Master Thesis on Brain and Cognition

Universitat Pompeu Fabra

Exploring the Underlying Representations of Infant Cognition using Connectionist Networks

Mark McGuill

Supervisor: Dr. Luca L. Bonatti

Co-Supervisor: Kinga Anna Bohus

July 2018

upf. Universitat
Pompeu Fabra
Barcelona

Abstract

Recent results in infant cognition research suggest that infants can possess early elementary logical abilities. The nature of the representations underlying these abilities is still controversial. Here we begin studying them by developing computer simulations of such results. Our initial strategy is to exploit earlier work on connectionist neural networks and their application to classic cases of infant cognition. We propose using these devices, despite their well-documented inherent limits, as yardsticks. That is, we seek to use them to probe the minimal character of representation required to elicit behaviour similar to that of young human infants when presented with cognitive tasks. In particular some of the tasks we are most interested in are those that potentially involve logical inferences. We present three experiments and their results with discussion on their shortcomings, and potential improvements. In experiment 1, we test an extremely rudimentary representation, which nonetheless acts as a proof of concept and as a base for comparison for later experiments. In experiment 2, we present a much more articulated and realistic representation, whose performance is quite distinct from that of our first experiment. In experiment 3, we apply an incoherent and physically impossible sequence of events to the network used in experiment 2, probing its 'cognitive' characteristics. We conclude with suggestions to further develop possibly promising lines of inquiry.

Acknowledgements

I would very much like to extend my sincerest gratitude to Dr. Luca L. Bonatti, Kinga Anna Bohus, and Ana Martín who have been more than generous with their time and advice from day one. I'd also like to thank all the faculty and staff at UPF for making the university such a pleasant place to learn. To my classmates for their support and camaraderie throughout the year and for their stoic patience, my parents, who are still wondering when their son is ever going to settle down, thank you and some day...

Contents

1 Introduction

> *"There is a gap between the mind and the world, and (as far as anybody knows) you need to posit internal representations if you are to have a hope of getting across it. Mind the gap. You'll regret it if you don't."*
>
> – Jerry A. Fodor

Infants from a very early age demonstrate many advanced capabilities which appear to be universal. In particular, they show knowledge of basic physical principles about the world such as *continuity* and *solidity*, often grouped under the broader term *object permanence* or *object concept* (Baillargeon, 1993). There has been much research in this area, particularly on object individuation (Xu, 1999), and on the role of labels and language (Xu & Carey, 1996). This work suggests a rich and subtle developing inner mental world but the character of representational structure being used is still an open question.

One possibility is that the richness of this inner mental world is really the result of extremely minimal representational structures. This is the stance taken by the connectionist framework, positing that these capabilities arise from weak internal representations supported by strong network architectures. From this point of view infants exhibit such an apparently rich inner world, not because they possess complex and articulated structural representations, but because they can better exploit their experience. Their behaviour is learned from their experience, e.g. from environmental stimuli and feedback. This learning is performed by simple configurations of elementary nodes that can take advantage of the associations between stimuli of varying kinds. If complexity exists, this lays in the configuration of the network (e.g. more layers, better node structure, improved training algorithms).

Investigations into the possibility of developing connectionist neural networks with these abilities began in the mid 1980s, with the resurgence of interest in connectionism. This was largely due to the discoveries of back-propagation and newer more powerful network types, especially that of the Parallel Distributed Processing (PDP) networks (Rumelhart & McClelland, 1986). These networks transcended the limits of earlier technologies such as the Perceptron (Rosenblatt, 1958).

A paradigm example of this work is due to Munakata(Munakata, 1998) who worked on the classic Piagetian A$\overline{B}$ task among others. In this task, an object is hidden in a location (A). Children repeatedly reach for it and successfully retrieve it from A several times before the experimenter hides it in location B, in full view of the infant.

While an adult would at this point reach for location B, infants before 7 months of age reach for it instead at A, even when they keep their eyes focused on the correct B location. This puzzling behaviour (technically called *perseverative reaching*) has been attributed to the infants' inability to inhibit a previously successful retrieval of an object; an object whose identity and location they can nevertheless perfectly represent. In her work, Munakata found certain interesting parallels in the way the networks behaved with the behaviour of infants. She experimented with varying representations and network architectures to simulate the infants behavioural idiosyncrasies. The interest of her proposal lies in the fact that while infant behavior in this task has generally been attributed to the fact that they possess a complex inner representation of objects, she tried to show that the same behavioral outcomes could be predicted by reasonably simple networks and how they process an abstract description of the input that a child could have received during the tasks (that is, the sequence of repeated successful retrieving events).

Despite the bold attempt, the results from this line of work were modest. The program itself has become mostly dormant for a number of reasons. For one, the network simulations diverged in significant ways from infant behaviour, particularly with regard to the familiarisation sequence.

Another more significant issue with this program arose due to Fodor & Pylyshyn. Their critical analysis of the connectionist framework (J. A. Fodor & Pylyshyn, 1988), like that of Minsky & Papert (Minsky & Papert, 1969) a generation before them, brought to light the limits of such a theoretical framework. The core point made in their analysis is that while connectionism & classical cognitive science are both representational systems, connectionism does not provide a framework for fluid combinatorial and syntactic organisation of these representations. It does not use a 'symbol-level' representation. These networks fail to capture basic cognitive symmetries of a *language of thought* (J. Fodor, 1975). To use their rather illustrative example, no one who understands "John loves Mary" can fail to understand "Mary loves John", or, more importantly, that they express very different states of the world. These are entirely different expressions with no underlying structural similarity in a connectionist net. Another limitation is that they fail to capture constituent or part/whole structures among representations. Broader philosophical critiques of empiricist philosophy also emphasise the shortcomings of this general approach to cognition (Chomsky, 1967). The limitations stemming from this conceptual commitment are seen by some as a fatal flaw.

Research Goal These limits prevented the connectionist framework from achieving its maximal objectives, namely, to show that simple networks deprived of rich inner structure are plausible models of cognition. However, even though the connectionist grand plan was not realized, there is a different fashion by which their networks could be used in cognitive research. Following the basic principles & guidelines laid down by the work of Munakata, the networks can be used as tools to probe the minimal representational structure required to achieve some of the basic capacities that infants show from a very young age. In this way connectionist simulations become a way to apply Occam's razor to the explanation offered of certain cognitive phenomena: even if they fail to account for them, they can give us indications of how rich a mental representation must be to overcome the networks shortcomings. This is how we intend to use network simulations in this work.

Specifically we are interested in the recent results from (Cesana-Arlotti et al., 2018). Our interest in this work is that, according to its authors, it reveals natural capacities such as logical deduction. Given the potential complexity of such a capacity, we believe it is a prime candidate for testing just what a network can do and just what kind of representational complexity is required for it to do so in an infant like way.

2 Methods

> *"The original question, 'Can machines think?' I believe to be too meaningless to deserve discussion."*
>
> – Alan Turing, *Mechanical Intelligence*

The aim of this work is to explore under what conditions a connectionist network can reproduce computational analogs to infants' behaviors during a task potentially involving a mental logical inference. We will explain the details of the results below. For starters, it is important to keep in mind that there are two main ways to alter connectionist networks to probe the complexity of a cognitive task: the architecture can be changed, or the input/output representation presented to the network for training and test can be changed. While not entirely orthogonal (e.g. the input and output nodes must be tailored to the representation), the hidden layers and feedback connections inter alia can be varied independently of the representation. In this work, which should be seen as the first step of a broader inquiry, only the representations are altered, keeping the network architecture static. The architecture is described in detail in Appendix A. This simplifies our approach a little.

We will focus on one particularly important recent result from the infant cognition literature (Cesana-Arlotti et al., 2018) and construct different plausible representations to apply to our networks. Measurements are taken of the effectiveness of the network at simulating infants' behavior (as described by the authors), under the heavy simplifications that the current approach involves. Our aim is to try to determine the character of the representations required to perform these capacities. In particular, our main target will be the infants' surprise reaction at the outcome of a scene which is logically inconsistent with a potential inference about the location of objects.

The representations are based on the scenarios (described below) tested in (Cesana-Arlotti et al., 2018). These form the basis of the training and testing sets for our neural networks.

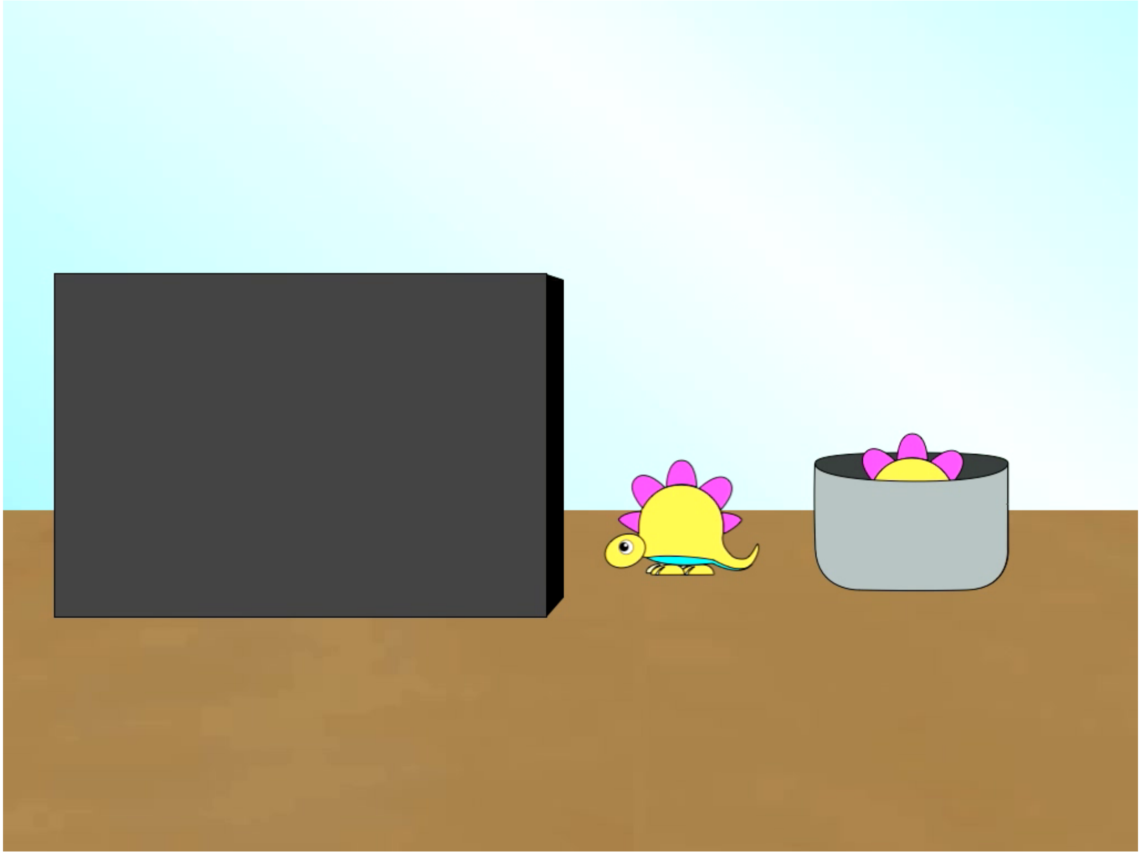


Figure 1: A typical snapshot from one of our scenarios

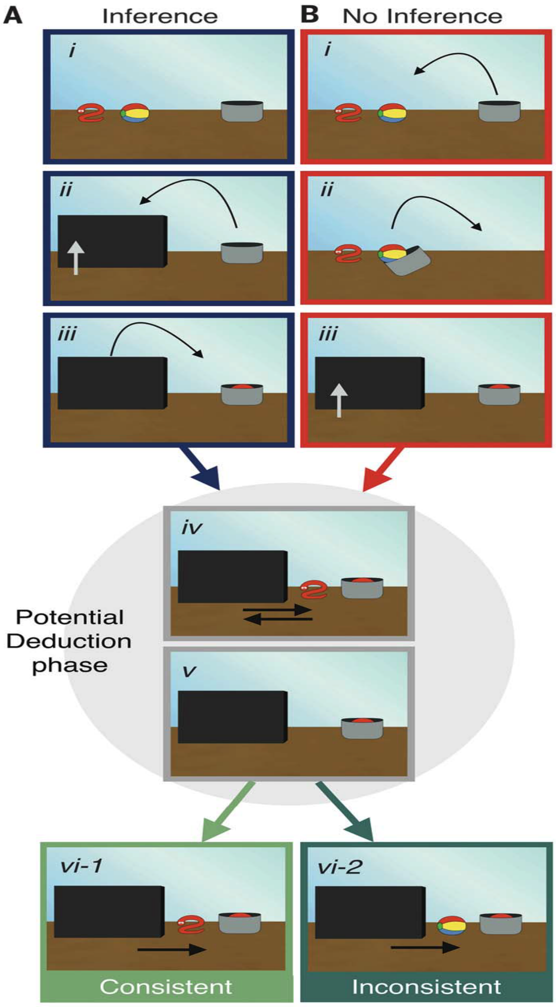


Figure 2: Scenario evolution and the potential deduction phase.

2.1 The Scenarios

The scenarios normally consist of a pair of (colourful and toy like) objects interacting with very basic but dynamic elements (e.g. an occluding wall, a scooping cup) of their environment in straightforward and intuitive fashions. An example of a typical scene is presented in Figure 1. Typically the scene opens with the scooping cup present on the right hand side. The occluding wall is not usually initially visible, though it will be animated in later.

The scenario then evolves with the pair of objects, call them object A and object B, introduced one after the other. They have an important invariant, which is that their top halves are always identical. This means if the bottom halves are hidden (by means of the occluding wall, or scooping cup) there is no way to tell A from B, a key part of the inference process to be examined.

There are two basic variations on the development of a scenario. They are called the Inference and No-Inference variations and they are described briefly below, and also presented more graphically in figure 2.

2.1.1 Inference Required Variation

This variation begins in the usual fashion, with two objects present on the scene. An occluding wall ascends obscuring both objects and the scooping cup moves behind the wall to scoop one of the objects and move it to the right hand side. Once in this position, it is impossible to tell which of the objects was scooped by the cup from behind the wall (because the bottom halves are occluded). The next step begins what is called the *Potential Deduction Phase*. The object that had remained behind the occluding wall slides out, revealing itself. At this point it is possible to deduce the identity of the object in the cup, e.g. dinosaur behind the wall therefore flower in the cup.

2.1.2 No Inference Required Variation

In the no inference variation, the cup scoops an object in plain sight, before the occluding wall comes up. This means the infant (or our network) does not need to draw any inference about the object locations, it has all the information necessary to know where object A and object B are located. The scenario will continue with the occluding wall raising as before, and the emergence of the object from behind the wall.

2.1.3 Consistent & Inconsistent Endings

There are two different ways a scenario can end. The consistent ending is when the expected object appears from behind the wall in line with expectations. The inconsistent ending has an object which could not have been behind the wall appear from behind it. These endings were used to test if the infant had made the correct inference (in the inference variation) and was surprised (*violation-of-expectation*) by the inconsistency. These endings can be seen in context in figure 2.

2.1.4 Familiarisation Scenarios

Familiarisation scenarios were used both in the original infant experiments (Cesana-Arlotti et al., 2018) and our own neural network training. These are simpler presentations of the objects in the scene, performing simple motions, with no emphasis on complex world logic, reasoning or inference. The networks are trained on these simpler scenarios only, in the same way that infants only view familiarisation videos before they are tested. Importantly, the familiarisation scenarios were all different, and did not contain any phase where making an inference was necessary or desirable. In these two aspects, our task and the nature of our simulations are sharply different from previous work, such as Munakata's simulations of the A$\overline{B}$ task. In that work, the network was trained with repeated identical experiences which matched the tasks required for the test phase. In contrast, and by principle, our case provides no such facilitation to the network. The training set is always composed of partial and inchoate segments of a potentially more complex scenario, never the full scenario on which testing is based.

2.1.5 Measuring Performance

In the original experiments (Cesana-Arlotti et al., 2018) infant looking time is recorded in the outcome phase (the final frame of the scene with consistent or inconsistent endings), in a *violation-of-expectation* paradigm. This is used as a measure of surprise at what has transpired in the movies.

The question then arises, how should we measure surprise in our networks? The answer is not so simple and each choice has its advantages and disadvantages. We decided to use the standard measure of Mean Squared Error (MSE) or 'loss' as a measure of accuracy and 'surprise'. The greater the loss the worse the performance, an analog of longer looking time or surprise. Similarly, the lower the loss, the better the performance. This idealization could provide an analog of shorter looking time and therefore less surprise. Notice, however, that in this way we disregard the fact that some parts or aspects of the scene are more salient for measuring cognitive performance than others. For example, the potential deduction phase has a particular relevance for infants solving their logical reasoning task; this phase of the scenario is more important than the initial introduction of Object A and B, a necessary but quotidian dimension of the task requiring no special logical calculus. In using MSE as our measure of surprise, we average over performance at all points in the scenario equally, losing detail where it may be most enlightening. This being

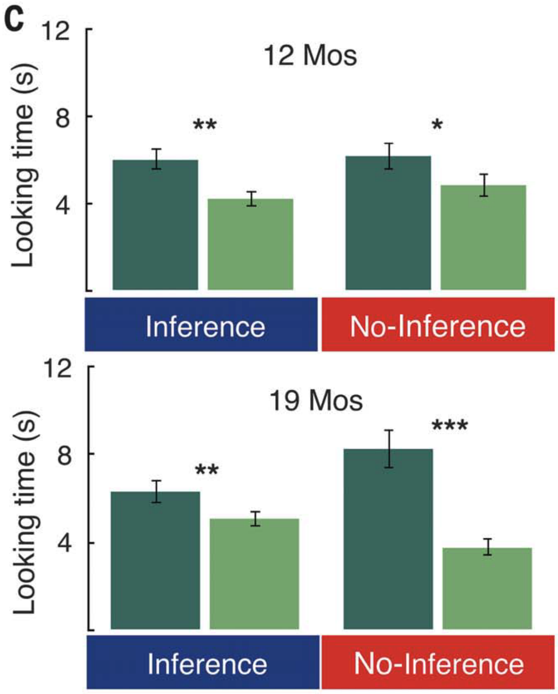


Figure 3: Infant Looking Time (Consistent (light-green) vs Inconsistent)

said, MSE strikes us as a reasonable first-pass attempt at finding a plausible analog to infants' surprise. We will revisit this compromise in the General Discussion.

NB: The movies can be found online in the supplementary material of the referenced work.

2.2 Scenario Representation Schemata

The representation of the movies as a sequence of discrete events appealed to us as the most natural format for our data and this suggested that a recurrent neural net architecture was most suitable for our experiments. This architecture is further described in Appendix A.

To describe a scenario a simple sequence of events is used. These events might be thought of as the important frames from the original movies. Each event contains a representation of a scene described briefly above and shown in figure 2, and these events are input to the network sequentially. The differences between the representations in each event are limited to those most salient for the inference process that we seek to test, e.g. *The cup has scooped an object* or *The occluding wall has appeared.*

In line with our research goal, we start with a basic model, and we then proceed to gradually increase the complexity of the representation to examine what effect it has on the networks performance. This way a progression of representations is built that can be used to measure the character of network performance against the amount or quality of representations used. What follows is a description of the experiments performed and their respective representational schemata.

2.3 Experiment 1 - A Consistency Detection Schema

"Colorless green ideas sleep furiously"

– Noam Chomsky, *Syntactic Structures*

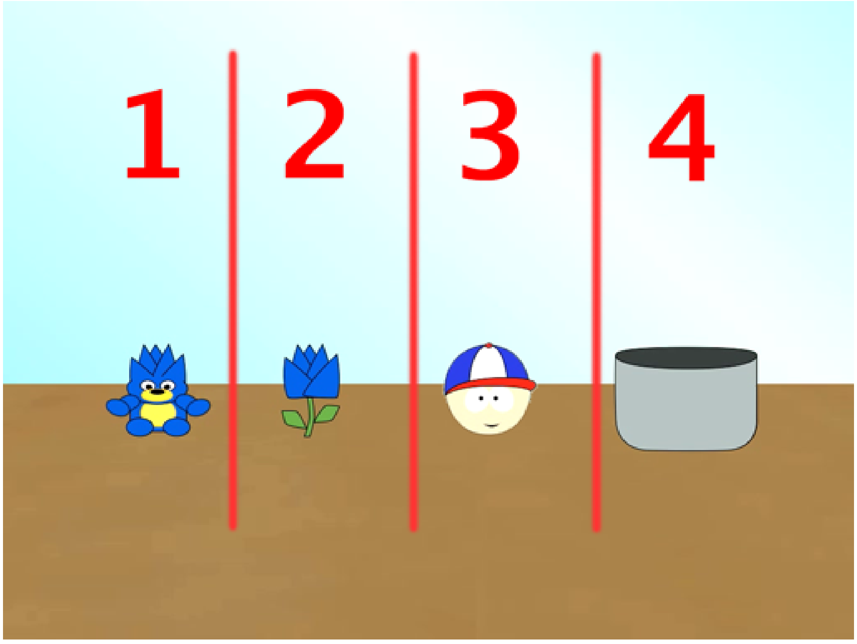
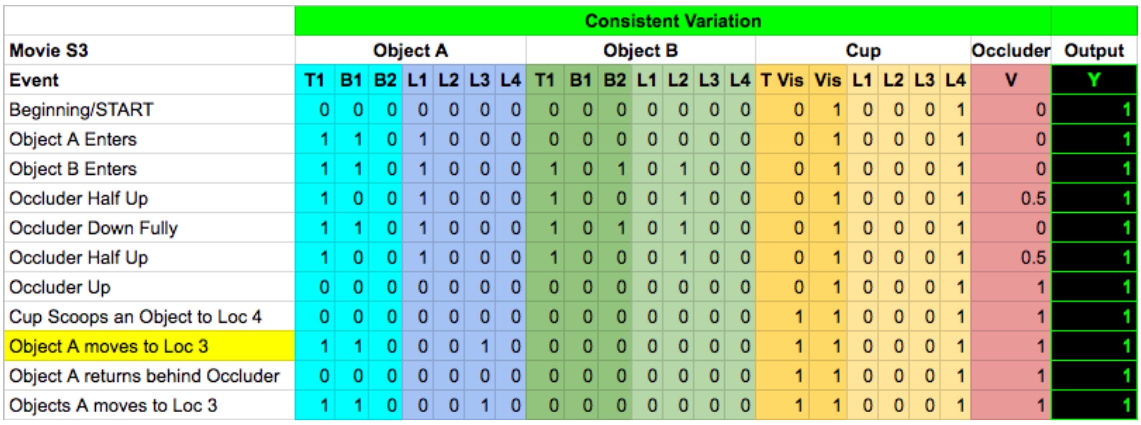


Figure 4: The four possible locations for objects

Material The first model is a consistency violation detector. This is a simple representation with a basic internal structure and a direct mapping from scenes to input data. It was selected in line with the aim of building incrementally from basic and limited towards more complex and articulated representations. Any elements or unnecessary detail which we felt could interfere with the inference process were removed. The representational system included the four basic elements of each scene. Spatial location is broken down into four positions (1-4 from left to right) as shown in figure 4. The objects can theoretically appear in any of these locations.

The output of this design is a single bit representation which indicates whether the scene is consistent or inconsistent. This corresponds to the infants 'surprise' response when something occurred that was inconsistent with her world knowledge. If this representation is a plausible proxy for the infants representation system we should expect that the network will show a 'surprise' response similar to the infants in the same situations.



	Consistent Variation																					
Movie S3	**Object A**							**Object B**							**Cup**						**Occluder**	**Output**
Event	**T1**	**B1**	**B2**	**L1**	**L2**	**L3**	**L4**	**T1**	**B1**	**B2**	**L1**	**L2**	**L3**	**L4**	**T Vis**	**Vis**	**L1**	**L2**	**L3**	**L4**	**V**	**Y**
Beginning/START	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1
Object A Enters	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1
Object B Enters	1	1	0	1	0	0	0	1	0	1	0	1	0	0	0	1	0	0	0	1	0	1
Occluder Half Up	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0.5	1
Occluder Down Fully	1	1	0	1	0	0	0	1	0	1	0	1	0	0	0	1	0	0	0	1	0	1
Occluder Half Up	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0.5	1
Occluder Up	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1
Cup Scoops an Object to Loc 4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	1
Object A moves to Loc 3	1	1	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	1
Object A returns behind Occluder	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	1
Objects A moves to Loc 3	1	1	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	1

Figure 5: Sample input and expected output test data

Figure 5 presents the schematic representation of the event sequences corresponding to the consistent sequence shown in figure 2 (Movie S3 of Cesana-Arlotti et al.'s supplementary material). The timeline runs chronologically, each row representing a further evolution of the scene. The columns describe different physical aspects of the scene.

A Brief Description of our data A full description of the representation used may be found in Appendix B, but briefly there are four main component parts. The two objects, the cup and the occluder. The output column is what we expect the network to output and is used for our (MSE) loss measurement. Each object has four locations it can be present in, corresponding to L1, L2, L3, and L4 columns. It also has 3 visibility components, T1, B1, and B2. T1 indicates to the network that the top half of this object is visible (1) or not (0). B1 and B2 indicate that the bottom half of this object is again visible or not, but this time B1/B2 indicates that the bottom half is of one type or another. The scooping cup has four location elements that operate in the same way as the objects' location elements. The other two components, TVis indicating that the top half of one of the objects (A or B) is visibly protruding from the cup, i.e. that the cup contains an object. Vis indicates whether or not the cup is visible on the scene. Finally the occluder is simply specified by a single element indicating whether the wall is down (0), half up (0.5) or fully up (1).

As presented in Figure 5, the timeline starts with an empty scene ('Beginning/START'). Next, Object A enters the scene, followed by Object B entering. The occluder then raises up half way, before retracting fully down. The occluder then comes back up, first half then fully, completely occluding the objects. Next, the scooping cup which had been resting at location 4, scoops an object. We have now reached the *potential deduction phase* highlighted in yellow in the Event column. Object A moves to location 3, allowing the infant or our network to infer the identity of the object in the scooping cup, and more obviously the identity of the object behind the wall. Object A returns behind the occluder, before finally reappearing. In the inconsistent variation of this scenario, at this final stage a different object appears from behind the wall, violating expectations.

A Note on Object Individuation An important further simplification is to deliberately exclude individual *object identity.* That is to say, it is not indicated to the network which particular object (e.g. flower, dinosaur, umbrella or snake) is on the scene, only that there are two different objects present in the scene, A and B. We believe this simplification can be justified due to findings concerning object individuation, particularly the *object first hypothesis* discussed in Xu & Carey 1996 (Xu & Carey, 1996). This research provides evidence that infants first learn to individuate objects based solely on spatio-temporal boundaries, developing later capacities to individuate objects based on other properties (shape, colour, function, and other more complex sortals). Infants do however have a more basic capacity, that of tracking the number of objects present, known as *numeric identity* which is captured in our representations explicitly. Object individuation could be an important factor for infants, and would possibly change the behaviour of our network and so it recommends itself as an important later addition to our progression of representations.

Procedure The experiment is run using a Jupyter Notebook (Python) with the help of the machine learning library PyTorch. Cesana-Arlotti et al tested approximately 24 participants in a typical experiment. Accordingly we aimed to test at least this amount. Given the ease with which participants can be instantiated in our software, we decided to use a group of one hundred (n=100), a generous but manageable quantity. This group of participants are instantiated (see below for further details on participant initialisation), and then exposed to the familiarisation data sequences. These familiarisation sequences are run through the networks, and a measure of error (see section 2.1.5), called *loss*, is computed. Learning is performed by the standard gradient descent technique (altering network weights in line with this loss). A run through all familiarisation sequences in this fashion continually altering the weights to reduce error is called a training *epoch*.

This process is repeated for multiple epochs and the network rapidly converges on an error plateau. At this point, the network is deemed sufficiently trained. Once the network is trained using our preferred hyper-parameters (see below), a final test is performed. Similar to Cesana-Arlotti et al(Cesana-Arlotti et al., 2018), inference and no-inference scenarios are tested separately. We record the results of these tests and our analysis and interpretation below.

A Note on Participant Initialisation To create participants, a random initialisation process was used. This is could be thought of as being somewhat biologically inspired, mimicking the random variation found in all individuals. Each participant instance begins with a randomly (normal) distributed set of weights which are then trained on the familiarisation sequences described above. For this experiment one hundred (n=100) participants were trained and tested.

Test Set The test data set consists of multiple instances of the scenarios described in section 2.1 above. These include both inference and no inference variations as well as the consistent/inconsistent endings.

Hyper Parameter Optimisation During training we periodically run the testing set through the network (these runs do not affect the weights, i.e. no learning takes place) to measure its accuracy at varying epochs, and also to try to optimise network hyper-parameters like *learning rate* and *activation function*. This process is described in detail in Appendix A. For this experiment it was found that the following hyper-parameters gave the lowest loss scores/training time trade-off: (activation=sigmoid, learning rate=0.4, epochs=100).

Results The differences at various stages of training progression (graphed in figure 6) are measured. To test if the our loss (accuracy) figures are normally distributed we used a Jarque-Bera test on each group. The results of these (Inference: p =

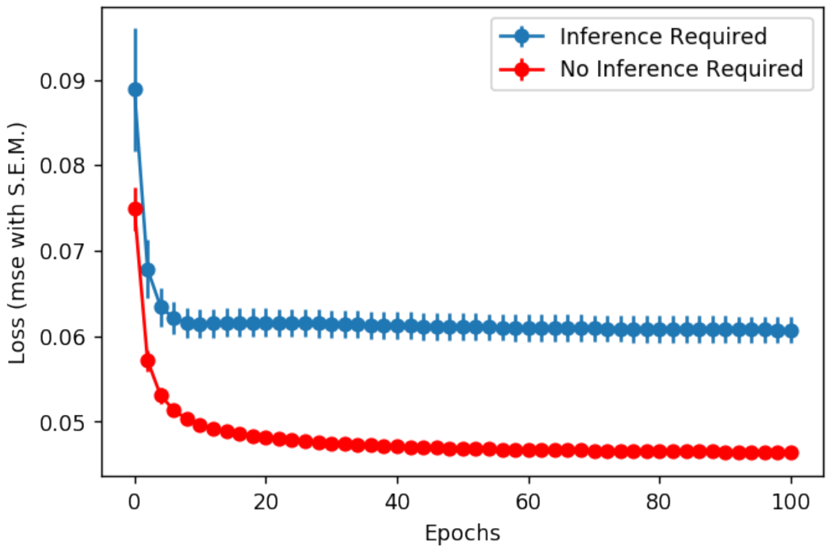


Figure 6: Inference vs No Inference Loss Progression

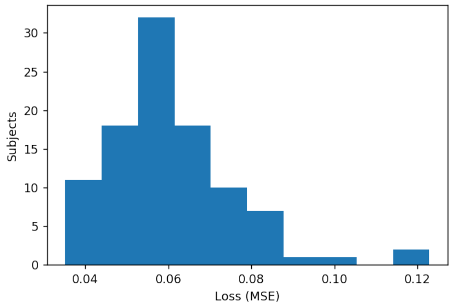


Figure 7: Inference Loss Histogram

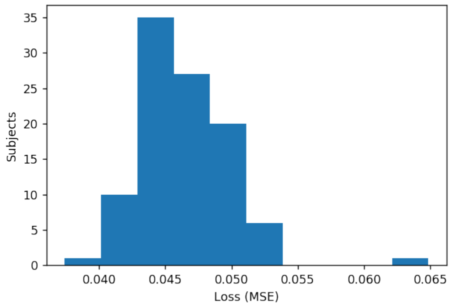


Figure 8: No Inference Loss Histogram

1.1e-16, No-Inference: p = ˜0, (also see figures 7 and 8)) indicate that this is so. An independent t-test was then performed on these groups. A significant (p = 1.46e-14. t = 8.89) difference in the way the networks treated these distinct scenarios was found.

Discussion In the original infant reasoning experiment (Cesana-Arlotti et al., 2018) the authors found, first, that infants were surprised at an inconsistent outcome both if when they were presented with scenarios that required an inference to be drawn in order to determine that the outcome was indeed inconsistent, and when they were presented with scenarios where no inference was required. Second, when comparing the degree of surprise between groups, it was found that 12 month old infants in the no-inference condition were as surprised as those in the inference condition, not showing any significant looking time differences at the outcome stage when presented with these distinct scenarios, see Figure 3. A small difference appeared in 19 month-old infants, who gave signs of being more surprised at the outcome when it directly violated a physical law than when they had to logically derive the inconsistency of the outcome.

This suggests that at least younger infants construct a model of the world that they update using various methods, inference-based updates being only one kind of update. They do not appear to differentiate between states of the world arrived at via inference vs states directly perceived, without the need for an inference. That is, their model of the world is independent of the method of arriving at that model. They forget the 'how' or 'why', and react based solely on the 'what' of their model.

The results above indicate that the network behaves significantly differently than an infant at her earliest stages of knowledge acquisition, when presented with our representations. It treats the inference variations quite distinctly. This is an interesting result. We believe that it is probably an artifact of the network architecture and training rather than being specifically related to the inference/no-inference properties of our test scenarios. Given some knowledge of the principals of neural networks and the gradient descent procedure, it can be said that they learn to approximate a function more and more accurately with training. In this case, manually examining the training data we discovered that the network might be trying to approximate a constant 'one' (consistent) output response no matter the input data. This makes sense since our familiarisation scenarios are not designed to include inconsistent world states or violations of normal world properties. Further, this is a realistic assumption, because an infant is almost always in that same position: she never experiences any breaks in the laws of physics, outside the realm of cartoons or fairy tales which are by nature exceptional (and thereby entertaining and surprising). It is however a substantial flaw in this representation that because of how impoverished our output representation is (i.e. a single bit), it is not possible to give the network a better hint of the true function it should be learning.

Thus, the results of this first representation are mixed. On the one hand, the rapid descent of the loss function over a small number of epochs suggests a reasonably quick convergence towards a solution. On the other hand, the solution converged upon seems to exhibit some potentially considerable differences with respect to infants' surprise behavior. In short, the network succeeds for reasons that have nothing to do with the inner mental processes of an infant. The fact that it is only at the end of each scenario that we can usefully examine the output of the network poses another difficulty for the interpretation. This is the only point in the experiment where the scene can become inconsistent. Ideally measuring the accuracy of the network throughout the scenario would allow us to reduce the gap between the network's behavior and the infants' behavior, but this analysis is not possible here.

What we can affirm with certainty is that the capabilities of the network are rudimentary in comparison with what it is known that infants can do. For example, the output does not include identity (object or numeric) or spatial location tracking representations. Overall, these considerations lead us to conclude that this is not a particularly good or likely model of infant cognition. In order to reduce the distance between infants' plausible representations and the functioning of a network, we decided to construct a different network characterized by a finer representational structure. Experiment 2 implements a different schema, described below. In that experiment we seek to ameliorate the single and constant output bit during training, and monitor accuracy throughout scenario evolution. We hope that the expanded output representation will improve the behaviour of the network, towards a more cognitively plausible one. The extra information provided in the training scenarios should also help the network to converge on a solution which does not show such categorical differences especially with regard to the inference/no-inference gap.

2.4 Experiment 2 - A World Modelling Schema

> *"No computer has ever been designed that is ever aware of what it's doing; but most of the time, we aren't either."*
>
> – Marvin Minsky

Material To address the shortcomings of the representation used in experiment 1, a different and somewhat more articulated representation was devised. This representation aims to have the network learn to keep an ongoing model of its world up to date as the scene changes. More specifically the aim is to train the network to keep track of the location of object A and B throughout the scenario. It must successfully predict the location and make the required inferences when appropriate.

A change is made to how the location is represented, reducing the number of independent locations to three, as it was felt that four was redundant in the first experiment. The locations are broken down as presented in Figure 9 and the data representation of those locations (L1, L2 and L3 for each object) are exemplified in Figure 10.

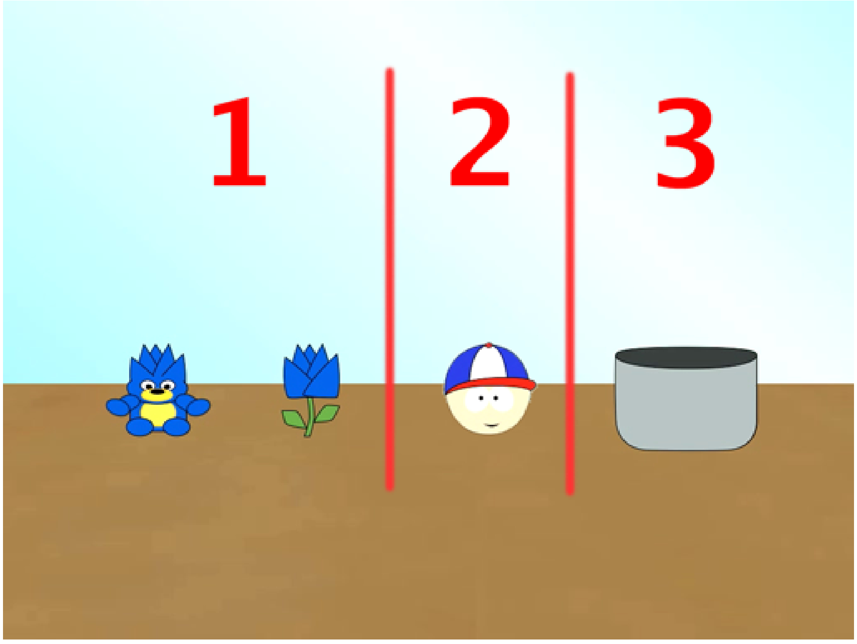
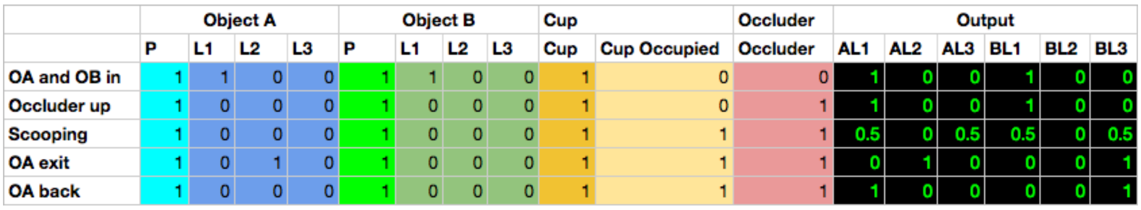


Figure 9: The three possible locations for objects



	Object A				Object B				Cup		Occluder	Output					
	P	L1	L2	L3	P	L1	L2	L3	Cup	Cup Occupied	Occluder	AL1	AL2	AL3	BL1	BL2	BL3
OA and OB in	1	1	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0
Occluder up	1	0	0	0	1	0	0	0	1	0	1	1	0	0	1	0	0
Scooping	1	0	0	0	1	0	0	0	1	1	1	0.5	0	0.5	0.5	0	0.5
OA exit	1	0	1	0	1	0	0	0	1	1	1	0	1	0	0	0	1
OA back	1	0	0	0	1	0	0	0	1	1	1	1	0	0	0	0	1

Figure 10: Sample input and expected output test data

We reduce the representation of the objects here by removing the elements indicating top and bottom half visibility, and replace the visibility components with a simpler 'presence' component (P). This indicates to the network that this object is present in the scene, but not its visibility status. That status is indicated by the location components. If it is possible to tell where an object is located then the location components (L1, L2, L3) will carry that information, otherwise, they will be zero. Cup location is also removed from this representation. It is represented here by a 'presence' component, and a component indicating whether something is visible inside the cup. For a fuller description see Appendix C. These modifications are, we believe, justified because it was felt they did not add any extra (logical) information to the first representation.

The output representation is however much expanded. We decided to explicitly represent the locations of each object throughout the scene. We use the same three locations, for each object A and B, leading to six separate elements of the output (AL1, AL2, AL3, BL1, BL2, and BL3). This appealed to us as the simplest way to directly represent the ongoing world, with the nice property of also exhibiting an ideal input/output location representation symmetry. This output format also allows the performance of the network to be tracked throughout the scenario, something we sought to achieve after our experiences in experiment 1.

Procedure The procedure in this experiment is identical to the first, with minor and necessary modifications to the architecture to change the input and output node counts. The key difference is the representational structure. The participants (n=100) are again created using the normally distributed random initialisation procedure. It was found that the same hyper-parameters (learning rate=0.4, epochs=100, activation=sigmoid) provided a satisfactory accuracy/time trade off during training. Training proceeds as previously described and when this is done, we measure the accuracy in the key inference and no inference scenarios.

Results The overall performance of the network measured using MSE was comparable if a little bit worse than the performance of experiment 1. The difference however is small, and overall the representation achieved a reasonable degree of accuracy.

A gap was found in the performance among the inference and no-inference variations when these groups were compared, but in contrast to the result in our first representation, here the network was found to perform worse in the no inference condition. Also noteworthy, is the progressive decrease in test accuracy (especially in the no-inference variation) as the network becomes more and more trained on the familiarisation set. Both of these aspects are presented in Figure 11.

To test if the final loss (accuracy) figures are normally distributed, a Jarque-Bera test was performed on each group. The results of these (Inference: p = 0.02, No-Inference: p = 0.05, (also see figures 12 and 13)) indicate that these groups are close to normal. An independent t-test was then performed on these groups giving a result of (p = 1.46e-14. t = 8.89). A separate non parametric test (Mann Whitney U) produced similar results (p = 9.15e-14. statistic = 1987.0). This indicates a significant difference in the way the networks treated these distinct variations.

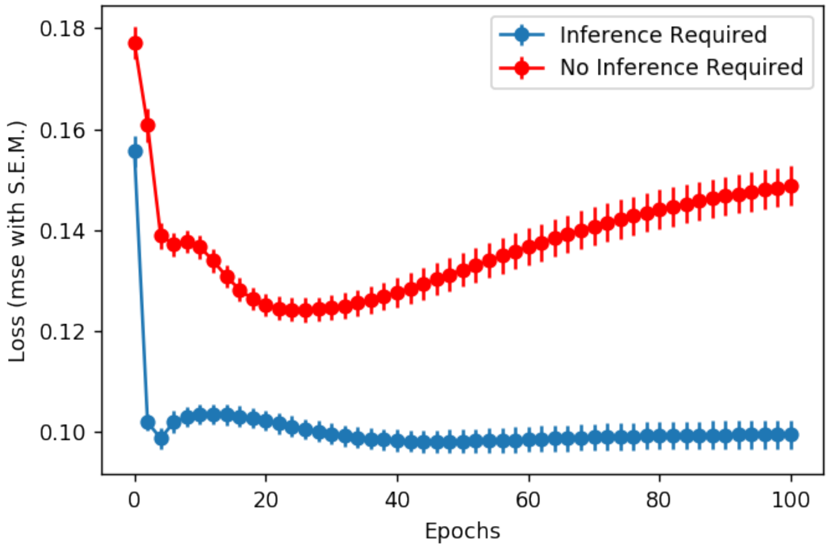


Figure 11: Inference vs No Inference loss (on test set) n=100

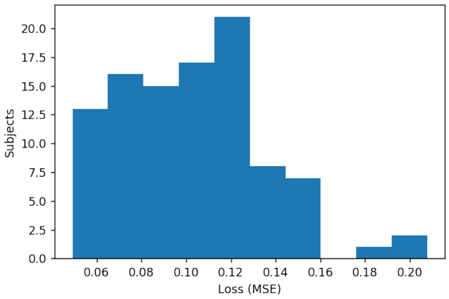


Figure 12: Inference Loss Histogram

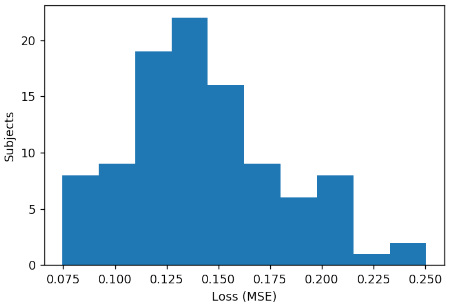


Figure 13: No Inference Loss Histogram

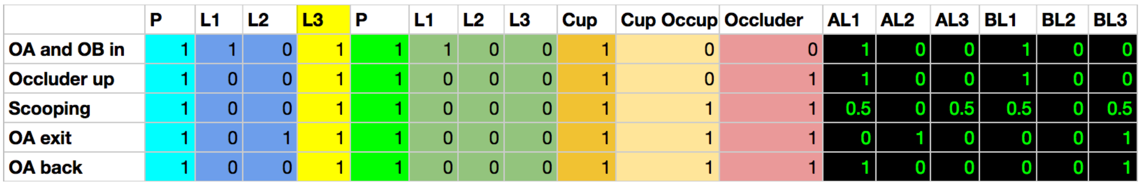
Discussion This model makes assessing progress possible at each step in the scenario, and provides a more transparent gauge on what the network 'believes' about the world. Overall accuracy is comparable with the first experiment. The network provides a more explicit output from a more compact input, but retains some of the major simplifications used in the first experiment (object identity in particular, see note above). The result indicates a difference in how the network views inference and no-inference variations. An indicator that the network is doing something distinct from what a human infant is doing. What the network has learned here, and what it is doing is more complex to determine. The function it approximates is less clear to us from an inspection of the data.

Another interesting aspect is the inversion in the pattern of inference & no inference scenario accuracy. It is difficult to interpret this inversion, and we can offer no particular interpretive insight here. What is important is the difference when compared with infant behaviour. This model is an unlikely candidate for 12 month old cognition as they show no difference between the either type. However for the 19 month old's this presents an extreme divergence, they show the opposite behaviour to our networks, a tendency to be less surprised at the no inference scenario, where our network predicts the opposite. We believe it is remarkable that as soon as the representational structure of the network takes a step closer to a plausible implementation of what infants may represent, the results of the simulations begin diverging from infants' behavior. This result seems to point at a trade-off between representational richness and real world network performance which may reveal an intrinsic limitation of these kinds of networks. Alternatively, it is conceivable that this divergence is related, not to the intrinsic limits of the network, but to one of the necessary abstractions that we imposed over it in constructing its basic representational structure, and in particular, to the *object-identity* simplification. This hypothesis presents another opportunity for further future experimentation.

2.5 Experiment 3 - An Impossible Situation

> *"Most people would sooner die than think; in fact, they do so."*
>
> – Bertrand Russell

To test if the divergence discovered in experiment 2 may be due to a general limit of the network or to how objects are represented, we tested it by presenting it with a highly unusual, incoherent and physically impossible situation, one which had could



	P	L1	L2	L3	P	L1	L2	L3	Cup	Cup Occup	Occluder	AL1	AL2	AL3	BL1	BL2	BL3
OA and OB in	1	1	0	1	1	1	0	0	1	0	0	1	0	0	1	0	0
Occluder up	1	0	0	1	1	0	0	0	1	0	1	1	0	0	1	0	0
Scooping	1	0	0	1	1	0	0	0	1	1	1	0.5	0	0.5	0.5	0	0.5
OA exit	1	0	1	1	1	0	0	0	1	1	1	0	1	0	0	0	1
OA back	1	0	0	1	1	0	0	0	1	1	1	1	0	0	0	0	1

Figure 14: An Impossible Situation Dataset (**NB:** Object A Location 3 highlighted)

never appear in the real world. Our reasoning behind this was that in so doing we could test if the network would react in a 'natural' (surprised) way or merely accept the situation the same as any other. If the network accepted the situation as normal, we could plausibly argue that the divergence between the network's behavior and infants' behavior could be due to the severe limitations we imposed upon the representation of the objects in the scene. If, instead, the network reacted with 'surprise' at the impossible situation, we would be more likely to believe that this was not the cause of the divergence, which might therefore be attributed to some deeper limits of its functioning.

Material To force the network to represent an impossible object, we decided to present it with object A located at both position 1 and 3, from the beginning of the scene. This situation can be created in a straightforward manner by altering the representation, which does not implicitly block such a configuration. This is presented in Figure 14, with Object A, Location 3 highlighted in yellow.

Procedure The network was trained as before on the familiarisation set. Once trained the network was presented with the constructed impossible scenario presented in Figure 14. The accuracy of the network was measured in the standard fashion and compared as training progresses with our standard test cases.

Results The results of this comparison are presented visually in Figure 15. The accuracy of the network looks worse (higher loss) for this strange situation.

A T-Test was run on the two distributions (fully trained at 100 epochs, with possible and impossible Jarque Bera tests indicating normality at ($p = 4.7e\text{-}06$) and ($p = 1.0e\text{-}07$) respectively). The results ($p = 3.8e\text{-}27$. $t = -14.61$) indicate that they are indeed significantly different.

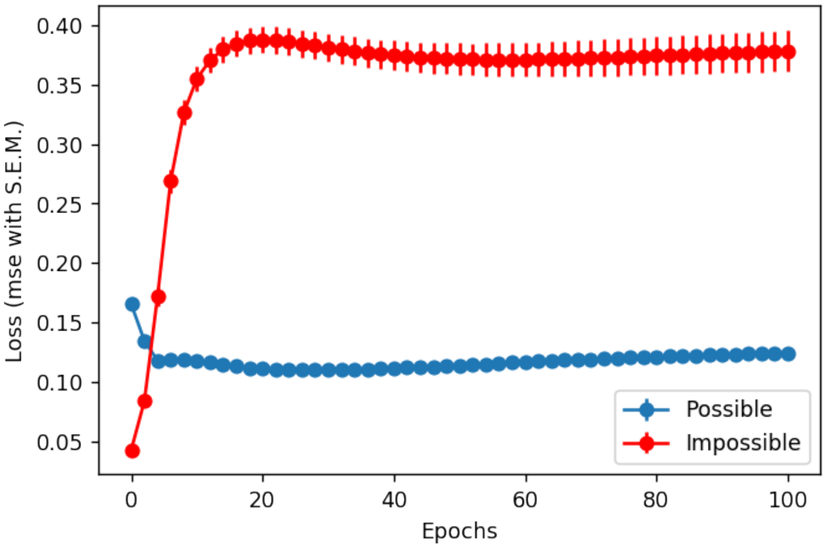


Figure 15: Possible vs impossible loss (n=100)

Discussion The significantly poorer performance in this situation suggests that the network does appear to detect an anomaly in the scenario. We can interpret this as surprise as we do when infant looking time is increased. This supports the conclusion that this representation is a more powerful and realistic candidate for minimal structural complexity because it exhibits some more human (infant) like cognitive behaviour. This conclusion is in a sense a double-edged sword. On the one hand it is possible to conclude that a network as simple as ours *already* embodies some basic object representation abilities. It indicates that, perhaps basic object representation is emerges rather naturally from these systems, and it also explains why other researchers have found reasonably positive results when testing how these simple connectionist neural networks mimic basic results in object representations (Munakata, 1998). On the other hand, it excludes simple explanations as to why our network was unable to pick up some basic differences in infants' behavior when simulating inference vs no-inference situations, thus possibly indicating a deeper limitation in how this architecture can be a plausible candidate to account for early human cognition.

3 General Discussion

> *"It has been a long road from Plato's Meno to the present, but it is perhaps encouraging that most of the progress along that road has been made since*

> *the turn of the twentieth century, and a large fraction of it since the midpoint of the century. Thought was still wholly intangible and ineffable until modern formal logic interpreted it as the manipulation of formal tokens. And it seemed still to inhabit mainly the heaven of Platonic ideals, or the equally obscure spaces of the human mind, until computers taught us how symbols could be processed by machines."*
>
> – Allen Newell

With the aim of probing what level of structural complexity is necessary to exhibit some basic infant like cognitive capacities we constructed multiple experiments. These experiments varied the representational format used to train and test the networks while keeping the underlying architecture constant.

The results were mixed with the networks exhibiting some patterns that did not parallel infant behaviour, and others that could be interpreted as successful and realistic reasoning about the world.

Our first representation had a very limited output. This output could only really be interrogated in the final outcome stage making judgment of performance difficult in intermediate stages of a scenario. It demonstrated reasonable accuracy, though a further inspection of the output suggests that network had learned a relatively constant output due to the nature of the familiarisation sequences (always consistent).

In the second representation a fuller more articulated output model proved to be much more amenable to analysis. This model provided a consistent view of the world for comparison at each step in the scenario. While this network had a comparable accuracy score on our test set, the difference between inference and no-inference conditions showed an inverted pattern compared with that of the first representation, i.e. the network performed worse on no-inference variations than inference variations.

This particular gap demonstrated between inference and no-inference conditions presents an anomaly. This difference does not emerge in 12 month old infants though there is a small difference with 19 month olds (Cesana-Arlotti et al., 2018). The gap is quite distinctive in our results and is evidence that can be drawn against a cognitive similarity between our networks and infant cognitive behaviour. Interestingly, the dissimilarity began to arise when the representation of events started moving closer to what infants may plausibly be said to represent in their structure of the events. It is difficult to know whether we are tapping into some fundamental limitation of the network used here, or whether the discrepancies are due to some of the radical simplifications we introduced in our initial representation. We are inclined to think that the latter possibility can be excluded. First, because our representation is not so different from what other similar studies have used to simulate cognitive phenomena, and second, because of the results of experiment 3 with the impossible scenario.

The networks performance in experiment 3 shows a degree of natural aversion to an impossible situation, that could be interpreted as demonstrating some basic physical knowledge and reasoning about the world. This suggests our second 'World Modelling' representation possesses a degree of structural complexity powerful enough to be considered a candidate for a minimal representation. As a consequence, the limits we encountered in how such networks can reproduce infants' behavior are probably not due to a deficient object representation.

As mentioned briefly in section 2.1.5, we have chosen MSE as our loss measure, a simple euclidean distance metric, which averages across the losses at each step in the scenario. In retrospect, we believe we are running up against the limits of this metric with these more advanced representations, and so this has turned out to be a less than ideal choice of performance metric. Particularly given the fact that one of our principle design considerations for this representation was continuous evaluation. The truly important steps (inference drawing stages) in our scenarios are the places that should be focused on the most performance-wise, so that the true character of the representation with regards to its cognitive similitude can be fairly judged. We believe a more focused and tailored metric should be used in future experiments, to avoid this rather simplified accuracy measure concealing possibly interesting aspects of our networks' capacities.

There are other concerns which are not just limited to our experimentation here but are a more general aspect of connectionist neural networks. As briefly discussed above in experiment 1, neural networks learn to approximate a function from input to output. They use the gradient descent technique to reduce the error between their current approximation of the function, and the function expressed by the input-output training data. The problem here is the well known one of finding a balance between over training and under training. If the network is over trained, it learns the training set too well, and approximates only that function, not being able to generalise to newer data it has not seen before. It misses the subtlety or generality of the function that was desired having swept past it in an over eager attempt to reduce training set error. Similarly if the network is not trained enough the function it approximates is either inaccurate or imprecise, both conditions leading to poor performance in most cases against the actual desired objective function.

One alternative explanation for our results, as discussed, is that our networks have learned something approaching the right function, but remain either over or under trained. The results we see then are possibly simply artefacts which we may in ignorance over interpret as evidence, pro or con, for actual cognition-like behaviour.

We have demonstrated the use of these older connectionist models as tools for probing the quality of structure required to reproduce some very rudimentary elements of infant cognition. While connectionist networks may not currently provide satisfactory computational or theoretical models for cognition, they do show some utility as implements, as a kind of yardstick of structural complexity.

This work then displays some promise for further lines of inquiry. In particular the architecture of the network has been kept static in this work. This presents an obvious and interesting dimension worth exploring, especially considering our use of an older and quite primitive Elman recurrent neural network. The significant recent developments in modern neural network technology (e.g. LTSM networks) suggest many further permutations and configurations deserving of greater examination.

Another possibility is to present the scenes to the network as a two dimensional matrix of pixel values. This method has the advantage of being free of subjective human judgments entering into the representation. The downside here is the amount of data and the computational resources needed to train the network on such a large input. Most of the large AI companies follow this approach in their video game based training (Mnih et al., 2015), although they do downsample quite significantly. Another downside that can be foreseen is how to explain what the network has learned and what it is reasoning about and how, another unsolved problem in the field.

This paper has presented a modest and somewhat novel use of an older tool for cognitive simulation and experimentation. The results are suggestive of further interesting exploratory work. There are many avenues forward and it would seem, much to be learned.

References

Baillargeon, R. (1993, 01). The object concept revisited: New directions in the investigation of infants"physical knowledge. , *23*.

Cesana-Arlotti, N., Martín, A., Téglás, E., Vorobyova, L., Cetnarski, R., & Bonatti, L. L. (2018). Precursors of logical reasoning in preverbal human infants. *Science*, *359*(6381), 1263–1266. Retrieved from http://science.sciencemag.org/content/359/6381/1263 doi: 10.1126/science.aao3539

Chomsky, N. (1967). A review of b.f. skinners verbal behavior. *Readings in the Philosophy of Psychology*, *1*, 142–143.

Elman, J. L. (1990). Finding structure in time. *COGNITIVE SCIENCE*, *14*(2), 179–211.

Fodor, J. (1975). *The language of thought.* Harvard Univ Pr.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1), 3 - 71. Retrieved from http://www.sciencedirect.com/science/article/pii/0010027788900315 doi: https://doi.org/10.1016/0010-0277(88)90031-5

Minsky, M., & Papert, S. (1969). *Perceptrons: An introduction to computational geometry.* Cambridge, MA, USA: MIT Press.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015, 02 25). Human-level control through deep reinforcement learning. *Nature*, *518*, 529 EP -. Retrieved from http://dx.doi.org/10.1038/nature14236

Munakata, Y. (1998). Infant perseveration and implications for object permanence theories: A pdp model of the ab task. *Developmental Science*, *1*(2), 161-184. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-7687.00021 doi: 10.1111/1467-7687.00021

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Rumelhart, D. E., & McClelland, J. L. (1986). Parallel distributed processing: explorations in the microstructure of cognition. volume 1. foundations.

Xu, F. (1999, Sep). Object individuation and object identity in infancy: the role of spatiotemporal information, object property information, and language. *Acta psychologica*, *102*(2-3), 113–36.

Xu, F., & Carey, S. (1996, Apr). Infants' metaphysics: the case of numerical identity. *Cognitive psychology*, *30*(2), 111–53.

A Neural Network Architecture

> *"The first principle is that you must not fool yourself and you are the easiest person to fool."*
>
> – Richard Feynman

The architecture chosen to model the infants reasoning process was a recursive neural network or RNN. This was based on the structure of the input data, which is presented as a series of discrete events over time, building an environmental context or narrative. In particular we chose an Elman Neural Network (Elman, 1990), one of the earliest and simplest RNN's available. There are other more advanced RNN architectures available (e.g. LSTM), and these could be examined as possible successors in our experimentation.

A.1 Network Initialisation

The network was initialised with normally distributed ($\mu = 0, \sigma = 0.3$) random values, as briefly described above in the participant initialisation section. Further investigation was not pursued on varying this standard initialisation procedure, though it is possible one could see some performance improvements on network convergence here, i.e. fewer training epochs maybe required. it is possible that a positive only normal distribution (with $\mu = 0.5$ for example) or a uniform distribution (in the range [0-1]) might lead to values more representative of the training and test sets initially.

A.2 Training Procedure

Training data is presented to the network as a series of events with varying representations. These representations are described fully in Appendices B and C. We used the standard gradient descent method to change the network's weights after each event, and after a full pass through all of our training scenarios, technically termed an *epoch*, the process is repeated. How often the process is repeated is determined by an accuracy/time trade off. Figure 16 and 17 present the decrease in the loss of the network with the number of epochs trained, but this tails off quickly after a few epochs and shows diminishing returns.

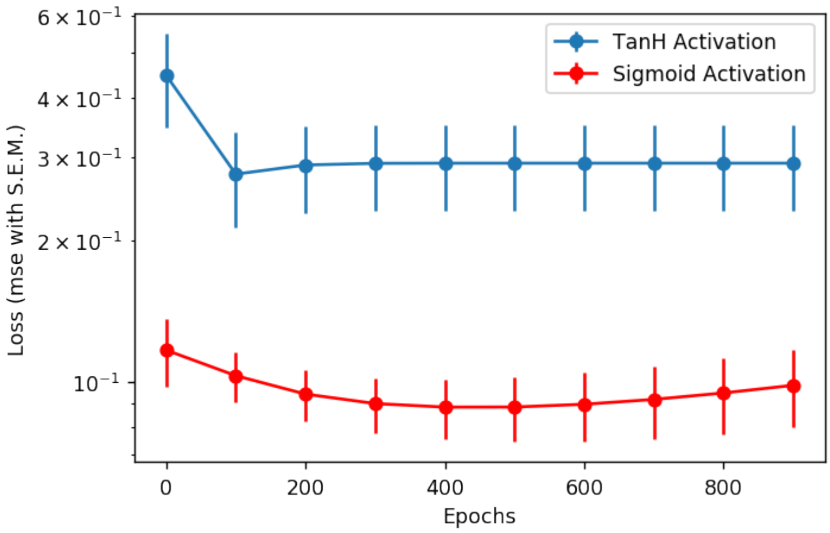


Figure 16: Loss (accuracy on test set) over epochs with different activation functions. (Learning Rate=0.4, Number of Participants (n=10))

A.3 Activation Functions

There are several activation functions that can be used with our network. Testing took place on the two most common to find the best performance. Figure 16 plots the TanH versus the Sigmoid activation functions, showing a noticeable difference in overall accuracy and spread across differing participants. For this reason the sigmoid activation function was chosen as the best option.

A.4 Learning Rate

The speed at which the network "learns", or more precisely the rate at which the weights are adjusted by the gradient descent algorithm can be varied. This learning rate parameter is important to optimise, though determining it a priori is not usually possible. If it is too low, the network will take a very long time to learn anything. If it is too high it is possible it will skip over or miss an important optimal weighting to solve the problem. Figure 17 presents the effects of various learning rates on the accuracy of the network.

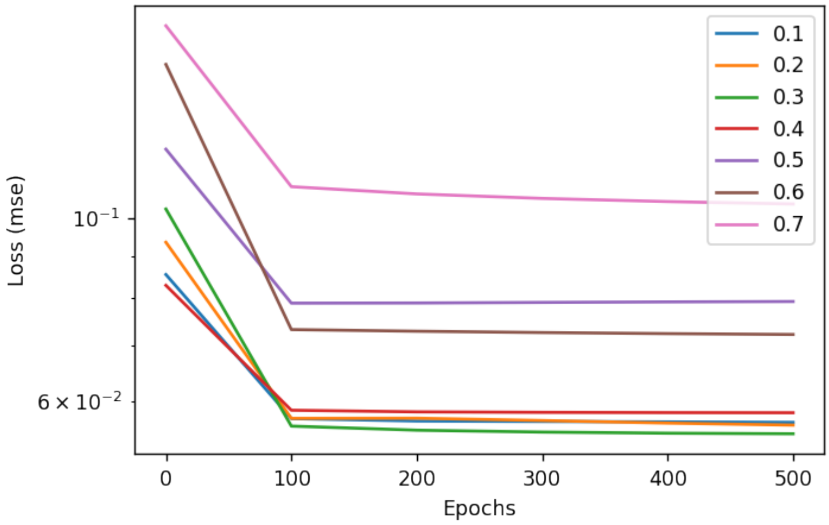
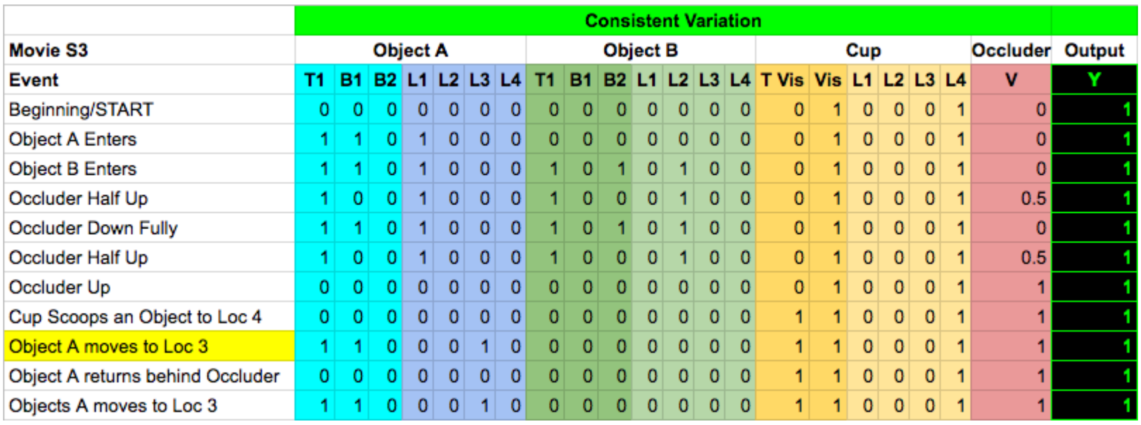


Figure 17: Loss (accuracy on test set) over epochs at various learning rates. (Activation=Sigmoid, Number of Participants (n=10))

B Consistency Detection Representation

> *"In the computer field, the moment of truth is a running program; all else is prophecy."*
>
> – Herbert A. Simon

Data for a sample scenario can be seen below (including input and expected output) below in Figure 18. It can be seen that there are twenty one different input symbols and a single output symbol. Included in the figure is a brief description of the event happening at that particular time step.



	Consistent Variation																					
Movie S3	**Object A**							**Object B**							**Cup**						**Occluder**	**Output**
Event	**T1**	**B1**	**B2**	**L1**	**L2**	**L3**	**L4**	**T1**	**B1**	**B2**	**L1**	**L2**	**L3**	**L4**	**T Vis**	**Vis**	**L1**	**L2**	**L3**	**L4**	**V**	**Y**
Beginning/START	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1
Object A Enters	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1
Object B Enters	1	1	0	1	0	0	0	1	0	1	0	1	0	0	0	1	0	0	0	1	0	1
Occluder Half Up	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0.5	1
Occluder Down Fully	1	1	0	1	0	0	0	1	0	1	0	1	0	0	0	1	0	0	0	1	0	1
Occluder Half Up	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0.5	1
Occluder Up	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1
Cup Scoops an Object to Loc 4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	1
Object A moves to Loc 3	1	1	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	1
Object A returns behind Occluder	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	1
Objects A moves to Loc 3	1	1	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	1

Figure 18: Input and expected output test data for one particular scenario

This representation can be broken down into different groups, which are shown in the headers of figure 18, and are documented below.

Locations There are four possible locations for an element in the scene. These are show in Figure 19.

Object A and B representations There are seven component elements used to represent the state of an object in a scene, three for visibility and four to represent the location. They are described in detail in Table 1 below.

Cup Representation The cup has a similar representational schema to the objects, it has four location elements, L1 to L4, and two visibility elements. The visibility element Vis describes the overall visibility of the cup itself within the scene. The TVis element indicates whether the top half of one of the objects within

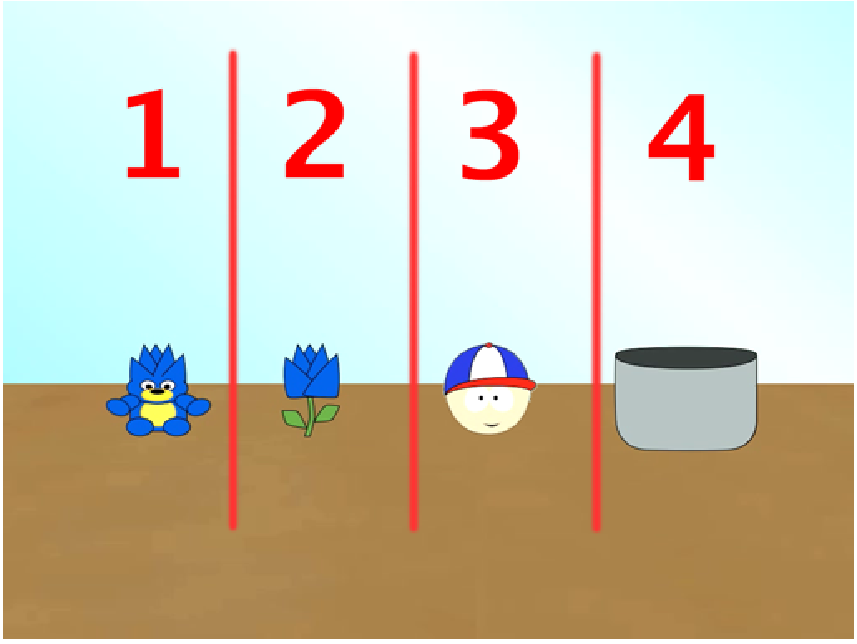


Figure 19: The four possible locations for objects in this representation

the scene is visible in the cup, i.e. whether there is an object in the cup. These elements are further specified in Table 3.

Occluding Wall Representation The occluding wall never changes position and therefore has a simpler representation than the other objects. However since the wall can be in three different states, down, half way up or fully up, it has three different values it can take on. These are described fully in Table 5.

Output Representation As discussed, we use a very minimal output representation, which was designed to simulate the infants level of surprise at the consistency of what she has seen. This simplifies our accuracy measure, so that we can judge the effectiveness of the network in a straightforward manner. We use 1 to indicate consistency or lack of surprise with the state of the world, and 0 to indicate surprise, that something does not conform to expectations. This is described briefly in Table 7.

Name	**Description**	**Values**
T	Top half visible	0,1
B1	Bottom half (Type 1) visible	0,1
B2	Bottom half (Type 2) visible	0,1
L1	Object is at location 1	0,1
L2	Object is at location 2	0,1
L3	Object is at location 3	0,1
L4	Object is at location 4	0,1

Table 1: Object Representation

Legend

0	=	**VISIBLE**
1	=	**NOT VISIBLE**

Name	**Description**	**Values**
TVis	Top half of object is visible in cup	0,1
Vis	Cup is visible	0,1
L1	Object is at location 1	0,1
L2	Object is at location 2	0,1
L3	Object is at location 3	0,1
L4	Object is at location 4	0,1

Table 3: Cup Representation

Legend

0	=	**VISIBLE**
1	=	**NOT VISIBLE**

Name	**Description**	**Values**
Vis	Visibility of Wall (Down, Half Up, Up)	0, 0.5, 1

Table 5: Occluding Wall Representation

Legend

0	=	**WALL DOWN**
0.5	=	**WALL HALF UP**
1	=	**WALL FULLY UP**

Name	**Description**	**Values**
O	Scenario Consistency	0, 1

Table 7: Output Representation

Legend

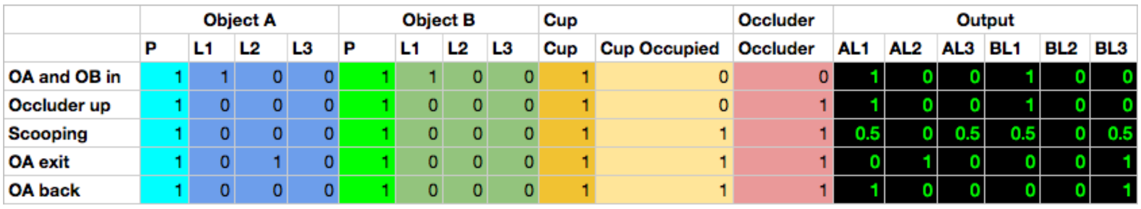
0	=	**SCENARIO INCONSISTENT**
1	=	**SCENARIO CONSISTENT**

C World Model Representation

"Every good mathematician is at least half a philosopher, and every good philosopher is at least half a mathematician."

– Gottlob Frege

Data for a sample scenario can be seen (including input and expected output) in Figure 20. It can be seen that there are seventeen different input symbols and six output symbols.



	Object A				Object B				Cup		Occluder	Output					
	P	L1	L2	L3	P	L1	L2	L3	Cup	Cup Occupied	Occluder	AL1	AL2	AL3	BL1	BL2	BL3
OA and OB in	1	1	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0
Occluder up	1	0	0	0	1	0	0	0	1	0	1	1	0	0	1	0	0
Scooping	1	0	0	0	1	0	0	0	1	1	1	0.5	0	0.5	0.5	0	0.5
OA exit	1	0	1	0	1	0	0	0	1	1	1	0	1	0	0	0	1
OA back	1	0	0	0	1	0	0	0	1	1	1	1	0	0	0	0	1

Figure 20: Sample input and expected output test data

This differs from the first representation not only in the size of the output but also in the semantics of the location components. These are described below.

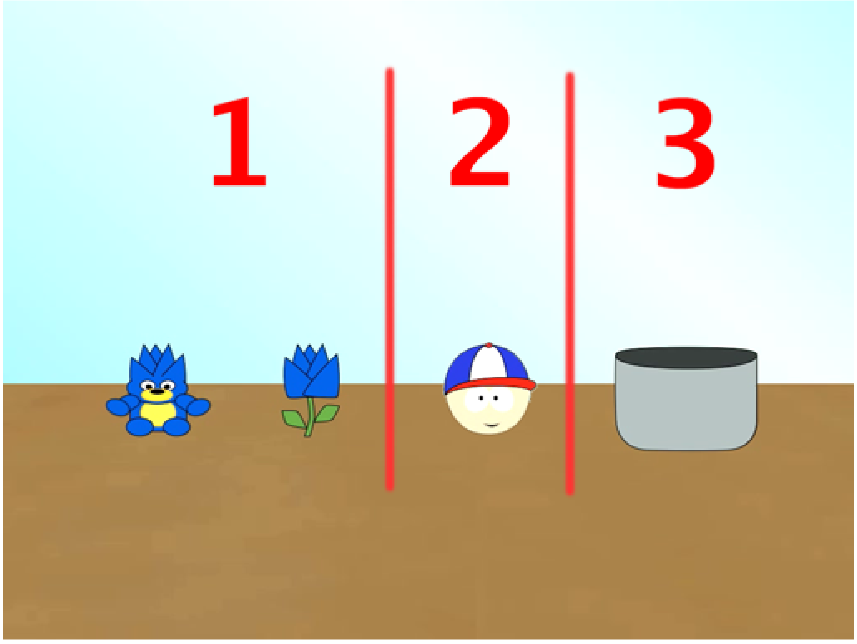


Figure 21: The three possible locations for objects in this representation

Locations There are three possible locations for an element in the scene. These are show in Figure 21. This differs from the first representation in that we merge locations 1 and 2 which are both behind the occluding wall.

Object A and B representations There are four components elements used to represent the state of an object in a scene, one for presence in the scene, and three to represent the location. They are described in detail in Table 9.

Name	**Description**	**Values**
P	Object is present in scene	0,1
L1	Object is at location 1	0,1
L2	Object is at location 2	0,1
L3	Object is at location 3	0,1

Table 9: Object Representation

Legend

0	=	**NOT PRESENT**
1	=	**PRESENT**

Cup Representation The cup representation is much reduced from the first representation, it doesn't contain any location component, purely the presence of the cup on the scene and whether it contains an object.

Name	**Description**	**Values**	**Legend**
Cup	Cup is present in scene	0,1	0=**NOT PRESENT**, 1=**PRESENT**
Cup Full	Cup contains an object	0,1	0=**CUP EMPTY**, 1=**CUP FULL**

Table 11: Cup Representation

Occluding Wall Representation The occluding wall representation is similar to the *Consistency Detector* representation, a very elemental single bit, indicating the presence or absense of the wall.

Output Representation The output representation is more articulated in this representation in comparison with our alternative. Here we try to capture the locations of the object after each step or scene. This is what gives rise to the name World Model representation, we aim to construct and maintain an internal model of the scene or world.

Name	**Description**	**Values**
Occluder	Occluding Wall is visible (Up) in scene	0,1

Table 12: Occluder Representation

Legend

0	=	**OCCLUDER DOWN**
1	=	**OCCLUDER UP**

Name	**Description**	**Values**
AL1	Object A is at location 1	0,0.5,1
AL2	Object A is at location 2	0,0.5,1
AL3	Object A is at location 3	0,0.5,1
BL1	Object B is at location 1	0,0.5,1
BL2	Object B is at location 2	0,0.5,1
BL3	Object B is at location 3	0,0.5,1

Table 14: Output Representation

Legend

0	=	**NOT** at this location
0.5	=	**UNKNOWN** if at this location
1	=	**AT** this location

D Resources

"Whereof one cannot speak, thereof one must be silent."

– Ludwig Wittgenstein

All code for the above experiments and analysis can be found online on at:

https://github.com/mmcguill/MBC-Precursors-NN

These can be run as Jupyter notebooks on any compatible system.