

# Measuring abstract reasoning in neural networks

David G.T. Barrett<sup>\*1</sup> Felix Hill<sup>\*1</sup> Adam Santoro<sup>\*1</sup> Ari S. Morcos<sup>1</sup> Timothy Lillicrap<sup>1</sup>

## Abstract

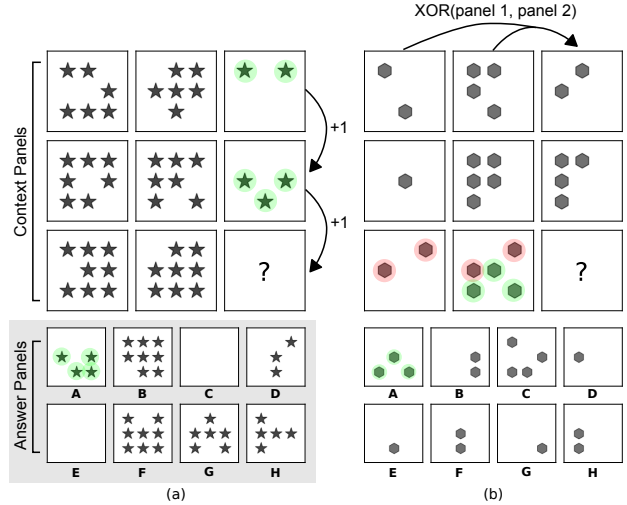
Whether neural networks can learn abstract reasoning or whether they merely rely on superficial statistics is a topic of recent debate. Here, we propose a dataset and challenge designed to probe abstract reasoning, inspired by a well-known human IQ test. To succeed at this challenge, models must cope with various generalisation ‘regimes’ in which the training and test data differ in clearly-defined ways. We show that popular models such as ResNets perform poorly, even when the training and test sets differ only minimally, and we present a novel architecture, with a structure designed to encourage reasoning, that does significantly better. When we vary the way in which the test questions and training data differ, we find that our model is notably proficient at certain forms of generalisation, but notably weak at others. We further show that the model’s ability to generalise improves markedly if it is trained to predict symbolic explanations for its answers. Altogether, we introduce and explore ways to both measure and induce stronger abstract reasoning in neural networks. Our freely-available dataset should motivate further progress in this direction.

## 1. Introduction

Abstract reasoning is a hallmark of human intelligence. A famous example is Einstein’s elevator thought experiment, in which Einstein reasoned that an equivalence relation exists between an observer falling in uniform acceleration and an observer in a uniform gravitational field. It was the ability to relate these two abstract concepts that allowed him to derive the surprising predictions of general relativity, such as the curvature of space-time.

A human’s capacity for abstract reasoning can be estimated

<sup>\*</sup>Equal contribution, ordered by surname. <sup>1</sup>DeepMind, London, United Kingdom. Correspondence to: <{barrettdavid; felixhill; adamsantoro}@google.com>.



**Figure 1. Raven-style Progressive Matrices.** In (a) the underlying abstract rule is an arithmetic progression on the number of shapes along the columns. In (b) there is an XOR relation on the shape positions along the rows (panel 3 = XOR(panel 1, panel 2)). Other features such as shape type do not factor in. A is the correct choice for both.

surprisingly effectively using simple visual IQ tests, such as Raven’s Progressive Matrices (RPMs) (Figure 1) (Raven et al., 1938). The premise behind RPMs is simple: one must reason about the relationships between perceptually obvious visual features – such as shape positions or line colors – to choose an image that completes the matrix. For example, perhaps the size of squares increases along the rows, and the correct image is that which adheres to this size relation. RPMs are strongly diagnostic of abstract verbal, spatial and mathematical reasoning ability, discriminating even among populations of highly educated subjects (Snow et al., 1984).

Since one of the goals of AI is to develop machines with similar abstract reasoning capabilities to humans, to aid scientific discovery for instance, it makes sense to ask whether visual IQ tests can help to understand learning machines. Unfortunately, even in the case of humans such tests can be invalidated if subjects prepare too much, since test-specific heuristics can be learned that shortcut the need for generally-applicable reasoning (Te Nijenhuis et al., 2001; Flynn, 1987). This potential pitfall is even more acute in the case of neural networks, given their striking capacity for memorization

(Zhang et al., 2016) and ability to exploit superficial statistical cues (Jo & Bengio, 2017; Szegedy et al., 2013).

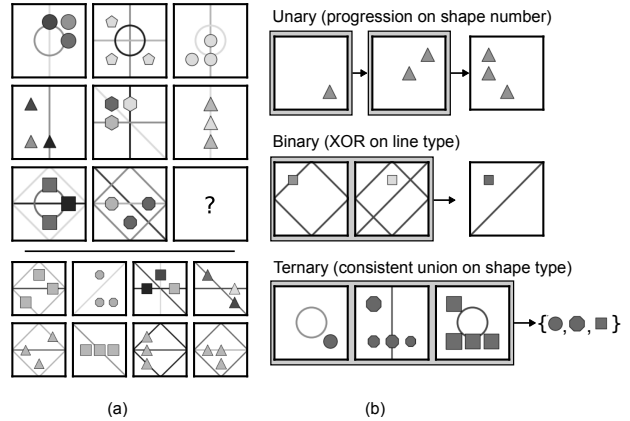
Nonetheless, we contend that visual intelligence tests can help to better understand learning and reasoning in machines (Fleuret et al., 2011), provided they are coupled with a principled treatment of generalisation. Suppose we are concerned with whether a model can robustly infer the notion of ‘monotonically increasing’. In its most abstract form, this principle can apply to the quantity of shapes or lines, or even the intensity of their colour. We can construct training data that instantiates this notion for increasing quantities or sizes and we can construct test data that only involves increasing colour intensities. Generalisation to the test set would then be evidence of an abstract and flexible application of what it means to monotonically increase. In this way, a dataset with explicitly defined abstract semantics (e.g., relations, attributes, pixels, etc.), allows us to curate training and testing sets that precisely probe the generalisation dimensions of abstract reasoning in which we are interested.

To this end, we have developed a large dataset of abstract visual reasoning questions where the underlying abstract semantics can be precisely controlled. This approach allows us to address the following questions: (1) Can state-of-the-art neural networks find solutions – *any* solutions – to complex, human-challenging abstract reasoning tasks if trained with plentiful training data? (2) If so, how well does this capacity generalise when the abstract content of training data is specifically controlled for?

To begin, we describe and motivate our dataset, outline a procedure for automatic generation of data, and detail the generalisation regimes we chose to explore. Next, we establish a number of strong baselines, and show that well known architectures that use only convolutions, such as ResNet-50 (He et al., 2016), struggle. We designed a novel variant of the Relation Network (Santoro et al., 2017; Raposo et al., 2017), a neural network with specific structure designed to encourage relation-level comparisons and reasoning. We found that this model substantially outperforms other well-known architectures. We then study this top-performing model on our proposed generalisation tests and find that it generalises well in certain test regimes (e.g. applying known abstract relationships in novel combinations), but fails notably in others (such as applying known abstract relationships to unfamiliar entities). Finally, we propose a means to improve generalisation: the use of auxiliary training to encourage our model to provide an explanation for its solutions.

## 2. Procedurally generating matrices

In 1936 the psychologist John Raven introduced the now famous human IQ test: Raven’s Progressive Matrices (RPM)



**Figure 2. A difficult PGM and a depiction of relation types.** (a) a challenging puzzle with multiple relations and distractor information. (b) a possible categorization of relation types based on how the panels are considered when computing the relation: for unary, a function is computed on one panel to produce the subsequent panel; for binary, two independently sampled panels are considered in conjunction to produce a third panel; and for ternary, all three panels adhere to some rule, such as all containing shapes from some common set, regardless of order.

(Raven et al., 1938). RPMs consist of an incomplete  $3 \times 3$  matrix of context images (see figure 1), and some (typically 8) candidate answer images. The subject must decide which of the candidate images is the most appropriate choice to complete the matrix.

It is thought that much of the power of RPMs as diagnostic of human intelligence derives from the way they probe *eductive* or *fluid* reasoning (Jaeggi et al., 2008). Since no definition of an ‘appropriate’ choice is provided, it is in principle possible to come up with a reason supporting any of the candidate answers. To succeed, however, the subject must assess all candidate answers, all plausible justifications for those answers, and identify the answer with the strongest justification. In practice, the right answer tends to be the one that can be explained with the simplest justification using the basic relations underlying the matrices.

Although Raven hand-designed each of the matrices in his tests, later research typically employed some structured generative model to create large numbers of questions. In this setting, a potential answer is correct if it is consistent with the underlying generative model, and success rests on the ability to invert the model.

### 2.1. Automatic generation of PGMs

Here we describe our process for creating RPM-like matrices. We call our dataset the *Procedurally Generated Matrices* (PGM) dataset. To generate PGMs, we take inspiration from Carpenter et al. (1990), who identified and catalogued

the relations that commonly underlie RPMs, as well as Wang & Su (2015), who outlined one process for creating an automatic generator.

The first step is to build an abstract structure for the matrices. This is done by randomly sampling from the following primitive sets:

- relation types ( $\mathcal{R}$ , with elements  $r$ ): progression, XOR, OR, AND, consistent union<sup>1</sup>
- object types ( $\mathcal{O}$ , with elements  $o$ ): shape, line
- attribute types ( $\mathcal{A}$ , with elements  $a$ ): size, type, colour, position, number

The structure  $\mathcal{S}$  of a PGM is a set of triples,  $\mathcal{S} = \{[r, o, a] : r \in \mathcal{R}, o \in \mathcal{O}, a \in \mathcal{A}\}$ . These triples determine the challenge posed by a particular matrix. For instance, if  $\mathcal{S}$  contains the triple [progression, shape, colour], the PGM will exhibit a progression relation, instantiated on the colour (greyscale intensity) of shapes. Challenging PGMs exhibit relations governed by multiple such triples: we permit up to four relations per matrix ( $1 \leq |\mathcal{S}| \leq 4$ ).

Each attribute type  $a \in \mathcal{A}$  (e.g. colour) can take one of a finite number of discrete values  $v \in \mathcal{V}$  (e.g. 10 integers between  $[0, 255]$  denoting greyscale intensity). So a given structure has multiple realisations depending on the randomly chosen values for the attribute types, but all of these realisations share the same underlying abstract challenge. The choice of  $r$  constrains the values of  $v$  that can be realized. For instance, if  $r$  is progression, the values of  $v$  must strictly increase along rows or columns in the matrix, but can vary randomly within this constraint. See the appendix for the full list of relations, attribute types, values, their hierarchical organisation, and other statistics of the dataset.

We use  $\mathcal{S}_a$  to denote the set of attributes among the triples in  $\mathcal{S}$ . After setting values for the colour attribute, we then choose values for all other attributes  $a \notin \mathcal{S}_a$  in one of two ways. In the *distracting* setting, we allow these values to vary at random provided that they do not induce any further meaningful relations. Otherwise, the  $a \notin \mathcal{S}_a$  take a single value that remains consistent across the matrix (for example, perhaps all the shapes are the exact same size). Randomly varying values across the matrix is a type of distraction common to Raven’s more difficult Progressive Matrices.

Thus, the generation process consists of: (1) Sampling 1-4 triples, (2) Sampling values  $v \in \mathcal{V}$  for each  $a \in \mathcal{S}_a$ , adhering to the associated relation  $r$ , (3) Sampling values  $v \in \mathcal{V}$  for each  $a \notin \mathcal{S}_a$ , ensuring no spurious relation is induced, (4) Rendering the symbolic form into pixels.

<sup>1</sup>Consistent union is a relation wherein the three panels contain elements from some common set, e.g., shape types {square, circle, triangle}. The ordering of the panels containing the elements does not matter.

## 2.2. Generalisation Regimes

Generalisation in neural networks has been subject of lots of recent debate, with some emphasising the successes (LeCun et al., 2015) and others the failures (Garnelo et al., 2016; Lake & Baroni, 2017; Marcus, 2018). Our choice of regimes is informed by this, but is in no way exhaustive.

**(1) Neutral** In both training and test sets, the structures  $\mathcal{S}$  can contain any triples  $[r, o, a]$  for  $r \in \mathcal{R}$ ,  $o \in \mathcal{O}$  and  $a \in \mathcal{A}$ . The training and test sets are disjoint, but this separation was at the level of the input variables (i.e., the pixel manifestations of the matrices).

**(2) Interpolation; (3) Extrapolation** As in the neutral split,  $\mathcal{S}$  consisted of any triples  $[r, o, a]$ . For interpolation, in the training set, when  $a = \text{colour}$  or  $a = \text{size}$  (the ordered attributes), the values of  $a$  were restricted to even-indexed members of the discrete set  $V_a$ , whereas in the test set only odd-indexed values were permitted. For extrapolation, the values of  $a$  were restricted to the lower half of their discrete set of values  $V_a$  during training, whereas in the test set they took values in the upper half. Note that all  $\mathcal{S}$  contained some triple  $[r, o, a]$  with  $a = \text{colour}$  or  $a = \text{size}$ . Thus, generalisation is required for every question in the test set.

**(4) Held-out Attribute shape-colour or (5) line-type**  $\mathcal{S}$  in the training set contained no triples with  $o = \text{shape}$  and  $a = \text{colour}$ . All structures governing puzzles in the test set contained at least one triple with  $o = \text{shape}$  and  $a = \text{colour}$ . For comparison, we included a similar split in which triples were held-out if  $o = \text{line}$  and  $a = \text{type}$ .

**6: Held-out Triples** In our dataset, there are 29 possible unique triples  $[r, o, a]$ . We allocated seven of these for the test set, at random, but such that each of the  $a \in \mathcal{A}$  was represented exactly once in this set. These held-out triples never occurred in questions in the training set, and every  $\mathcal{S}$  in the test set contained at least one of them.

**7: Held-out Pairs of Triples** All  $\mathcal{S}$  contained at least two triples, of which 400 are viable<sup>2</sup> ( $[r_1, o_1, a_1], [r_2, o_2, a_2]$ ) =  $(t_1, t_2)$ . We randomly allocated 360 to the training set and 40 to the test set. Members  $(t_1, t_2)$  of the 40 held-out pairs did not occur together in structures  $\mathcal{S}$  in the training set, and all structures  $\mathcal{S}$  had at least one such pair  $(t_1, t_2)$  as a subset.

<sup>2</sup>Certain triples, such as [progression, shape, number] and [progression, shape, XOR] cannot occur together in the same PGM

**8: Held-out Attribute Pairs**  $\mathcal{S}$  contained at least two triples. There are 20 (unordered) viable pairs of attributes  $(a_1, a_2)$  such that for some  $r_i, o_i$ ,  $([r_1, o_1, a_1], [r_2, o_2, a_2])$  is a viable triple pair.  $([r_1, o_1, a_1], [r_2, o_2, a_2]) = (t_1, t_2)$ . We allocated 16 of these pairs for training and four for testing. For a pair  $(a_1, a_2)$  in the test set,  $\mathcal{S}$  in the training set contained triples with  $a_1$  and  $a_2$ . In the test set, all  $\mathcal{S}$  contained triples with  $a_1$  and  $a_2$ .

### 3. Models and Experimental Setup

We first compared the performance of several standard deep neural networks on the neutral split of the PGM dataset. We also developed a novel architecture based on Relation Networks (Santoro et al., 2017), that we call the Wild Relation Network (WReN), named in recognition of Mary Wild who contributed to the development of Raven’s progressive matrices along with her husband John Raven.

The input consisted of the eight context panels and eight multiple-choice panels. Each panel is an  $80 \times 80$  pixel image; so, the panels were presented as a set of 16 feature maps.

Models were trained to produce the label of the correct missing panel as an output answer by optimising a softmax cross entropy loss. We trained all networks by stochastic gradient descent using the ADAM optimiser (Kingma & Ba, 2014). For each model, hyper-parameters were chosen using a grid sweep to select the model with smallest loss estimated on a held-out validation set. We used the validation loss for early-stopping and we report performance values on a held-out test set. For hyper-parameter settings and further details on all models see appendix A.

**CNN-MLP:** We implemented a standard four layer convolutional neural network with batch normalization and ReLU non-linearities (LeCun et al., 2015). The set of PGM input panels was treated as a set of separate greyscale input feature maps for the CNN. The convolved output was passed through a two-layer, fully connected MLP using a ReLU non-linearity between linear layers and dropout of 0.5 on the penultimate layer. Note that this is the type of model applied to Raven-style sequential reasoning questions by Hoshen & Werman (2017).

**ResNet:** We used a standard implementation of the ResNet-50 architecture as described in He et al. (2016). As before, each of the context panels and multiple-choice panels was treated as an input feature map. We also trained a selection of ResNet variants, including ResNet-101, ResNet-152, and several custom-built smaller ResNets. The best performing model was ResNet-50.

**LSTM:** We implemented a standard LSTM module (Hochreiter & Schmidhuber, 1997), based on Zaremba et al. (2014). Since LSTMs are designed to process inputs sequentially, we first passed each panel (context panels and multiple choice panels) sequentially and independently through a small 4-layer CNN, tagged the CNN’s output with a one-hot label indicating the panel’s position (the top left PGM panel is tagged with label 1, the top-middle PGM panel is tagged with label 2 etc.), and passed the resulting sequence of labelled embeddings to the LSTM. The final hidden state of the LSTM was passed through a linear layer to produce logits for the softmax cross entropy loss. The network was trained using batch normalization after each convolutional layer and drop-out was applied to the LSTM hidden state.

**Wild Relation Network (WReN):** Our novel WReN model (fig. 3) applied a Relation Network module (Santoro et al., 2017) multiple times to infer the inter-panel relationships.

The model output a 1-d score  $s_k$  for a given candidate multiple-choice panel, with label  $k \in [1, 8]$ . The choice with the highest score was selected as the answer  $a$  using a softmax function  $\sigma$  across all scores:  $a = \sigma([s_1, \dots, s_8])$ . The score of a given multiple-choice panel was evaluated using a Relation Network (RN):

$$s_k = \text{RN}(\mathcal{X}_k) = f_\phi \left( \sum_{y, z \in \mathcal{X}_k} g_\theta(y, z) \right), \quad (1)$$

where  $\mathcal{X}_k = \{x_1, x_2, \dots, x_8\} \cup \{c_k\}$ ,  $c_k$  is the vector representation of the multiple choice panel  $k$ , and  $x_i$  the representation of context panel  $i$ . The input vector representations were produced by processing each panel independently through a small CNN and tagging it with a panel label, similar to the LSTM processing described above, followed by a linear projection. The functions  $f_\phi$  and  $g_\theta$  are MLPs.

The structure of the WReN model is well matched to the problem of abstract reasoning, because it forms representations of pair-wise relations (using  $g_\theta$ ), in this case, between each context panel and a given multiple choice candidate, and between context panels themselves. The function  $f_\phi$  integrates information about context-context relations and context-multiple-choice relations to provide a score. Also the WReN model calculates a score for each multiple-choice candidate independently, allowing the network to exploit weight-sharing across multiple-choice candidates.

**Wild-ResNet:** We also implemented a novel variant of the ResNet architecture in which one multiple-choice candidate panel, along with the eight context panels were provided as input, instead of providing all eight multiple-choices and eight context panels as input as in the standard ResNet. In

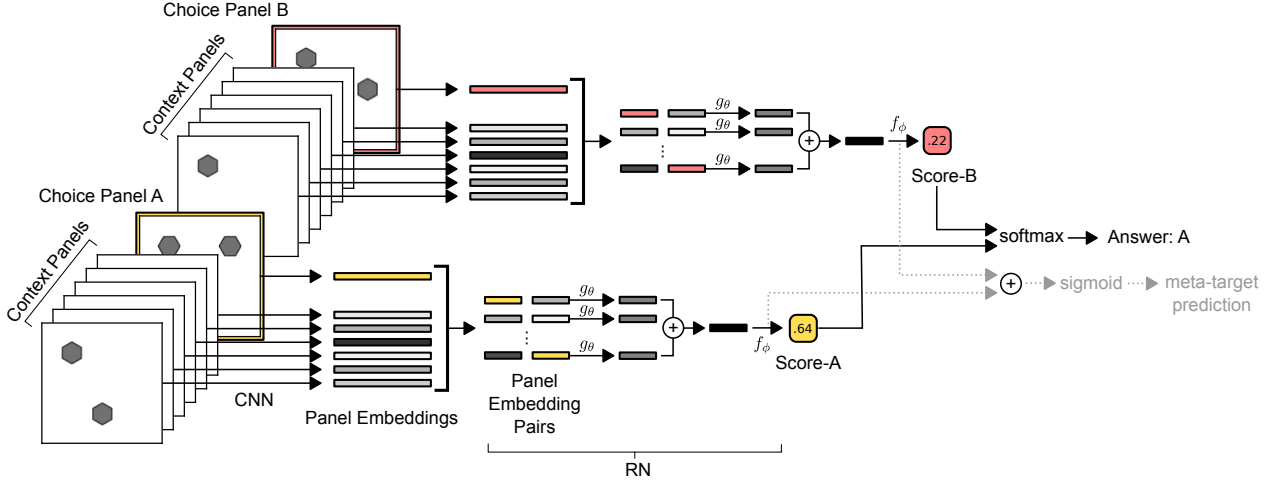


Figure 3. **WReN model** A CNN processes each context panel and an individual answer choice panel independently to produce 9 vector embeddings. This set of embeddings is then passed to an RN, whose output is a single sigmoid unit encoding the “score” for the associated answer choice panel. 8 such passes are made through this network (here we only depict 2 for clarity), one for each answer choice, and the scores are put through a softmax function to determine the model’s predicted answer.

this way, the Wild-ResNet is designed to provide a score for each candidate panel, independent of the other candidates. The candidate with the highest score is the output answer. This is similar to the WReN model described above, but using a ResNet instead of a Relation Network for computing a candidate score.

**Context-blind ResNet:** A fully-blind model should be at chance performance level, which for the PGM task is 12.5%. However, sufficiently strong models can learn to exploit statistical regularities in multiple-choice problems using the choice inputs alone, without considering the context (Johnson et al., 2017). To understand the extent to which this was possible, we trained a ResNet-50 model with only the eight multiple-choice panels as input.

### 3.1. Training on auxiliary information

We explored auxiliary training as a means to improve generalisation performance. We hypothesized that a model trained to predict the relevant relation, object and attribute types involved in each PGM might develop representations that were more amenable to generalisation. To test this, we constructed “meta-targets” encoding the relation, object and attribute types present in PGMs as a binary string. The strings were of length 12, with elements following the syntax: (shape, line, color, number, position, size, type, progression, XOR, OR, AND, consistent union). We encoded each triple in this binary form, then performed an OR operation across all binary-encoded triple to produce the meta-target. That is,  $\text{OR}([101000010000], [100100010000]) = [101100010000]$ . The models then predicted these labels

using a sigmoid unit for each element, trained with cross entropy. A scaling factor  $\beta$  determined the influence of this loss relative to the loss computed for the answer panel targets:  $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{target}} + \beta \mathcal{L}_{\text{meta-target}}$ . We set  $\beta$  to a non-zero value when we wish to explore the impact of auxiliary meta-target training.

## 4. Experiments

### 4.1. Comparing models on PGM questions

We first compared all models on the Neutral train/test split, which corresponds most closely to traditional supervised learning regimes. Perhaps surprisingly given their effectiveness as powerful image processors, CNN models failed almost completely at PGM reasoning problems (Table 1), achieving performance marginally better than our baseline - the context-blind ResNet model which is blind to the context and trained on only the eight candidate answers. The ability of the LSTM to consider individual candidate panels in sequence yielded a small improvement relative to the CNN. The best performing ResNet variant was ResNet-50, which outperformed the LSTM. ResNet-50 has significantly more convolutional layers than our simple CNN model, and hence has a greater capacity for reasoning about its input features.

The best performing model was the WReN model. This strong performance may be partly due to the Relation Network module, which was designed explicitly for reasoning about the relations between objects, and partly due to the scoring structure. Note that the scoring structure is not sufficient to explain the improved performance as

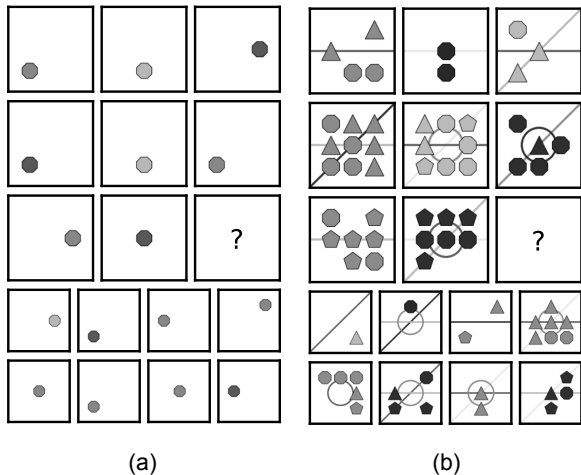


Figure 4. **The effect of distraction.** In both PGMs, the underlying structure  $S$  is  $[[\text{shape}, \text{colour}, \text{consistent union}]]$ , but (b) includes distraction on shape-number, shape-type, line-color, and line-type.

the WReN model substantially outperformed the best Wild-ResNet model, which also had a scoring structure.

#### 4.2. Performance on different question types

Questions involving a single  $[r, o, a]$  triple were easier than those involving multiple triples. Interestingly, PGMs with three triples proved more difficult than those with four. Although the problem is apparently more complex with four triples, there is also more available evidence for any solution. Among PGMs involving a single triple, OR (64.7%) proved to be an easier relation than XOR (53.2%). PGMs with structures involving lines (78.3%) were easier than those involving shapes (46.2%) and those involving shape-number were much easier (80.1%) than those involving shape-size (26.4%). This suggests that the model struggled to discern fine-grained differences in size compared to more salient changes such as the absence or presence of lines, or the quantity of shapes. For more details of performance by question type, see Appendix Tables 7, 8.

#### 4.3. Effect of distractors

The results reported thus far were on questions that included distractor attribute values (see Fig. 4). The WReN model performed notably better when these distractors were removed (79.3% on the validation and 78.3% on the test set, compared with 63.0% and 62.6% with distractors).

#### 4.4. Generalisation

We compared the best performing WReN model on each of the generalisation regimes (Table 1), and observed notable differences in the ability of the model to generalise. Interpo-

lation was the least problematic regime (generalisation error 14.6%). Note that performance on both the Interpolation and Extrapolation training sets was higher than on the neutral training set because certain attributes (size, colour) have half as many values in those cases, which reduces the complexity of the task.<sup>3</sup>

After Interpolation, the model generalised best in regimes where the test questions involved novel combinations of otherwise familiar  $[r, o, a]$  triples (Held-out Attribute Pairs and Held-out Triple Pairs). This indicates that the model learned to combine relations and attributes, and did not simply memorize combinations of triples as distinct structures in their own right. However, worse generalisation in the case of Held-out Triples suggests that the model was less able to induce the meaning of unfamiliar triples from its knowledge of their constituent components. Moreover, it could not understand relations instantiated on entirely novel attributes (Held-out line-type, Held-out shape-colour). The worst generalisation was observed on the Extrapolation regime. Given that these questions have the same abstract semantic structure as interpolation questions, the failure to generalise may stem from the model’s failure to perceive inputs outside of the range of its prior experience.

#### 4.5. Effect of auxiliary training

We then explored the impact of auxiliary training on abstract reasoning and generalisation by training our models with symbolic meta targets as described in Section 3.1. In the neutral regime, we found that auxiliary training led to a 13.9% improvement in test accuracy. Critically, this improvement in the overall ability of the model to capture the data also applied to other generalisation regimes. The difference was clearest in the cases where the model was required to recombine familiar triples into novel combinations: (56.3% accuracy on Held-out triple pairs, up from 41.9%, and 51.7% accuracy on Held-out attribute pairs, up from 27.2%). Thus, the pressure to represent abstract semantic principles such that they can be decoded simply into discrete symbolic explanations seems to improve the ability of the model to productively compose its knowledge. This finding aligns with previous observations about the benefits of discrete channels for knowledge representation (Andreas et al., 2016) and the benefit of inducing explanations or rationales (Ling et al., 2017).

#### 4.6. Analysis of auxiliary training

In addition to improving performance, training with meta-targets provides a means to measure which shapes, attributes,

<sup>3</sup>Since test questions focus on held-out phenomena, test sets in different regimes may have differing underlying complexity. Absolute performance cannot therefore be compared across different regimes.

Model	Test (%)	Regime	$\beta = 0$			$\beta = 10$		
			Val. (%)	Test (%)	Diff.	Val. (%)	Test (%)	Diff.
WReN	<b>62.6</b>	Neutral	63.0	62.6	-0.6	77.2	76.9	-0.3
Wild-ResNet	48.0	Interpolation	79.0	64.4	-14.6	92.3	67.4	-24.9
ResNet-50	42.0	H.O. Attribute Pairs	46.7	27.2	-19.5	73.4	51.7	-21.7
LSTM	35.8	H.O. Triple Pairs	63.9	41.9	-22.0	74.5	56.3	-18.2
CNN + MLP	33.0	H.O. Triples	63.4	19.0	-44.4	80.0	20.1	-59.9
Blind ResNet	22.4	H.O. line-type	59.5	14.4	-45.1	78.1	16.4	-61.7
		H.O. shape-colour	59.1	12.5	-46.6	85.2	13.0	-72.2
		Extrapolation	69.3	17.2	-52.1	93.6	15.5	-78.1

Table 1. Performance of all models on the neutral split (left), and generalisation performance of the WReN model (right) with generalisation regimes ordered according to generalisation error for  $\beta = 0$ . Context-blind ResNet generalisation test performances for all regimes is given in Table 9 of the Appendix. (**Diff**: difference between test and validation performance, H.O:“Held-out”)

and relations the model believes are present in a given PGM, providing insight into the model’s decisions. Using these predictions, we asked how the WReN model’s accuracy varied as a function of its meta-target predictions. Unsurprisingly, the WReN model achieved a test accuracy of 87.4% when its meta-target predictions were correct, compared to only 34.8% when its predictions were incorrect.

The meta-target prediction can be broken down into predictions of object, attribute, and relation types. We leveraged these fine-grained predictions to ask how the WReN model’s accuracy varied as a function of its predictions on each of these properties independently. The model accuracy increased somewhat when the shape meta-target prediction was correct (78.2%) compared to being incorrect (62.2%), and when attribute meta-target prediction was correct (79.5%) compared to being incorrect (49.0%). However, for the relation property, the difference between a correct and incorrect meta-target prediction was substantial (86.8% vs. 32.1%). This result suggests that predicting the relation property correctly is most critical to task success.

The model’s prediction certainty, defined as the mean absolute difference of the meta-target predictions from 0.5, was predictive of the model’s performance, suggesting that the meta-target prediction certainty is an accurate measure of the model’s confidence in an answer choice (Figure 5; qualitatively similar for sub-targets; Appendix Figures 6-8).

## 5. Related work

Various computational models for solving RPMs have been proposed in the cognitive science literature (see (Lovett & Forbus, 2017) for a thorough review). The emphasis in these studies is on understanding the operations and comparisons commonly applied by humans. They typically factor out raw perception in favour of symbolic inputs, and hard-code strategies described by cognitive theories. In contrast, we

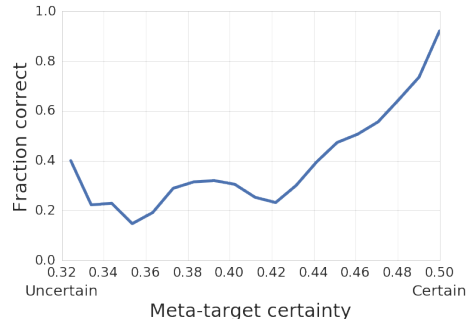


Figure 5. Relationship between answer accuracy and meta-target prediction certainty for the WReN model ( $\beta = 10$ ). The WReN model is more accurate when it is more confident about its meta-target predictions. Certainty was defined as the mean absolute difference of the meta-target predictions from 0.5.

consider models that process input from raw pixels and study how they infer, from knowledge of the correct answer, the processes and representations necessary to resolve the task. Much as we do, Hoshen & Werman (2017) trained neural networks to complete the rows or columns of Raven-style matrices from raw pixels. They found that a CNN-based model induced visual relations such as rotation or reflection, but they did not address the problem of resolving complete RPMs. Our experiments showed that such models perform poorly on full RPM questions. Moreover, Hoshen & Werman (2017) do not study generalisation to questions that differ substantively from their training data. Wang & Su (2015) present a method for automatically generating Raven-style matrices and verify their generator on humans, but do not attempt any modelling. Our method for automatically generating RPM-style questions borrowed extensively from the insights in that work.

There is prior work emphasising both the advantages (Clark & Etzioni, 2016) and limitations (Davis, 2014) of apply-



ing standardized tests in AI (see Marcus et al. (2016) and contributed articles for a review). Approaches based on standardized testing generally focus on measuring the general knowledge of systems, while we focus on models’ abilities to generalize learned information.

## 6. Discussion

One of the long-standing goals of artificial intelligence is to develop machines with abstract reasoning capabilities that equal or better those of humans. Though there has also been substantial progress in both reasoning and abstract representation learning in neural nets (Botvinick et al., 2017; LeCun et al., 2015; Higgins et al., 2016; 2017), the extent to which these models exhibit anything like general abstract reasoning is the subject of much debate (Garnelo et al., 2016; Lake & Baroni, 2017; Marcus, 2018). The research presented here was therefore motivated by two main goals. (1) To understand whether, and (2) to understand how, deep neural networks might be able to solve abstract visual reasoning problems.

Our answer to (1) is that, with important caveats, neural networks can indeed learn to infer and apply abstract reasoning principles. Our best performing model learned to solve complex visual reasoning questions, and to do so, it needed to induce and detect from raw pixel input the presence of abstract notions such as logical operations and arithmetic progressions, and apply these principles to never-before observed stimuli. Importantly, we found that the architecture of the model made a critical difference to its ability to learn and execute such processes. While standard visual-processing models such as CNNs and ResNets performed poorly, a model that promoted the representation of, and comparison between parts of the stimuli performed very well. We found ways to improve this performance via additional supervision: the training outcomes and the model’s ability to generalise were improved if it was required to decode its representations into symbols corresponding to the reason behind the correct answer.

When considering (2), it is important to note that our models were solving a very different problem from that solved by human subjects taking Raven-style IQ tests. The model’s world was highly constrained, and its experience consisted of a small number of possible relations instantiated in finite sets of attributes and values across hundreds of thousands of examples. It is highly unlikely that the model’s solutions match those applied by successful humans. This difference becomes clear when we study the ability of the model to generalise. Unlike humans, who must transfer knowledge distilled from their experience in everyday life to the unfamiliar setting of visual reasoning problems, our models exhibited transfer across question sets with a high degree of perceptual and structural uniformity. When required to

interpolate between known attribute values, and also when applying known abstract content in unfamiliar combinations, the models generalised notably well. Even within this constrained domain, however, they performed strikingly poorly when required to extrapolate to inputs beyond their experience, or to deal with entirely unfamiliar attributes.

In this latter behaviour, the model differs in a crucial way from humans; a human that could apply a relation such as XOR to the colour of lines would almost certainly have no trouble applying it to the colour of shapes. On the other hand, even the human ability to extend apparently well-defined principles to novel objects has limits; this is precisely why RPMs are such an effective discriminator of human IQ. For instance, a human subject might be uncertain what it means to apply XOR to the size or shape of sets of objects, even if he or she had learned to do so perfectly in the case of colors.

An important contribution of this work is the introduction of the PGM dataset, as a tool for studying both abstract reasoning and generalisation in models. Generalisation is a multi-faceted phenomenon; there is no single, objective way in which models can or should generalise beyond their experience. The PGM dataset provides a means to measure the generalization ability of models in different ways, each of which may be more or less interesting to researchers depending on their intended training setup and applications.

Designing and instantiating meaningful train/test distinctions to study generalisation in the PGM dataset was simplified by the objective semantics of the underlying generative model. Similar principles could be applied to more naturalistic data, particularly with crowdsourced human input. For instance, image processing models could be trained to identify black horses and tested on whether they can detect white horses, or trained to detect flying seagulls, flying sparrows and nesting seagulls, and tested on the detection of nesting sparrows. This approach was taken for one particular generalisation regime by Ramakrishnan et al. (2017), who tested VQA models on images containing objects that were not observed in the training data. The PGM dataset extends and formalises this approach, with regimes that focus not only on how models could respond to novel factors or classes in the data, but also novel combinations of known factors etc.

In the next stage of this research, we will explore strategies for improving generalisation, such as meta-learning, and will further explore the use of richly structured, yet generally applicable, inductive biases. We also hope to develop a deeper understanding of the solutions learned by the WReN model when solving Raven-style matrices. Finally, we wish to end by inviting our colleagues across the machine learning community to participate in our new abstract reasoning challenge.



## ACKNOWLEDGMENTS

We would like to thank David Raposo, Daniel Zoran, Murray Shanahan, Sergio Gomez, Yee Whye Teh and Daan Wierstra for helpful discussions and all the DeepMind team for their support.

## References

- Andreas, J., Klein, D., and Levine, S. Modular multitask reinforcement learning with policy sketches. *arXiv preprint arXiv:1611.01796*, 2016.
- Botvinick, M., Barrett, D., Battaglia, P., de Freitas, N., Kumaran, D., Leibo, J., Lillicrap, T., Modayil, J., Mohamed, S., Rabinowitz, N., et al. Building machines that learn and think for themselves: Commentary on lake et al., behavioral and brain sciences, 2017. *arXiv preprint arXiv:1711.08378*, 2017.
- Carpenter, P. A., Just, M. A., and Shell, P. What one intelligence test measures: a theoretical account of the processing in the raven progressive matrices test. *Psychological review*, 97(3):404, 1990.
- Clark, P. and Etzioni, O. My computer is an honor student-but how intelligent is it? standardized tests as a measure of ai. *AI Magazine*, 37(1):5–12, 2016.
- Davis, E. The limitations of standardized science tests as benchmarks for artificial intelligence research: Position paper. *arXiv preprint arXiv:1411.1629*, 2014.
- Fleuret, F., Li, T., Dubout, C., Wampler, E. K., Yantis, S., and Geman, D. Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences*, 108(43):17621–17625, 2011.
- Flynn, J. R. Massive iq gains in 14 nations: What iq tests really measure. *Psychological bulletin*, 101(2):171, 1987.
- Garnelo, M., Arulkumaran, K., and Shanahan, M. Towards deep symbolic reinforcement learning. *arXiv preprint arXiv:1609.05518*, 2016.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vaes: Learning basic visual concepts with a constrained variational framework. 2016.
- Higgins, I., Sonnerat, N., Matthey, L., Pal, A., Burgess, C. P., Botvinick, M., Hassabis, D., and Lerchner, A. Scan: learning abstract hierarchical compositional visual concepts. *arXiv preprint arXiv:1707.03389*, 2017.
- Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735.
- Hoshen, D. and Werman, M. Iq of neural networks. *arXiv preprint arXiv:1710.01692*, 2017.
- Jaeggi, S. M., Buschkuhl, M., Jonides, J., and Perrig, W. J. Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, 105(19):6829–6833, 2008.
- Jo, J. and Bengio, Y. Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017.
- Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., and Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pp. 1988–1997. IEEE, 2017.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Lake, B. M. and Baroni, M. Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks. *arXiv preprint arXiv:1711.00350*, 2017.
- LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436, 2015.
- Ling, W., Yogatama, D., Dyer, C., and Blunsom, P. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint arXiv:1705.04146*, 2017.
- Lovett, A. and Forbus, K. Modeling visual problem solving as analogical reasoning. *Psychological review*, 124(1):60, 2017.
- Marcus, G. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.
- Marcus, G., Rossi, F., and Veloso, M. Beyond the turing test. *Ai Magazine*, 37(1):3–4, 2016.
- Ramakrishnan, S. K., Pal, A., Sharma, G., and Mittal, A. An empirical evaluation of visual question answering for novel objects. *arXiv preprint arXiv:1704.02516*, 2017.
- Raposo, D., Santoro, A., Barrett, D., Pascanu, R., Lillicrap, T., and Battaglia, P. Discovering objects and their relations from entangled scene representations. 2017.
- Raven, J. C. et al. *Raven’s progressive matrices*. Western Psychological Services, 1938.

- Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., and Lillicrap, T. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pp. 4974–4983, 2017.
- Snow, R. E., Kyllonen, P. C., and Marshalek, B. The topography of ability and learning correlations. *Advances in the psychology of human intelligence*, 2(S 47):103, 1984.
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- Te Nijenhuis, J., Voskuijl, O. F., and Schijve, N. B. Practice and coaching on iq tests: Quite a lot of g. *International Journal of Selection and Assessment*, 9(4):302–308, 2001.
- Wang, K. and Su, Z. Automatic generation of ravens progressive matrices. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.
- Zaremba, W., Sutskever, I., and Vinyals, O. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.
- Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

## A. Appendix

### A.1. PGM Dataset

Altogether there are 1.2M training set questions, 20K validation set questions, and 200K testing set questions.

When creating the matrices we aimed to use the full Cartesian product  $\mathcal{R} \times \mathcal{A}$  for construction structures  $\mathcal{S}$ . However, some relation-attribute combinations are problematic, such as a progression on line type, and some attributes interact in interesting ways (such as number and position, which are in some sense tied), restricting the type of relations we can apply to these attributes. The final list of relevant relations per attribute type, broken down by object type (shape vs. line) is:

**shape:**

**size:** progression, XOR, OR, AND, consistent union  
**color:** progression, XOR, OR, AND, consistent union  
**number:** progression, consistent union  
**position:** XOR, OR, AND  
**type:** progression, XOR, OR, AND, consistent union

**line:**

**color:** progression, XOR, OR, AND, consistent union  
**type:** XOR, OR, AND, consistent union

Since the number and position attribute types are tied (for example, having an arithmetic progression on number whilst having an XOR relation on position is not possible), we forbid number and position from co-occurring in the same matrix. Otherwise, all other  $((r, o, a), (r, o, a))$  combinations occurred unless specifically controlled for in the generalisation regime.

We created a similar list for possible values for a given attribute:

**shape:**

**color:** 10 evenly spaced greyscale intensities in  $[0, 1]$   
**size:** 10 scaling factors evenly spaced in  $[0, 1]$ <sup>4</sup>  
**number:** 0, 1, 2, 3, 4, 5, 6, 7, 8, 9  
**position** ((x, y) coordinates in a (0, 1) plot):  
(0.25, 0.75),  
(0.75, 0.75),  
(0.75, 0.25),  
(0.25, 0.25),  
(0.5, 0.5),  
(0.5, 0.25),  
(0.5, 0.75),  
(0.25, 0.5),  
(0.75, 0.5)  
**type:** circle, triangle, square, pentagon, hexagon,

octagon, star

**line:**

**color:** 10 evenly spaced greyscale intensity in  $[0, 1]$

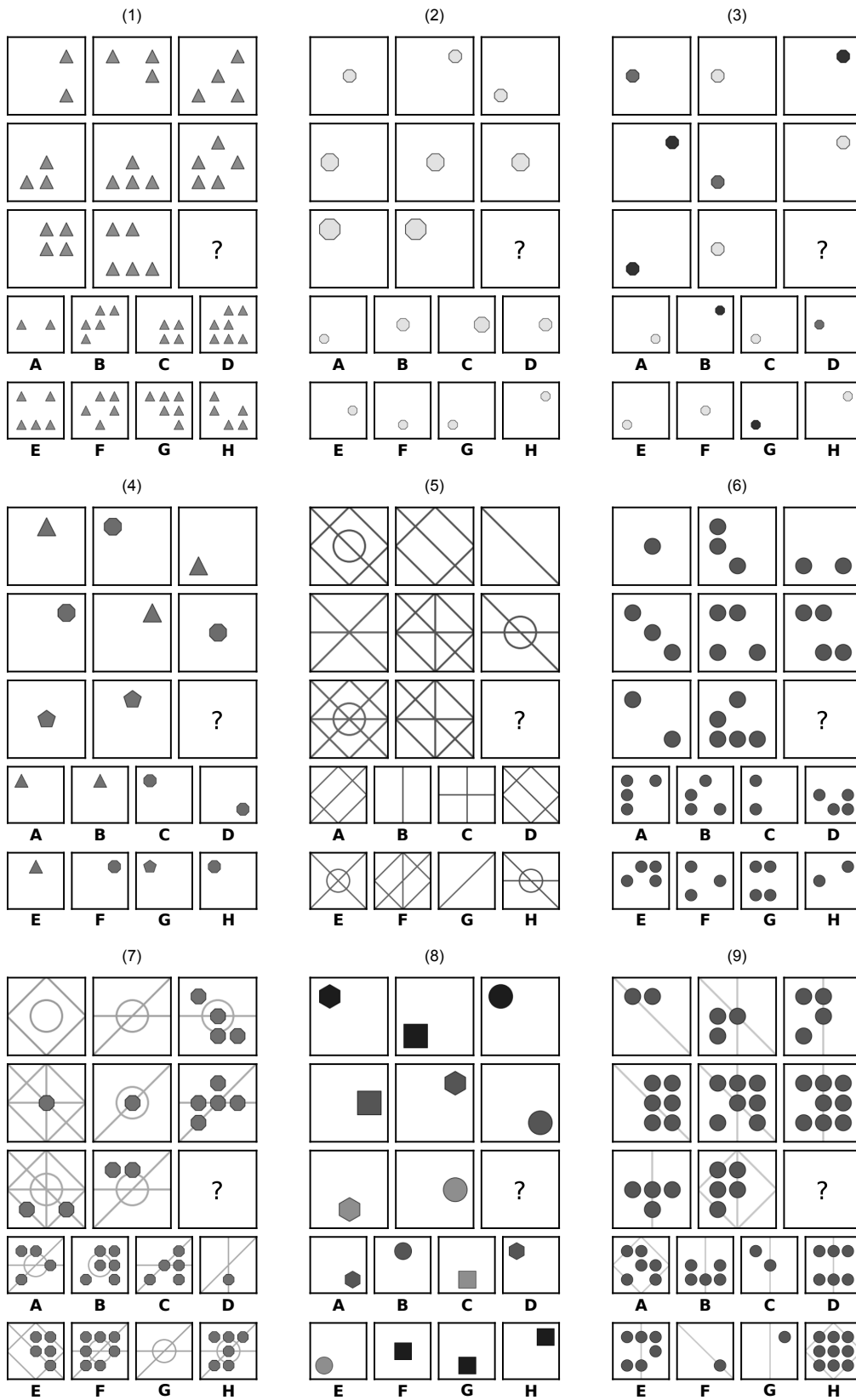
**type:** diagonal down, diagonal up, vertical, horizontal, diamond, circle

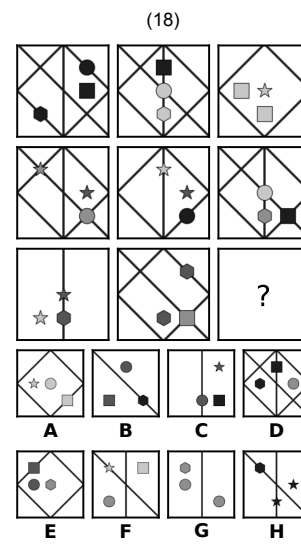
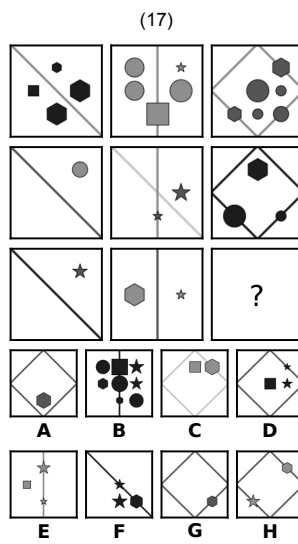
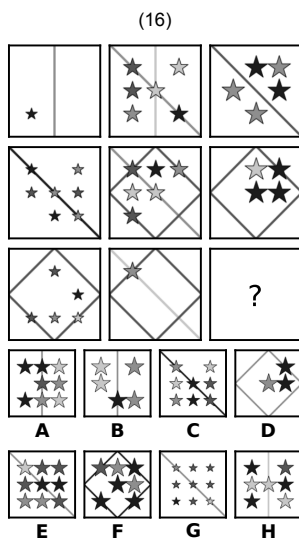
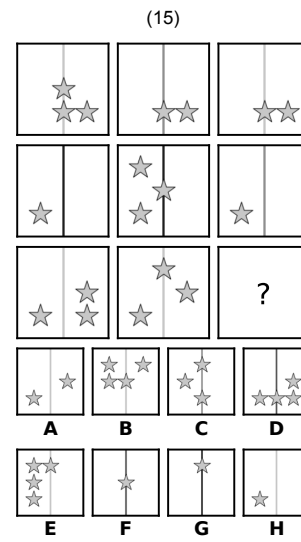
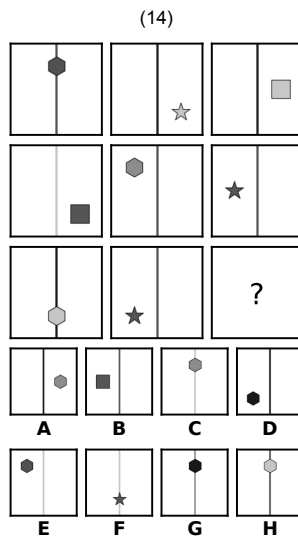
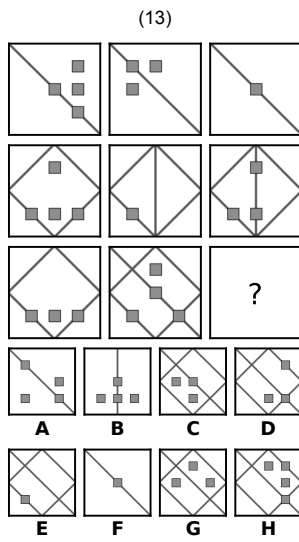
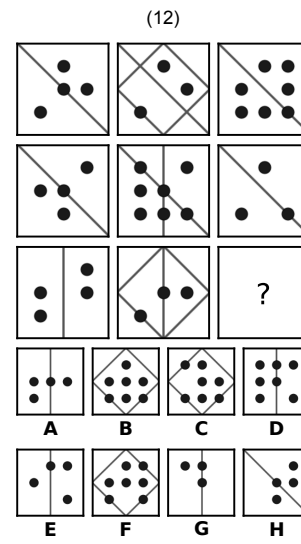
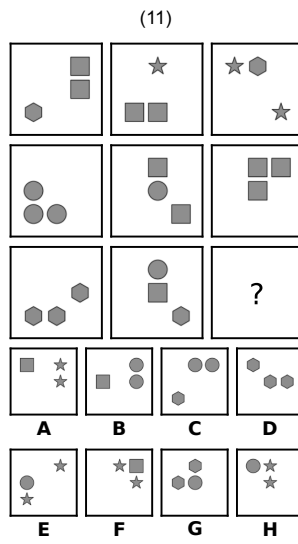
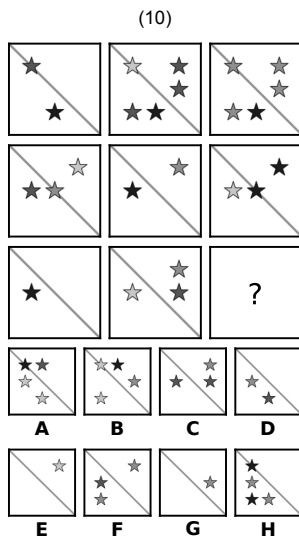
### A.2. Examples of Raven-style PGMs

Given the radically different way in which visual reasoning tests are applied to humans (no prior experience) and to our models (controlled training and test splits), we believe it would be misleading to provide a human baseline for our results. However, for a sense of the difficulty of the task, we present here a set of 18 questions generated from the neutral splits. Note that the values are filtered for human readability. In the dataset there are 10 greyscale intensity values for shape and line colour and 10 sizes for each shape. In the following, we restrict to 4 clearly-distinct values for each of these attributes. Best viewed on a digital monitor, zoomed in (see next page). Informal human testing revealed wide variability: participants with a lot of experience with the tests could score well ( $> 80\%$ ), while others who came to the test blind would often fail to answer all the questions.

---

<sup>4</sup>The actual specific values used for size are numbers particular to the matplotlib implementation of the plots, and hence depend on the scale of the plot and axes, etc.





## B. Model details

Here we provide additional details for all our models, including the exact hyper-parameter settings that we considered. Throughout this section, we will use the notation  $[x, y, z, w]$  to describe CNN and MLP size. For a CNN, this notation refers to the number of kernels per layer:  $x$  kernels in the first layer,  $y$  kernels in the second layer,  $z$  kernels in the third layer and  $w$  kernels in the fourth layer. For the MLP, it refers to the number of units per layer:  $x$  units in the first layer,  $y$  units in the second layer,  $z$  units in the third layer and  $w$  units in the fourth layer.

All models were trained using the Adam optimiser, with exponential decay rate parameters  $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$ . We also used a distributed training setup, using 4 GPU-workers per model.

	hyper-parameters
CNN kernels	[64, 64, 64, 64]
CNN kernel size	$3 \times 3$
CNN kernel stride	2
MLP hidden-layer size	1500
MLP drop-out fraction	0.5
Batch Size	16
Learning rate	0.0003

Table 2. CNN-MLP hyper-parameters

	hyper-parameters
Batch Size	32
Learning rate	0.0003

Table 3. ResNet-50 and context-blind ResNet hyper-parameters

	hyper-parameters
CNN kernels	[8, 8, 8, 8]
CNN kernel size	$3 \times 3$
CNN kernel stride	2
LSTM hidden layer size	96
Drop-out fraction	0.5
Batch Size	16
Learning rate	0.0001

Table 4. LSTM hyper-parameters

	hyper-parameters
CNN kernels	[32, 32, 32, 32]
CNN kernel size	$3 \times 3$
CNN kernel stride	2
RN embedding size	256
RN $g_\theta$ MLP	[512, 512, 512, 512]
RN $f_\phi$ MLP	[256, 256, 13]
Drop-out fraction	0.5
Batch Size	32
Learning rate	0.0001

Table 5. WReN hyper-parameters

	hyper-parameters
Batch Size	16
Learning rate	0.0003

Table 6. Wild-ResNet hyper-parameters

## C. Results

# Relations	WReN (%)	Blind (%)
One	68.5	23.6
Two	51.1	21.2
Three	44.5	22.1
Four	48.4	23.5
All	62.6	22.8

Table 7. WReN test performance and Context-Blind ResNet performance after training on the neutral PGM dataset, broken down according to the number of relations per matrix.

	WReN (%)	Blind (%)
OR	64.7	30.1
AND	63.2	17.2
consistent union	60.1	28.0
progression	55.4	15.7
XOR	53.2	20.2
number	80.1	18.1
position	77.3	27.5
type	61.0	28.1
color	58.9	18.7
size	26.4	16.3
line	78.3	27.5
shape	46.2	18.6
All Single Relations	68.5	23.6

Table 8. WReN test performance and Context-Blind ResNet performance for single-relation PGM questions after training on the neutral PGM dataset, broken down according to the relation type, attribute type and object type in a given matrix.

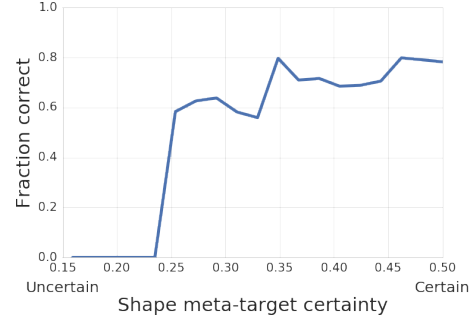


Figure 6. Relationship between answer accuracy and shape meta-target prediction certainty. The WReN model ( $\beta = 10$ ) is more accurate when confident about its meta-target predictions. Certainty was defined as the mean absolute difference of the predictions from 0.5.

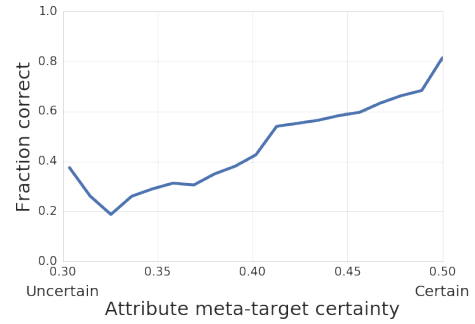


Figure 7. Relationship between answer accuracy and attribute meta-target prediction certainty

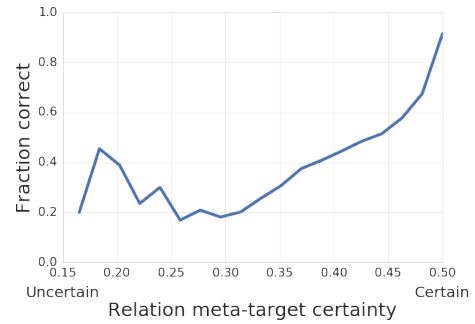


Figure 8. Relationship between answer accuracy and relation meta-target prediction certainty



Regime	Test (%)	
	$\beta = 0$	$\beta = 10$
Neutral	22.4	13.5
Interpolation	18.4	12.2
H.O. Attribute Pairs	12.7	12.3
H.O. Triple Pairs	15.0	12.6
H.O. Triples	11.6	12.4
H.O. line-type	14.4	12.6
H.O. shape-colour	12.5	12.3
Extrapolation	14.1	13.0

Table 9. Performance of the Context-blind Resnet model for all the generalization regimes, in the case where there is an additional auxiliary meta-target ( $\beta = 10$ ) and in the case where there is no auxiliary meta-target ( $\beta = 0$ ). Note that most of these values are either close to chance or slightly above chance, indicating that this baseline model struggles to learn solutions that generalise better than a random guessing solution. For several generalisation regimes such as Interpolation, H.O. Attribute Pairs, H.O. Triples and H.O. Triple Pairs the generalisation performance of the WReN model reported in Table 1 is far greater than the generalisation performance of our context-blind baseline, indicating that the WReN generalisation cannot be accounted for with a context-blind solution.

Answer Key:

- (1) G; [progression, shape, number]
- (2) C; [progression, shape, size]
- (3) D; [consistent union, shape, color]
- (4) G; [consistent union, shape, type]
- (5) H; [OR, line, type]
- (6) G; [XOR, shape, position]
- (7) H; [progression, shape, number], [OR, line, type]
- (8) C; [consistent union, shape, type], [progression, shape, color]
- (9) A; [progression, shape, number], [XOR, line, type]
- (10) C; [OR, shape, position]
- (11) B; [XOR, shape, type]
- (12) D; [consistent union, shape, number], [AND, line, type]
- (13) E; [consistent union, shape, number], [OR, line, type]
- (14) B; [consistent union, shape, type]
- (15) A; [AND, shape, position]
- (16) E; [progression, shape, size], [XOR, line, type]
- (17) D; [progression, shape, number], [AND, line, type], [consistent union, shape, color]
- (18) F; [XOR, shape, color]

Figure 9. Answer key to puzzles in section A.2