thought. These tests are much better suited to measuring abstract non-anthropomorphic notions of intelligence, which is not our goal here. The non-language based abilities we are interested in are those involving continuous interaction with an environment.

Another attempt would be to collect together as many non-language based tasks that we expect our exemplar thinking machines to be able to solve. Instead of creating this non-language based operationalisation from scratch, we can draw on the wealth of behavioural tests and ideas from research in comparative cognition in non-language or pre-language entities, such as non-human animals and children. The overall goal of this research is not explicitly to answer the question 'can animals/children think?', but to explore some of their capabilities through experiment. Similarly, our overall goal is not explicitly to answer the question, 'can machines think?', but to create a test that is as *relevant, expansive, unpredictable*, solvable by as many *exemplars* as possible, and that provides an *actionable* research tool.

Capabilities investigated in comparative cognition include understanding the hidden causal relations required for effective tool use (Bluff et al. 2010), the ability to forgo an immediate reward for a later, larger, reward (also known as delayed gratification) (Beran 2002), and possessing theory of mind (Penn and Povinelli 2007). With in each case disputes as to animals even do possess the relevant abilities. These are all areas that have also been studied in AI in recent years. Tool use was found to be an emergent property in Open-AI's hide and seek experiments using large scale standard reinforcement learning algorithms (Baker et al. 2019), a variation of delayed gratification, accepting a small negative reward for a large later positive reward, is included in DeepMind Lab in the 'Stairway to Melon' level (Beattie et al. 2016), and I have already briefly introduced the Machine Theory of Mind work (Rabinowit et al. 2018).

A particular task that involves tool use and causal reasoning is the Aesop's Fable task. In the basic version of this task, an animal must retrieve food from a container of water that has been fixed in place. The food floats on top of the water just out of reach of the animal, but with some conveniently placed rocks (or other sinking objects) in the vicinity. The solution is to drop the rocks into the container to raise the water level high enough to retrieve the food floating on top. Behaviour that can solve this task has been observed in corvids (Jelbert et al. 2014). However, it is unclear just what kind of reasoning, if any, is involved in solving the task (Hennefield et al. 2018). Consider a single happy accident in which the frustrated animal accidentally drops a heavy object in the container while annoyed it cannot retrieve the food. This has the effect of raising the water level and bringing the food closer, and the animal may associate the sequence of actions with the positive reward of more obtainable food without the properties of water and objects ever mattering to the solution. In the ideal case though, an entity presented for the first time with a situation containing some rocks, a beaker full of sand with inaccessible food on top, and a beaker full of water with inaccessible food on top would be displaying the ability to think if it, perhaps after a pause to take in the scene, picked up a rock and dropped it in the beaker of water.

Taken individually, comparative cognition tests do not have the *expansiveness* quality, and unfortunately they are often performed in isolation due to the large costs in time and effort of working with animals and children. This has been rectified somewhat recently, with the introduction of test batteries (Herrmann et al. 2007; Shaw and Schmelz 2017), but the problems of cost still remain. We should be wary of any individual behavioural test results that supposedly confirm some higher thought process or capability. This was recognised when as comparative cognition was first founded as a field, with Morgan arguing that many apparent demonstrations of 'thinking' can be explained by intelligent adaptation (Lloyd Morgan 1894). More recently, many have argued that many high-level capabilities such as planning and tool use can be explained by associative learning (similar to the reinforcement learning employed in many of the AI examples given so far) (Lind 2018). Farrar and Ostojić (2019) give several reasons why standard claims of cognitive abilities in comparative cognition fail. These include the publishing bias towards finding exceptional cognitive abilities, or trophy hunting (Allen 2014), which leads to confirmatory research methods, and lack of high enough sample sizes or rigorous statistical analysis.

Fortunately, we can mitigate these problems when porting tests to AI. AI systems are considerably easier to work with; they can be tested unsupervised without breaks, and at very little cost and can be restarted to experience environments afresh when required. This means that not only can a single system be tested on all tests that have been ported over, but can also be retested on many variations to build up statistically significant sample sizes. The quality of *expansiveness*, arguably missing in comparative cognition, can be met when performing similar tests on AI systems. The bias towards reporting, and therefore even finding (perhaps erroneously) also becomes a non-issue when individual tests are placed in the context of a large test battery and positive results are seen purely as reason for further study.

Most studies on animals and children automatically pass the *unpredictability* constraint. In most cases, the animal or child could not have expected the test, and probably does not even conceptualise it as a test. As long as good practice is maintained by the experimenters, there is no reason that the animal or child should be over-prepared. This is a much more significant problem in AI. Once a test has been revealed, a team of engineers can work to create a system with the sole goal of passing the test. One way to combat this is to use hidden test problems which are fair, but could not be predicted by the engineers. This will be the main topic of the next section.

The only remaining quality is that the tests must lead to an *actionable* research plan. Opening up research paths is less relevant as a concern in comparative cognition, which does not aim to genetically engineer or build biological creatures to pass the tests, but only to determine the capabilities of existing creatures. However, when adapting tests for AI systems it is useful to provide a wide range of tests of different difficulties, and to cohere with existing AI research paradigms. This way, there will always be a next problem to solve (a subset of the currently unsolved tasks) that can be isolated for research.

In summary, experiments in comparative cognition form a good basis for operationalising nonverbal components of thinking. They are naturally *relevant* (assuming a representationalist version of embodied cognition such as active inference) and *solvable by exemplars*. The issues with *expansiveness* in comparative cognition can be solved with the move to AI by including a wide variety of tests and performing extensive testing that is not practical with animals or children. The remaining challenge is maintaining *unpredictability*, which comes automatically when working with animals, but is lost when teams of engineers deliberately set out to solve tests. I will discuss strategies for this in the next section.

4 From Comparative Cognition to AI

When translating tests designed for animals to AI they must take into account current AI capabilities and practices. In 1950, it was impossible to predict precisely how AI would progress. The field has been heavily influenced by the hardware that has been developed. For example, much of machine learning has been shaped by the utilisation of parallel processing from GPUs, primarily developed as graphics engines. Nevertheless, there are some surprisingly accurate predictions from the era, including Turing's own suggestion for creating learning machines built on reinforcement signals of reward and punishment.

One statement that has continued to hold is Moravec's paradox, formulated in the 1980s:

> It is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility (Moravec 1988).

This makes sense in the context of evolutionary history; there is a much longer period of developing sensorimotor skills than the relatively recent higher-level thinking. We have also seen this play out in AI (in reverse temporal order), where we now have superhuman Go and chess players, but are still a long way from a robot that can solve the Aesop's Fable task.

One problem that modern machine learning excels at is i.i.d. testing of labelled data. In machine learning, the i.i.d. assumption (that the datapoints are *independent and identically distributed*), is used when the data (both for training and testing) has been generated in the same way. A classic example where i.i.d. testing is common is with ImageNet (Deng et al. 2009), which contains hierarchically labelled images with over 20,000 classes, though subsets of the full database are more commonly used. AI now outperforms humans at image classification on this dataset (He et al. 2015) and in general does well at i.i.d. tasks.

The application of the i.i.d. label to ImageNet is fairly loose, relying on an assumption that the data collection method used for the images is principled enough to be determined constant and that the images contained represent an unbiased sample of that method. Recht et al. (2019) show that classifiers with high accuracy on ImageNet show a significant drop in accuracy when tested on new images generated in a way that attempts to match the generation process of ImageNet. This suggests that there is overfitting, even in this large dataset. Such overfitting is the exact reason to include the *unpredictability* constraint and shows the tendency for systems to be built purely to pass a test and not to have the generalisable properties that the test was designed for. One solution to this problem proposed by Recht et al. is to have a 'super hold-out' set of data that is kept apart from the released dataset for multiple years and used only to test for overfitting to the released part of the dataset.

In contrast to i.i.d., *out-of-distribution* (*o.o.d.*) testing is much less common. Though it is starting to become more popular. One technique for o.o.d. testing in datasets like ImageNet is to separate out different contexts in which a labelled object can appear (Arjovsky et al. 2019; Geirhos et al. 2020). For example, if a dataset only contained cows that appear on grass and camels that appear on sand, then a classifier will likely find that the easiest method to accurately classify the images is to check for the presence of green or yellow pixels respectively. If green is in the background it is a cow, if yellow, a camel. An o.o.d. test of a camel on grass would be incorrectly labelled as being a cow. Of course, not all o.o.d. testing is good, and there is no point in taking just anything from outside the expected i.i.d. data. We can use a camel on grass because this is something that could naturally occur in a dataset of images, but it would not be fair to expect an ImageNet classifier to be able to read the word camel for example. Ideally, we want o.o.d. tests that illuminate relevant underlying properties of the intended solution, but this can often be hard to define.

The use of o.o.d. data will always be hard to define precisely, as what is fair, and what were the underlying generative processes of the original dataset is context dependent and in many cases somewhat subjective. Furthermore, the distinction already rests on unstable foundations with the way the i.i.d. assumption is applied quite loosely in machine learning. Nevertheless, o.o.d. tests have the benefit of being automatically *unpredictable*, and, when implemented fairly, are a good way of guaranteeing that solutions actually implement the underyling properties intended to be tested for. One way to achieve this is to define an environment within which the tests will take place along with a language within which the tests will be written, but not their exact content. This is similar to the process of setting out a syllabus and providing past papers for an exam. This constrains the subject matter (analagous to the environment for physical tests), and also sets out some conventions which the unseen test will follow.

As well as using unseen o.o.d. tests, we also need the test to lead to a practical research plan. This means that it needs to conform to many of the standard practices in AI. The environment within which we will perform the tests must be well-defined and available for open-ended training. It also must be sufficiently simplified enough that some progress can be made. It is not yet possible to solve animal cognition tasks in the real world, so a simplified simulated environment must be used with reduced observation size, simpler physics, and less noise. However, it must still be rich enough to incorporate a wide range of types of test in order to be *expansive*.

To summarise the last two sections, animal cognition tasks provide a good starting point for designing a testbed for AI systems, but some care must be made when translating them for use. It should be possible to translate as many different types of test as possible into a single well-defined environment that is suitable for research.

The environment should be simplified to the point that it is usable by current and near-future AI systems. The tests should be kept secret, but a syllabus should be given. Finally, to facilitate research, it should be easy to generate and work with test-like configurations within the environment. The next section presents the Animal-AI testbed, which was designed with these principles in mind.

5 The Animal-AI testbed

In this section I discuss the Animal-AI testbed as a weak operationalisation of the question 'can machines think?' for nonverbal intelligence. A presentation of the testbed from an AI perspective can be found in Crosby et al. (2020). All tests mentioned here can be played online at http://animalaiolympics.com, which also includes further details and links to all code needed to run the environment. An overview of the environment is shown in Fig. 1. The environment is a 3D simulated arena with simple physics in which an agent must navigate to retrieve rewards, similar to the food rewards commonly used in animal experiments. The agent receives pixel-based inputs each frame representing its visual point of view, along with some information about its current velocity. The agent then returns an action: move forwards/backwards and/or rotate left/right. To succeed in a task, the agent has to retrieve as much food as possible in the environment within the time limit (a given number of perception/action loops) by translating the visual inputs into successful action sequences. In many of the tests, this requires an intermediary step of interpreting the scene and predicting future states or reasoning based on the known properties of the objects and environment.

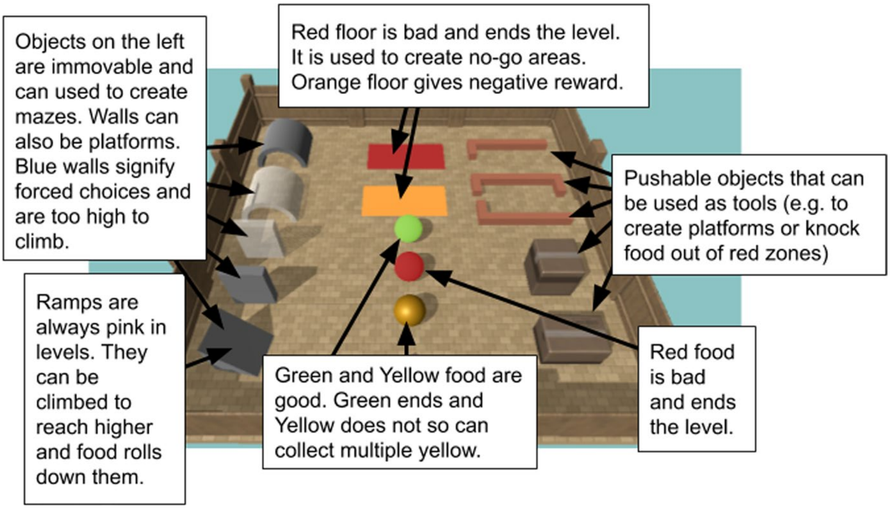
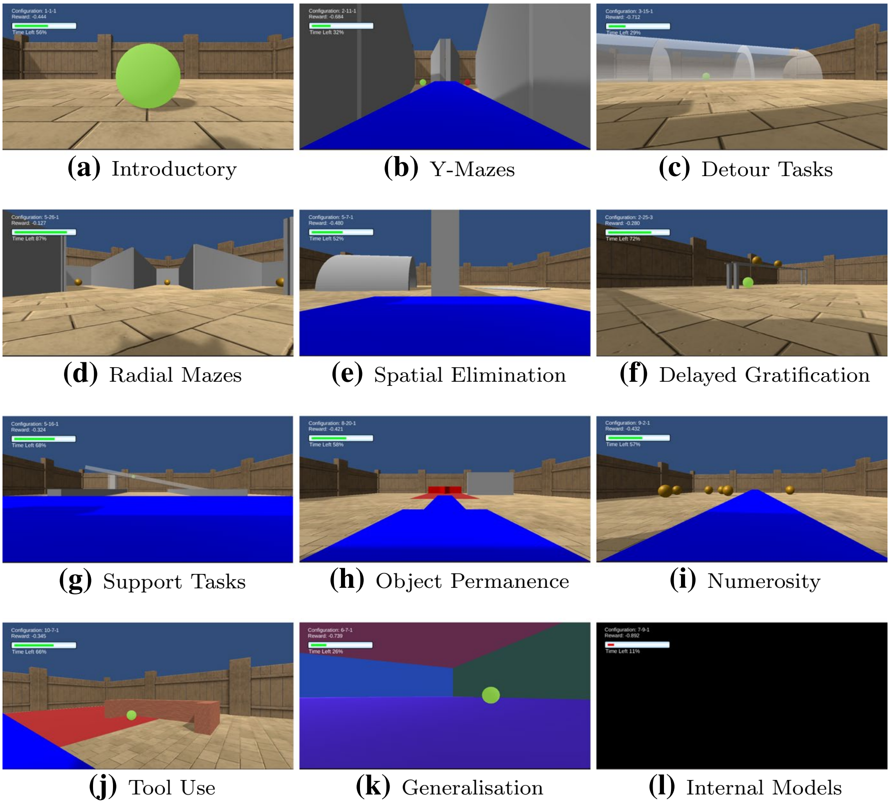

Objects on the left are immovable and can used to create mazes. Walls can also be platforms. Blue walls signify forced choices and are too high to climb.
Red floor is bad and ends the level. It is used to create no-go areas. Orange floor gives negative reward.
Pushable objects that can be used as tools (e.g. to create platforms or knock food out of red zones)
Ramps are always pink in levels. They can be climbed to reach higher and food rolls down them.
Green and Yellow food are good. Green ends and Yellow does not so can collect multiple yellow.
Red food is bad and ends the level.

Fig. 1 The Animal-AI environment. The objects shown are the building blocks used to create levels. They can be combined to make complex scenarios for the agent to solve

The objects in the environment can be used as building blocks to create many kinds of possible experiment. They can all be resized and rotated and the environment can hold any number of objects. This allows the associated testbed to be both *expansive* and *unpredictable*. To maintain its unpredictability, when used to measure intelligence, the exact configurations used for testing are kept secret. To keep them fair, all tests are built using a single configuration file that places all the objects at the start of the test. There are no interventions except for the agent's actions and the results of the physics engine. *Relevance* and *exemplar plausibility* are ensured by using tests that are solvable by animals and/or humans. Finally, the tests are designed to lead to an *actionable* research plan. The environment has been created to be compatible with standard machine learning frameworks (Beyret et al. 2019; Bellemare et al. 2012; Julian et al. 2018), and the test difficulties range from currently solvable, to requiring multiple new breakthroughs in research. The tests are arranged such that subsets can be tackled independently so that different research goals can separate out a new skill to tackle.



(a) Introductory **(b)** Y-Mazes **(c)** Detour Tasks

(d) Radial Mazes **(e)** Spatial Elimination **(f)** Delayed Gratification

(g) Support Tasks **(h)** Object Permanence **(i)** Numerosity

(j) Tool Use **(k)** Generalisation **(l)** Internal Models

Fig. 2 Example problems from each of the 12 task types. **a** contains familiarisation tasks. **b–j** have direct links to animal tests. **k**, **l** Are AI-specific. There are 900 total problems

Figure 2 shows a breakdown by type of the experiments included in the testbed. There are 900 variations in all, representing a wide range of different testing types. To maintain *unpredictability*, all the tests were originally kept secret and used as part of the Animal-AI Olympics competition. They have all be made publicly available to make research on them easier with a new set of hidden tasks made available in 2021. A large part of the testbed is comprised of introductory tests. These include tasks with single food items placed in front of the agent and tasks with a single ramp that the agent needs to navigate up to reach the food. This category can be broken down into many subcategories based on the type of object introduced, but for brevity we collect them all here. These are important to provide an *actionable* research goal, and provide some easier starting points to build research ideas from. If none of the tests were solvable then it would be hard to get started. They are also important to help analyse the capabilities of agents on more complex tasks. For example, consider an agent that manages to climb multiple ramps in a row to retrieve the food. If this agent cannot also solve the simpler task of climbing one ramp, then it probably just got lucky in the more complex case.

I am not claiming this to be anywhere close to a complete breakdown of all the elements required to build a thinking machine. I do not believe this to be even possible. If it was then there would have been much more progress in the last 70 years. I am instead claiming that this is an extensive set of tests that captures enough relevant abilities that solving them all would necessitate the conclusion that the agent is a thinking machine. The later tests have been specifically chosen to require maintaining temporally extended internal models of the environment that can be used for reasoning and prediction. These kind of abilities are similar to those given by Adam's mark of the cognitive in that the features of the models must have non-derived content suitable for causal reasoning over multiple possible futures (Adams 2010).

Y-Mazes are simple mazes where there are two concurrently visible choices available to the agent. These measure the preferences of the agent. Usually, one arm of the maze contains a positive reward, and the other is empty or contains a negative reward, but they can also be used to determine preferences between more subtle choices. The testbed includes three different types of Y-Maze (built in different ways), and both left- and right-reward conditions for a series of comparisons between positive and negative and different sized foods (larger food gives more reward) of increasing difficulty. These are a simplified setting for determining an agent's ability to choose between different possible outcomes, arguably a fundamental component of thought.

Detour tasks, radial mazes, and spatial elimination tasks are types of navigation tasks that require either path planning or spatial memory. They can all easily be varied in dimensions such as occluding wall size, number of arms and by the addition of landmarks that guide navigation. In radial mazes it is standard to hide food at the end of each arm so that the agent cannot use sight or smell to determine which arms it has already visited, and needs to form memories to ensure efficient navigation. In the testbed, food is hidden behind two small back-to-back ramps which allow passage, but block vision, in both directions. In these tasks, the time limit is often use to ensure somewhat efficient navigation.

An interesting variation is described in Hughes and Blight (1999), where fish are trained to retrieve food from an 8-arm radial maze. Interestingly, with spatial cues available the fish appear to rely on spatial memory, while in the absence of cues they adopt algorithmic behaviour such as visiting every third arm, which ultimately solves the maze without revisiting any arms. In one experiment of note, a delay is introduced mid-trial (by closing off the arms of the maze), which impairs performance only in the non-cue case. As we do not use any mid-test interventions in the testbed, this is achieved by setting a box to start 'falling' from high above the arena in the initial configuration. The box lands in the centre of the maze midway through the time-limit, creating a barrier and changing the structure of the centre of the maze.

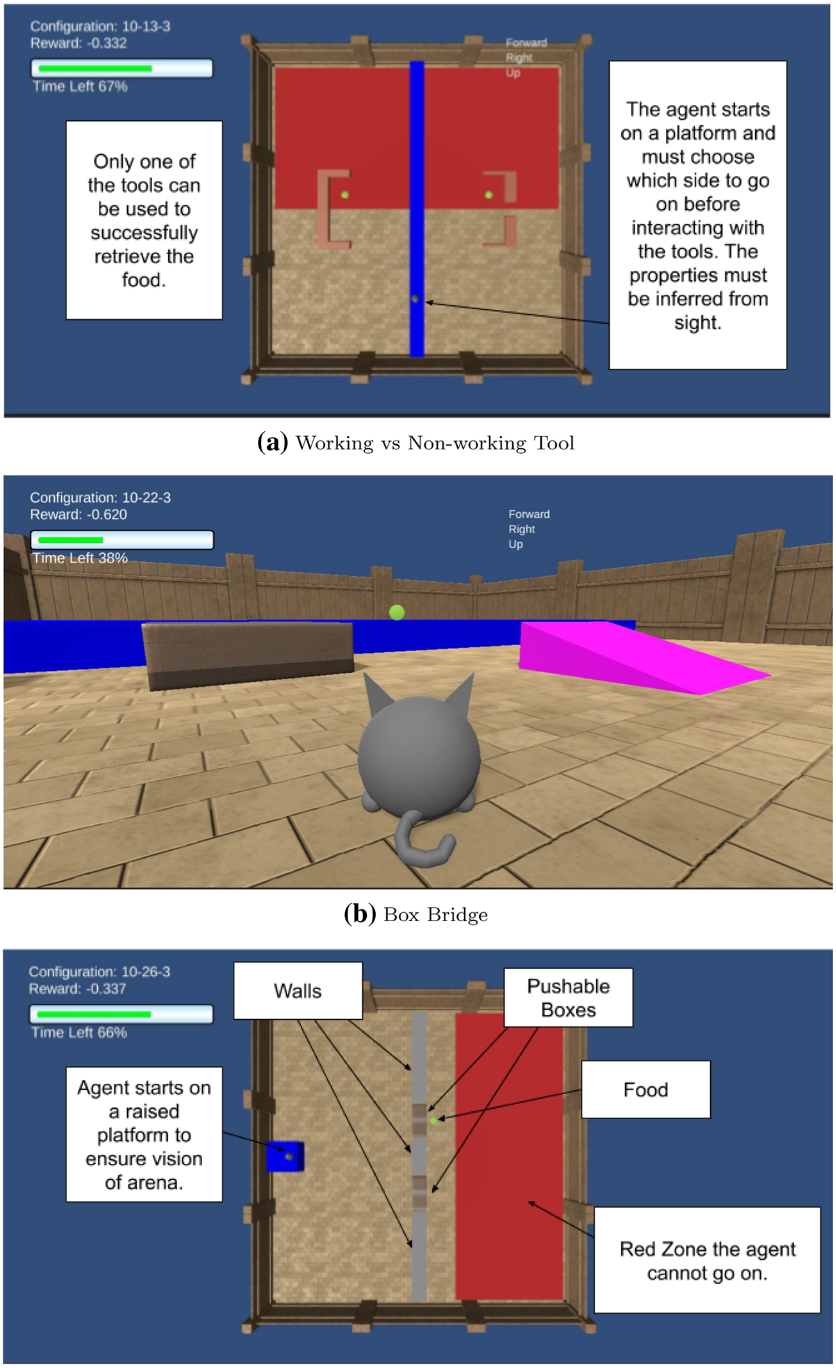
Delayed gratification is the ability to forgo an immediate small reward to obtain a larger future reward. In the example shown in Fig. 2f a green food is placed under a set of rails that are slowly dispensing yellow food. An impatient agent, or one simply trained to retrieve any food available to it, will head to the green food, thus ending the level without the maximum possible score and failing. A successful agent will wait for the yellow food to drop and then collect this before finally also getting the green food. This requires the ability to predict future states and act on preferences over those states to maximise reward. It is certainly possible to come up with strategies for passing that would pass this task without requiring any of the above abilities now that it has been revealed. However, in the context of a secret task this would be next to impossible.

Object permanence tests the understanding that objects still persist when they go out of sight. This requires a model of the world that is not just purely reactive to current inputs, but also takes into account properties that can be inferred, such as the location of an object that has just moved behind a wall. One set of tasks is based on the work of Chiandetti and Vallortigara (2011), which briefly occludes days old chicks from an object they have been imprinted on and hides it in an environment. Figure 2h shows a setup where there is only one place the object could have gone and, in this condition, a statistically significant proportion of chicks correctly infer the correct location.

The internal models category also contains tests that can only be solved robustly with a model of the environment that is maintained in the absence of current sensory information. In these, the lights in the environment are turned off (replacing the inputs with darkness). Tasks include navigating a previously seen path and predicting the final location of food rolling down a ramp.

Numerosity tasks involve counting to choose an option with more rewards. Figure 2i shows a setup where the left side has more rewards than the right side, which can be solved with the ability to compare relative density of food. Other tasks in this category (which comes after the object permanence category) emulate addition tasks form the primate cognition test battery (Herrmann et al. 2007), and involve food moving out of sight to join 'plates' with visible numbers such that success involves choosing, for example, 0 + 3 over 1 + 1.

Tool use tasks test the agent on its ability both to use tools, and to select tools appropriately. Figure 3 shows three examples of tool use tasks. The first is adapted from the string pulling paradigm and tests (Jacobs and Osvath 2015). The agent


Configuration: 10-13-3
Reward: -0.332
Time Left 67%
Forward
Right
Up
Only one of the tools can be used to successfully retrieve the food.
The agent starts on a platform and must choose which side to go on before interacting with the tools. The properties must be inferred from sight.
Configuration: 10-22-3
Reward: -0.620
Time Left 38%
Forward
Right
Up
Configuration: 10-26-3
Reward: -0.337
Time Left 66%
Walls
Pushable Boxes
Food
Agent starts on a raised platform to ensure vision of arena.
Red Zone the agent cannot go on.

(a) Working vs Non-working Tool

(b) Box Bridge

(c) Swing Door Task

Fig. 3 Three of the tool use tasks in the testbed. In each it is possible to get (at least one of) the food by pushing some objects around

must figure out which of the two available tools can be used to pull the food off the red zone. The second is an adaptation of Kohler's famous box and banana test (Köhler 1917). The agent must push the box to form a platform that it can navigate across to get the food. The final is a version of the swing door task from the primate cognition test battery (Herrmann et al. 2007). In this task, there are two pushable boxes (that the agent can see over but cannot climb over). If the agent pushes the box closest to the food, then the food will move into the red zone and cannot be retrieved. The agent has to infer that this will happen and push the other box out of the way to retrieve the food.

I hope I have shown that the testbed contains *relevant* tests and is somewhat *expansive*. The tests can be presented such that they are *unpredictable*, and they are designed to be solvable by our human *exemplars*. The testbed also leads to an *actionable* research plan, with many different difficulty levels and subsets that can be tackled independently at first. If an agent passed all the tests then it could, amongst other things:

- Robustly choose to obtain larger food items over smaller ones.
- Navigate around obstructing objects (even when transparent).
- Find efficient paths through the environment.
- Navigate in a way that takes into account the locations it has previously visited.
- Avoid easily available small rewards when it is possible to infer that a further reward will become available later.
- Act in a way which respects the continued existence of objects that have moved out of sight.
- Maintain a useful model of the environment when inputs are removed.
- Choose containers with more food even when food is added after the initial presentation.
- Use objects as tools to get inaccessible food.
- Choose the appropriate tool when only one has the properties required.

all from interpreting sensory inputs and acting in previously unseen situations. It would be hard to conclude that an agent capable of all these things is not capable of thinking, therefore making it an acceptable marker for the presense of thinking in machines. Of course, it will be a long time before we are able to build an agent that can pass all these tests.

6 Current Progress

In 2019 the Animal-AI testbed was presented as a competition which had over $30,000 worth of prizes and ran for 4 months. More than 60 teams from around the world entered and over 500,000 tests were performed on the private testing server. All of the tests were kept secret during the competition, but the full environment was released, with participants able to create any kind of tests of their own to work with. Information was provided about the types of abilities to be tested as part of the syllabus required to keep the hidden tests fair.

There was no expectation that anybody would solve the harder tasks in the competition. These are incredibly challenging. For example, the Box Bridge task requires an agent to identify that food is unreachable, that a ramp could potentially be used to get to it, and that this solution does not work at the moment but could do so if the conveniently placed box were pushed to the correct place. The agent must then execute the requisite actions within the time limit. Of course, if this task were known it would be relatively easy to train an agent to solve it, just like it is easy to hardcode answers to predictable questions in the Turing Test. The harder tests are included both to ensure that the testbed assessed a wide range of abilities (is *relevant* and *extensive*), and to confirm (for future iterations of the competition) that difficult tests are not susceptible to unexpected 'shortcut' solutions.

The top agents solved 40% of the tasks, roughly corresponding to those that were amenable to associative rules such as 'head towards yellow and green and away from red' with some small improvements. Between them, the top three agents also learned how to use ramps (but not when to use them), how to traverse around transparent walls (though not 100% reliably), and how to escape from very simple mazes. For comparison, I get 100% on the tasks. A purely reactive agent I wrote that acts like a slightly more sophisticated Braitenberg vehicle responsive to green, yellow, and red scores 34%.

The cylinder task is a good example of a task that sits on the edge of current research capabilities (at least within the tight constraints of the competition). The best agent solved 7 of the 9 variations. In the cylinder task, food is placed in a transparent cylinder orientated such that the entrances are perpendicular to the animal or agent. In animal studies this is preceded by trials in which the apparatus is presented with the entrances facing the animal, so that the animal can obtain food by moving directly towards it. To solve the task when the cylinder is angled perpendicularly, the animal must take a longer route that involves moving away from the food in order to eventually retrieve it. This task has been completed successfully by many animals, but even within a single species there can be variation in how well animals perform (MacLean et al. 2014). The top performing agent from the competition would immediately attempt a new route after bumping into the edge of cylinder, and found the entrance in all but the two hardest variations. Other agents would exhibit failure behaviour similar to some animals, repeatedly bumping into the edge of the cylinder until the time ran out.

The earlier categories provide increasingly difficult targest for current research. The best score on the detour tasks was 33% and the best on spatial elimination was 26%, whereas 90% could be considered a plausible cut-off point for robust solutions. As these tasks require navigating around obstacles the Braitenburg agent scores 0 here. Our lab is currently attempting to solve these categories with various research projects including improving curiosity metrics, adding episodic memory to active inference, learning object-level feature representations, automatically generating curriculum of increasing difficulty for training, and combining neural and symbolic approaches. All of these seem promising directions for research, but time will tell which are the most fruitful for developing agents that can solve these particular tasks, and which can then be extended to the more complex tasks on the path towards building thinking machines.

7 Objections

I have suggested that the Animal-AI testbed is a better operationalisation of 'can machines think' than the Turing Test. I now consider some possible objections to this claim. The first set of objections are variations on the claim that a bottom-up approach like this misses some key elements of thought.

1a. You can not ignore language as it is a key part of thinking.

This is a completely reasonable point and I concede that there are elements of our thinking that are impossible without language (Carey 2009). Bermúdez argues that while non-linguistic creatures are capable of sophisticated thinking about the physical environment, there are certain second-order thoughts that require language (Bermúdez 2017). While this may be true, we were originally interested in the question 'can machines think?', and this can be answered positively even if only a subset of human thinking is covered—thinking about the physical environment provides the sufficiency condition.

Furthermore, purely language-based tests are arguably even more limited. It is a common argument that the capacity to think can only develop if an agent is situated in an environment with which it can causally interact (Pearl and Mackenzie 2018), which is not necessary in the standard Turing Test unless faced with a particularly creative judge.

Even if language remains an ultimate goal, research-wise, it seems time to give the bottom-up approach a try. It is more likely that language is scaffolded on our physical understanding of the world (at least in known exemplars of thinking) than the other way round. Lake et al. (2017) similarly propose a primarily non-language based research path for building thinking machines (Lake et al. 2017) as a pragmatic step forward even if, ultimately, language needs to be reintroduced.

1b. The testbed is not *expansive* enough. You cannot ignore non-language property *x* as *x* is a necessary component of thinking.

It is true that the testbed is missing some key problem types, even when only considering experiments from animal cognition. For example, there are no tests that require online learning (learning after or during presentation with a new environment). The properties of all the objects are known in advance, no new objects are introduced for the hidden tests, and no new associations are required to be made at test time. Even so, the testbed still contains configurations that result in previously unseen combinations of objects that have new derived properties. For example, 2f shows 6 wall objects placed such that they form a ramp that food slowly rolls down. Countless such contraptions can be built within the environment.

Another possible missing feature comes from the very shallow embodiment involved in the environment. The agent is effectively a sphere that can push and bounce off objects, but do little more. This was a deliberate decision made in order to simplify the environment such that the easier tests are feasible and also to avoid introducing too many complicating factors at once. I personally believe that shallow embodiment can go a long way as it is the possible interactions available to the agent that are the most important aspect of embodiment and these are still available. That said, as capability improves on the testbed, more complex agents can be introduced. Doing this too soon would make the tests too complex, preventing *actionable* research plans. Other potentially missing elements include social interaction and theory of mind tasks as well as any extended cultural or societal embedding. If these cases are truly necessary then this means that the list of capabilities given at the end of Sect. 5 has not been convincing.

For the purposes of the competition, we decided that these tests—which are already very challenging—offered sufficient evaluation of agents' adaptability. It will always be possible to make the testbed more expansive and, for practical reasons as well as to make research possible, a line must be drawn somewhere. It is hard to imagine what the architecture of an agent that solves all the tasks will look like. As we get closer to this time and more information about the capabilities such an agent has it will become easier to assess if any currently 'missing' components are essential.

1c. There is a difference in kind between humans and animals and only humans are capable of thinking. You cannot build up to thought from solving animal-cognition tasks.

This is a view that has been dormant since Descartes version of it was discredited and replaced by variations of Darwin's mental continuity view. However, there has been having a recent empirically-backed resurgence in recent years (Penn et al. 2008; Ghirlanda et al. 2017). It should be noted that at no point do I need to be committed to the view that animals that pass these tests individually show continuity with human thought. For each of the animal-inspired tests in the testbed there is an example of an animal that has been claimed to pass it at a rate significantly above chance. However, for attribution of an associated cognitive ability in each case requires thorough investigation of the methods used and is not the aim here. The key aspects that are being used are the ideas behind the tests and not the results that may have been achieved or any conclusions that can be drawn from them alone.

More importantly, there is no one type of animal (even great apes) that has passed all the tests. This is largely a contingent fact that they have yet to tested, but ongoing work to translate the AI environment back in to testing for animals either via joystick control or virtual reality will hopefully provide full empirical support to this claim in the future. Nevertheless the full test is not to pass these individually, but to have a single agent that can pass all the tests. Similarly, the goal in the Turing Test is not to answer a single question indistinguishably from a human, but to hold a sustained conversation.

Even so, it could be objected that an extra leap would still be needed to get to a true thinking machine. But even when Penn et al. (2008) argue for a discontinuity between humans and animals they suggest that "this cognitive gap must have evolved largely through incremental, Darwinian processes." This again suggests that the first step to human thought is an agent capable of solving the Animal-AI testbed, from which perhaps only incremental additions will be needed to unlock the full range of human thought. As I have argued throughout, Animal-AI is an alternative, more practical, weak operationalisation that can guide research towards building thinking machines and not a definitive infallible test for their existence. There are many breakthroughs required on the path towards solving the complete testbed. Not least it will have to be possible to build up and reason with object-level properties of the environment from the sensory inputs, something that we currently do not know how to do robustly and that aims to bridge the gap between continuous and symbolic reasoning (Garnelo and Shanahan 2019).

2. Now that you have given all this information, hasn't the testbed failed *unpredictability* and become subject to Goodhart's law?

This is a good point, and it is correct that the now fully released testbed no longer constitutes a good operationalisation because it fails *unpredictability*. Releasing the full set was the best way to allow research on the environment to progress as fast as possible. Of course, with access to all the information it is both much easier to solve (though still will require many innovations), and no longer a good operationalisation. A new set of hidden tests is scheduled for release in 2021, including new scenarios to solve that require reasoning steps that do not exist in the original dataset. This way *unpredictability* can be regained, and at the same time improvements and new types of tests can be added.

3. Better alternatives for bottom-up tests already exist.

In AI, there have been many new test environments developed in recent years. There has also been a move towards more open-ended testbeds that require agents to learn the kinds of skills tested for in Animal-AI. These can be arranged in terms of how unpredictable the tests they contain can be. At one end of the spectrum are environments like CoinRun (Cobbe et al. 2018) and Obstacle Tower (Juliani et al. 2019), which use procedural generation to create previously unseen levels for testing. Often, generated levels are used to train on. Whilst these will not match exactly the ones used for testing, they are generally created by the same generative process making them i.i.d. under the standard usage. At the other end of the spectrum, many modern machine learning environments are designed to be customisable, so that o.o.d. tests can be generated for them by hand. For example, the previously mentioned DeepMind lab (Beattie et al. 2016) can have many levels built for it, as can Microsoft's MALMO research platform, which is built on top of the block-based world of Minecraft (Johnson et al. 2016). The video game definition language (VGDL) takes this a step further and is used in the General Video Game AI competitions to generate novel games to solve (Perez-Liebana et al. 2016). Each of these is a step in the right direction for AI, but none intend to fill the role claimed here for Animal-AI.

Perhaps the best existing candidate for meeting our criteria of an operationalisation of Turing's question is the Abstraction and Reasoning testbed (Chollet 2019). This was designed as a challenge to expose elements missing from current AI approaches aiming at AGI. Two example problems from this testbed are shown in Fig. 4. In this test, a few simple grids of colours are presented with representative input–output pairs that exhibit some rule for transforming input to output. The test is to infer the rule and produce the correct output for a new input after seeing only a few examples. For example, in Fig. 4, a human might articulate the rule for the first problem as 'you place the strip that's underneath on top'. This is easy only when using a natural object-based interpretation of the image. The second test is more complicated. The rule appears to be to place a blue block in a 3 × 3 grid depending on the number of unfilled squares in the 'glass' (note that the grid sizes of each input example are different). Again, this requires a very heavyweight interpretation of the image; even the colours used seem to help. Note that it is unclear what to do in this case when presented with 5 unfilled levels, but the solution only requires 1 so avoids this ambiguity.

Figure 4 shows just 2 of the 800 released samples. The full set contains a large range of different types of abstraction involved making the test *expansive*, at least in terms of abstract reasoning. 200 tests have been kept back as hidden tasks to be used in an ongoing competition in a similar process that I have argued is necessary for *unpredictability*. This testbed was also presented as a competition, with the top result being ≈ 20% success on unseen problems. This kind of figure, like for

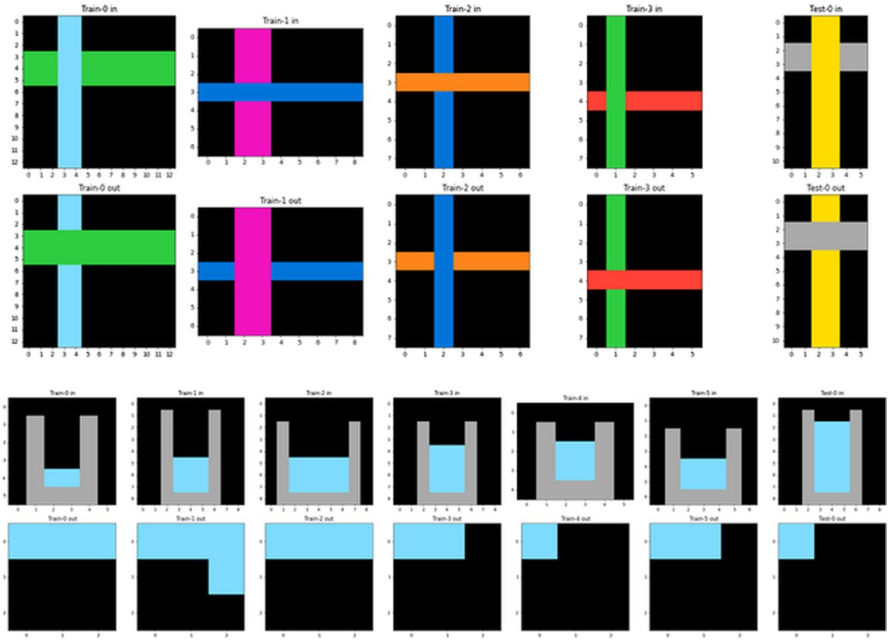


Fig. 4 Two problems from the abstraction and reasoning challenge. After being presented with the first $n-1$ input–output pairs and the n'th input, you must provide the correct output

Animal-AI, suggests the testbed is suitable for leading to *actionable* research. There is a long way to go, but it is possible to get some purchase on the problem. In terms of being solvable by *exemplars*, the testbed may even be a little too complicated as problems such as the second in Fig. 4 are tricky even for humans used to having their thinking abilities tested. In Sect. 2 I argued that the test should aim for the lowest common denominator amongst exemplars of the property at hand.

The *relevance* of this testbed depends on the possibility for these abstract aspects of thinking to be isolated from the kinds of embedded interaction I have suggested are necessary underpinnings of thought. It is relevant to note that the winning entry on this testbed was a program that searched through combinations of a list of 142 common transformations to the grids that was compiled by hand from working through the released examples. It is unclear just how far such a method can get and it is unlikely that even the most sophisticated version of this approach would be able to 'solve' the testbed. Nevertheless, there is a strong difference between how a human solves the tests (a process which includes 'thinking') and how an algorithm might. It is less clear that there will be such a difference between a hypothetical agent that solves Animal-AI and a human solving the same problems. The Abstraction and Reasoning challenge was intended as a measure of intelligence, and not of 'thinking' in machines, but I present this discussion to further motivate the need for a bottom-up approach with tests built up from sensory information.

4. This is a very bio-centric account.

Finally, I will briefly address the concern that this is a very bio-centric account of what it takes to be a thinking machine. Should I have been attempting to give a general intelligence test, then this would be a fair objection. However, that was not the intention here, and neither do I believe was it the intention in Turing (1950), despite the misleading title of his article. Whilst abstract definitions and measures of intelligence can be given (Legg and Hutter 2007; Hernández-Orallo 2000), it is also possible to link the concept to an entities ability to interact with its environment (Wang 2019). As we move from intelligence to thought, it becomes harder to keep the discussion purely abstract. I expect we will only ever be able to recognise the capacity for thought in systems that can interact with their environment in ways that we can interpret as purposeful in some sense. This makes it sensible to test for it in environments that share properties (such as their intuitive physics) with our own and to rely, as I have done in this paper, on the properties of our only known exemplars as a guide.

8 Conclusion

I started with the question 'can machines think?'. Like Turing, I quickly discarded it in favour of an empirically verifiable test, designed to be closely related to the original question. Because of the ill-defined nature of the initial question, and because it is a hypothetical question about future machines I argued that this move is one of *weak operationalism*. The original question is replaced with an imperfect empirical measure that picks out systems that are thinking machines. As the move is imperfect, the measure must have pragmatic value whilst still being as accurate as possible.

To be accurate I argued that the test needs to be *relevant, expansive, and solvable by exemplars*. To be useful it needs to be *unpredictable* and lead to *actionable* research. Turing's test does not score well against these conditions, primarily because it sets the bar too high. It is important that as many exemplars as possible can solve the test otherwise it will be too exclusive. If we assume that our set of exemplars is all humans above a certain age, then a reasonable set of tests that they can be expected to be able to solve are those that have been found to be solvable by non-human animals. Of course, there are tests that non-human animals can pass that humans cannot, and these are excluded.

The Animal-AI testbed combines an *extensive* set of animal cognition tasks into an environment that is designed to be amenable for research in AI. I showed that the test holds up to the research community's attempts to solve using current methods in AI through an open competition. If the testbed had been solved here then this would likely be due to a flaws in its design and not due to the unexpected existence of thinking machines. Fortunately, this did not happen.

If the goal of AI is to build general intelligence (AGI), then it is unclear if the route it should go down involves solving the tasks in Animal-AI. Solving the tasks presented here certainly does not directly help with solving games like Go or chess or to build systems capable of working with large sequences of symbols. On the other hand, if the goal is to build thinking machines, then solving Animal-AI is not only a sensible goal, but one that will likely also teach us more about what it is like to be thinking machines ourselves. It may even turn out that this is a necessary component of AGI, and that the progress we've seen to date on narrow intelligence tasks is held back by the inability to solve the simple physical tasks that many animals are capable of.

After 70 years we can see a little further. There is still plenty that needs to be done.

Acknowledgements This work was supported by the Leverhulme Centre for the Future of Intelligence, Leverhulme Trust, under Grant RC-2015-067. The Animal-AI Olympics competition was supported by Amazon AWS, GoodAI, Unity, the WBA-Initiative and AIJ.

Compliance with ethical standards

Conflicts of interest The authors declare that they have no conflict of interest.

References

Adams, F. (2010). Why we still need a mark of the cognitive. *Cognitive Systems Research*, *11*(4), 324–331.

Akagi, M. (2018). Rethinking the problem of cognition. *Synthese*, *195*(8), 3547–3570.

Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., & Ribas, R., et al. (2019). *Solving rubik's cube with a robot hand*. arXiv preprint arXiv:191007113.

Allen, C. (2014). Models, mechanisms, and animal minds. *The Southern Journal of Philosophy*, *52*, 75–97.

Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). *Invariant risk minimization*. arXiv preprint arXiv:190702893.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2019). *Emergent tool use from multi-agent autocurricula*. arXiv preprint arXiv:190907528.

Beattie, C., Leibo, JZ., Teplyashin, D., Ward, T., Wainwright, M., Küttler, H., Lefrancq, A., Green, S., Valdés, V., & Sadik, A., et al. (2016). *Deepmind lab*. arXiv preprint arXiv:161203801.

Bellemare, M. G., Naddaf, Y., Veness, J., & Bowling, M. (2012). *The arcade learning environment: An evaluation platform for general agents*. CoRR abs/1207.4708, arXiv:1207.4708.

Beran, M. J. (2002). Maintenance of self-imposed delay of gratification by four chimpanzees (pan troglodytes) and an orangutan (pongo pygmaeus). *The Journal of General Psychology*, *129*(1), 49–66.

Bermúdez, J. L. (2017). *Can nonlinguistic animals think about thinking?*., The Routledge Handbook of Philosophy of Animal Minds London: Routledge.

Beyret, B., Hernández-Orallo, J., Cheke, L., Halina, M., Shanahan, M., & Crosby, M. (2019). *The animal-ai environment: Training and testing animal-like artificial cognition*. arXiv preprint arXiv:190907483.

Block, N. (1981). Psychologism and behaviorism. *The Philosophical Review*, *90*(1), 5–43.

Bluff, L. A., Troscianko, J., Weir, A. A., Kacelnik, A., & Rutz, C. (2010). Tool use by wild new caledonian crows corvus moneduloides at natural foraging sites. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1686), 1377–1385.

Buckner, C. (2015). A property cluster theory of cognition. *Philosophical Psychology*, *28*(3), 307–336.

Carey, S. (2009). *The origin of concepts*. Oxford: Oxford University Press.

Chiandetti, C., & Vallortigara, G. (2011). Intuitive physical reasoning about occluded objects by inexperienced chicks. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1718), 2621–2627.

Chollet, F. (2019). *The measure of intelligence*. arXiv preprint arXiv:191101547.

Clark, A. (2015). *Predicting peace: The end of the representation wars. Open MIND*. Frankfurt a. M.: MIND Group.

Cobbe, K., Klimov, O., Hesse, C., Kim, T., & Schulman, J. (2018). *Quantifying generalization in reinforcement learning*. arXiv preprint arXiv:181202341.

Crosby, M., Beyret, B., Shanahan, M., Hernández-Orallo, J., Cheke, L., & Halina, M. (2020). The animal-ai testbed and competition. In *Proceedings of Machine Learning Research*

Deng, J., Dong, W., Socher, R., Li, LJ., Li, K., Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE* (pp. 248–255).

Dennett, D. C. (1984). Can machines think? In M. G. Shafto (Ed.), *How We Know*. New York: Harper & Row.

Epstein, R., Roberts, G., & Beber, G. (2009). *Parsing the turing test*. Berlin: Springer.

Farrar, B. G., Ostojić, L. (2019). *The illusion of science incomparativecognition*. PsyArXiv October 2

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, *104*(1–2), 137–160.

Garnelo, M., & Shanahan, M. (2019). Reconciling deep learning with symbolic artificial intelligence: Representing objects and relations. *Current Opinion in Behavioral Sciences*, *29*, 17–23.

tGeirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. arXiv:2004.07780.

Ghirlanda, S., Lind, J., & Enquist, M. (2017). Memory for stimulus sequences: A divide between humans and other animals? *Royal Society Open Science*, *4*(6), 161011.

Goodhart, C. A. (1984). Problems of monetary management: The UK experience. *Monetary theory and practice* (pp. 91–121). Berlin: Springer.

Guss, WH., Codel, C., Hofmann, K., Houghton, B., Kuno, N., Milani, S., Mohanty, S., Liebana, DP., Salakhutdinov, R., & Topin, N., et al. (2019). *The minerl competition on sample efficient reinforcement learning using human priors*. arXiv preprint arXiv:190410079.

Harnad, S. (1991). Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines*, *1*(1), 43–54.

He, K., Zhang, X., Ren, S., & Sun, J. (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1026–1034).

Hennefield, L., Hwang, H. G., Weston, S. J., & Povinelli, D. J. (2018). Meta-analytic techniques reveal that corvid causal reasoning in the aesop's fable paradigm is driven by trial-and-error learning. *Animal Cognition*, *21*(6), 735–748.

Hernández-Orallo, J. (2000). Beyond the turing test. *Journal of Logic, Language and Information*, *9*(4), 447–466.

Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science*, *317*(5843), 1360–1366.

Hughes, R. N., & Blight, C. M. (1999). Algorithmic behaviour and spatial memory are used by two intertidal fish species to solve the radial maze. *Animal Behaviour*, *58*(3), 601–613.

Hutter, M. (2000). *A theory of universal artificial intelligence based on algorithmic complexity*. arXiv preprint cs/0004001.

Jacobs, I. F., & Osvath, M. (2015). The string-pulling paradigm in comparative psychology. *Journal of Comparative Psychology*, *129*(2), 89.

Jelbert, S. A., Taylor, A. H., Cheke, L. G., Clayton, N. S., & Gray, R. D. (2014). Using the Aesop's fable paradigm to investigate causal understanding of water displacement by new caledonian crows. *PLoS ONE*, *9*(3), e92895.

Johnson, M., Hofmann, K., Hutton, T., Bignell, D. (2016). The malmo platform for artificial intelligence experimentation. In *IJCAI* (pp. 4246–4247).

Juliani, A., Berges, V., Vckay, E., Gao, Y., Henry, H., Mattar, M., & Lange, D. (2018). *Unity: A general platform for intelligent agents*. CoRR abs/1809.02627, arXiv:1809.02627.

Juliani, A., Khalifa, A., Berges, VP., Harper, J., Teng, E., Henry, H., Crespi, A., Togelius, J., & Lange, D. (2019). *Obstacle tower: A generalization challenge in vision, control, and planning*. arXiv preprint arXiv:190201378.

Köhler, W. (1917). *Intelligenzprüfungen an anthropoiden. 1.-. 1, Königl*. akademie der wissenschaften.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253.

Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and Machines*, *17*(4), 391–444.

Lind, J. (2018). What can associative learning do for planning? *Royal Society Open Science*, *5*(11), 180778.

Lloyd Morgan, C. (1894). *An introduction to comparative psychology*. London: W Scott.

MacLean, E. L., Hare, B., Nunn, C. L., Addessi, E., Amici, F., Anderson, R. C., et al. (2014). The evolution of self-control. *Proceedings of the National Academy of Sciences*, *111*(20), E2140–E2148.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Moravec, H. (1988). *Mind children: The future of robot and human intelligence*. Cambridge: Harvard University Press.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. New York: Basic Books.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*(2), 109–130.

Penn, D. C., & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 731–744.

Perez-Liebana, D., Samothrakis, S., Togelius, J., et al. (2016). The 2014 general video game playing competition. *IEEE Transactions on Computational Intelligence and AI in Games*, *8*(3), 229–243.

Proudfoot, D. (2011). Anthropomorphism and ai: Turing's much misunderstood imitation game. *Artificial Intelligence*, *175*(5–6), 950–957.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018). Machine theory of mind. In *International Conference on Machine Learning* (pp. 4218–4227).

Racanière, S., Weber, T., Reichert, D., Buesing, L., Guez, A., Jimenez Rezende, D., Puigdomènech Badia, A., Vinyals, O., Heess, N., Li, Y., Pascanu, R., Battaglia, P., Hassabis D., Silver, D., Wierstra, D. (2017). Imagination-augmented agents for deep reinforcement learning. In: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., (pp. 5690–5701). http://papers.nips.cc/paper/7152-imagination-augmented-agents-for-deep-reinforcement-learning.pdf.

Recht, B., Roelofs, R., Schmidt, L., Shankar, V. (2019). *Do imagenet classifiers generalize to imagenet?* arXiv preprint arXiv:190210811.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756.

Shapiro, L. (2019). *Embodied cognition*. Abingdon: Routledge.

Shaw, R. C., & Schmelz, M. (2017). Cognitive test batteries in animal cognition research: Evaluating the past, present and future of comparative psychometrics. *Animal Cognition*, *20*(6), 1003–1018.

Shettleworth, S. J. (2009). *Cognition, evolution, and behavior*. Oxford: Oxford University Press.

Shevlin, H., & Halina, M. (2019). Apply rich psychological terms in AI with care. *Nature Machine Intelligence*, *1*(4), 165–167.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., et al. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354–359.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.

Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence*, *10*(2), 1–37.

Williams, D. (2018). Predictive processing and the representation wars. *Minds and Machines*, *28*(1), 141–172.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.