1 Appendix: Data Generation

Our artificial language can be glossed as having a lexicon of 12 words, which are presented with their English gloss in table 1. It is to be noted that the English glosses, especially for the shapes and colours, should not be taken at face value. The network is never provided with any information about what makes a triangle different from a square. A sentence always takes the form: attitude verb + colour + shape + relation + colour + shape. Accordingly, all sentences are of length 6.

The world and mind representations share a format. The format can be understood as a serialisation of a 3-by-3 grid, where each cell content is described by a single token. This token conveys both colour and shape of content or that the cell. That there is only a single token means that the network faces the challenge the mapping from the composed noun phrases, e.g. "red triangle" to a single cell token.

We only used a subset of all possible data that can be generated from all possible combinations of artificial language expressions, mind, and world representations. The data is balanced so that the model sees roughly the same number statements for each attitude verb and roughly the same number of false and true statements for each of the three attitude verbs. (The sampling of the data based on attitude verb, the train-test split, and the splitting of the train data into cross-validation folds lead to minor divergences.)

2 Appendix: Training

For training the data is split in to a training and a test set, size 633981 and 70443, respectively. On the training data, 51 hyperparameter settings were explored during a randomised grid-search of the hyperparameter space. The hyperparameter grid explored is documented in table 2. For each explored hyperparameter setting, a 5-fold cross-validation was performed on the training data. The same random validation splits were used for all hyperparameter settings, i.e. independently of the random seed.

The best hyperparameter setting according to this search was then used for evaluating on the hold-out test set. The best hyperparameters can also be found in table 2. The entire training set was used for training the model in this final evaluation.

Type	Count	English Gloss
Attitude verb	3	factive, non-factive, contrafactive
Shape	3	circle, triangle, square
Colour	3	red, blue, yellow
Relation	4	above, below, left, right

Table 1: Vocabulary of artificial language

Hyperparameter	Best	Available Settings
Random seed	348882	integers from 0 to 9999999
# encoder layers	7	5, 7, 10
Embedding size	100	100, 150, 200
Hidden layer size	50	50, 100
# attention heads	10	2, 5, 10
Prop. dropout	.2	.1,, .2, .3, .4, .5
Learning rate	1e-4	1e-2, 1e-3, 5e-4, 1e-4, 5e-5, 1e-5, 5e-6, 1e-6, 5e-7, 1e-7
# Epochs	25	5, 10, 25, 50
Batch size	5	5, 10, 50, 75, 150, 200, 250, 300

Table 2: Best and hyperparameters and search space

3 Further Architecture Details

The weights of the model were initialised with a Xavier uniform distribution. The biases were 0 initialised.

The gradients were cliped to a norm of 1.

The classification used the mean of all tokens of the last encoding layer. We used a Binary Cross Entropy Loss and the Adam optimization algorithm.