Learning Dynamics of Semantic Knowledge: An Investigation in Humans and Neural Networks

Candidate Number: 1061432

Supervisors: Dr. Andrew Saxe & Professor Christopher Summerfield

Degree Programme: Master of Science in Psychological Research (JRPS)

September 2022

Word count: 9189

Contents

List of Figures

Abstract

Semantic knowledge forms a central component of human cognition and allows the flexible acquisition, storage, and employment of conceptual representations. Previously, different theoretical models have attempted to explain a wealth of empirical phenomena documented in semantic cognition. Here, we focus on analytical results obtained in the study of semantic learning in deep linear networks. We investigate if the key phenomena of progressive differentiation and stage-like transitions are reflected in human learning of hierarchically structured semantic properties. A semantic learning experiment invoking hierarchical constraints on semantic properties was used to investigate these phenomena. Given our theoretical predictions, we analysed human data by comparing learning to several classes of differently initialised neural networks and by fitting functions to individual participants' learning trajectories. Findings indicate that human learning respects the hierarchical constraints invoked in the semantic learning task. Furthermore, we find that human learning is most closely mirrored by neural networks which learn from small random weights, which are known to display patterns of progressive differentiation and stage-like transitions. A Bayesian model comparison between participant-wise function fits revealed that, contrary to our expectations, linear functions produced superior approximation of participant learning trajectories. However, an analysis of parameters of fitted non-linear functions appears to be in accordance with results obtained in simple neural networks. Despite ambiguities, we find our results to empirically validate simple neural networks as useful models of semantic cognition.

1 Introduction

Semantic abilities form an integral component of human cognition which allow us to flexibly acquire and employ conceptual representations related to objects in the physical world (Collins & Quillian, 1969; Keil, 2013; Martin & Chao, 2001; Rogers & McClelland, 2004; Saxe et al., 2019). In the context of the parallel distributed processing (PDP) framework, Rogers and McClelland (2004) define *semantic tasks* as those which demand the retrieval of semantic information from memory in response to an object or a query. Furthermore *semantic information* is defined as information that is associated with a particular object and that is not immediately accessible from perceptual inputs alone.

Classical work in psychology and cognitive neuroscience established semantic memory and semantic knowledge as key component of human declarative memory. Semantic memory is contrasted with those in episodic memory in information content, operation, and applications (Squire, 1986; Tulving, 1984) and has been claimed as key component in the design of any intelligent system (Tulving, 1984). Semantic knowledge and conceptual ontologies are acquired progressively throughout childhood while being governed by processes independent of perceptual learning (Mandler & McDonough, 1993). How we acquire, organise, and employ semantic knowledge has been subject to extensive research documenting a host of empirical phenomena. In an attempt to frame and explain these phenomena, several theories have attempted to explain the workings of semantic systems.

In the following sections we will review empirical findings and past theories in semantic cognition before turning to analytical and empirical findings by Saxe et al. (2019) which are of particular relevance for the approach taken in the current project.

1.1 Key empirical findings

Previous studies have documented numerous empirical findings with respect to the development of semantic knowledge. Semantic knowledge often appears to be structured in a hierarchical fashion and knowledge acquisition has been argued to proceed in accordance with hierarchical taxonomies (Keil, 2013; Rogers & McClelland, 2004). Keil (2013)

postulated that semantic knowledge is organised in so called *predictability trees*. In this framework, semantic properties (or predicates) of objects are thought to be shared across objects within hierarchies. Objects are uniquely identified by their particular predicates while subsets of predicates can be shared across objects. In an anomaly study, adult subjects classified sentences generated from a previously designed predictability tree as being either normal or abnormal. The results showed that participant judgements of abnormal sentences did respect the hierarchical structure from which the stimulus sentences were generated. In a modified version of the task, children of different age groups were instructed to judge the acceptability of particular properties in reference to different objects drawn from a predictability tree. Children were, for example, asked if it was 'silly' or 'OK' to state that milk can be alive. Tree representations constructed from children's answers displayed particular developmental trajectories in their ontological distinctions. Knowledge of hierarchically organised properties developed progressively, whereby broader distinctions were attained earliest while knowledge of specific properties was attained later during development. In late development (grade 6) trees aligned perfectly with those of adult participants (Keil, 2013). These results seem to suggest that semantic knowledge is broadly organised and learned in a hierarchical fashion, whereby participants deem property assignments only as acceptable if they are associated with a particular concept in such hierarchical tree structures.

Similar results also come from the work of Mandler and McDonough (1993). In studies of 7- to 11- month old children, the authors studied the ability of infants to distinguish between objects via their semantic-conceptual properties while controlling for perceptual similarities between objects. In a familiarisation-dishabituation paradigm, infants were first tasked to examine a series of objects drawn from the same semantic category. In a following test phase, participants were given one object belonging to the same category as in the familiarisation phase and one item belonging to a different category. The authors inferred successful conceptual distinctions of the classes if infants displayed prolonged exploration of novel item at test time.

The results of such paradigms showed that infants are able to distinguish the global

domains of animal and vehicle. In contrast, infants appeared insensitive to differences within the animal category and did not distinguish between subsets of the animal class such as dogs or fish. These results persisted even when controlling for perceptual features of used stimuli (Mandler & McDonough, 1993). Mandler (2000) subsequently proposed that in conceptual-semantic categorisation tasks infants are sensitive to broad, high-level distinctions first, whereby they categorise objects using broad boundaries defined by few high-level properties. Later in development, fine-grained categories and semantic distinctions are progressively learned. The empirical results by Mandler are not without contention. Main criticisms have been concerned with the difficulty of adequately controlling for perceptual similarity of objects of different categories (Mandler, 2000; Rogers & McClelland, 2004).

A further empirical hallmark of human semantic learning is observable during the developmental acquisition of semantic knowledge. Learning of hierarchically structured semantic knowledge has been described to occur via abrupt improvements during which new levels of taxonomical structures are learned (Inhelder & Piaget, 1958; Keil, 2013; Saxe et al., 2019; Siegler, 1976). Keil (2013) found evidence for such stage-like transitions in studies on child development. Specific groups of children were found to be in a state of transition between semantic taxonomies as identified by particular error patterns in a predictability tree task. Keil (2013) assumed that a particular reason for the rarity of such error patterns might be the speed of transitions when acquiring a new part of the semantic structure. Rapid stage-like transitions during semantic learning also explain the grave changes in predictability trees seen in differently aged children. Similarly, Carey (1985) identified transitions in conceptual knowledge via the inductive judgements of children during the learning of novel animal properties. Children of different age groups and adults displayed differential profiles when generalising newly learned properties to other animals. Younger children displayed strong tendencies to overgeneralise learned properties during generalisation tasks. Given these different patterns of induction Carey (1985) argued that children appear to undergo rapid stages of conceptual reorganisation when

acquiring new semantic knowledge.

A further empirical phenomenon in the acquisition of semantic knowledge relates to characteristic inductive failures – also termed illusory correlations. During the acquisition of semantic knowledge, children often ascribe to false beliefs or illusory correlations about the semantic properties of particular object classes (Carey, 1985; Saxe et al., 2019). Crucially, these illusory correlations are connected to the hierarchical structure of semantic knowledge. Carey (1985) captured this phenomenon well in their study of property inference in differently aged children. When making judgements about unfamiliar animals, younger children drew inferences that were closely connected to the perceived similarity of these animals to humans. Older children and adult participants organised their inferences based on perceived categorical taxonomies (categorising animals as vertebrate or invertebrate). The resulting false beliefs in the inferences of younger children therefore reflect false beliefs connected to the hierarchical conceptual groupings learned during semantic development.

These three key phenomena – progressive differentiation, stage-like transitions, and illusory correlations – must be seen as separate but connected phenomena that largely stem from the progressive learning of hierarchical semantic ontologies (Keil, 2013; Rogers & McClelland, 2004). Several theoretical frameworks have attempted to account for these and other empirical findings in the study of semantic knowledge (Carey, 1985; Collins & Quillian, 1969; Keil, 2013; Rogers & McClelland, 2004).

1.2 Previous theoretical frameworks

Several classical theoretical frameworks have attempted to explain the workings of semantic memory and cognition. We can categorise theories of semantic cognition into the three dominant frameworks: categorisation based approaches, the theory-theory framework, and – of particular relevance for the current project – the parallel distributed

processing (PDP) approach.

The categorisation framework assumes that semantic knowledge and cognition are mediated via categorisation of conceptual classes (Rogers & McClelland, 2004). These approaches assume – at least implicitly – that categorical representations are mediating storage, access, and retrieval of semantic information. Categorisation based theories propose that the classification of encountered objects subsequently allows access to stored semantic knowledge (Collins & Quillian, 1969; Keil, 2013; Rogers & McClelland, 2004). To adequately explain a wealth of empirical phenomena, the categorisation based framework frequently invokes three constraints: hierarchically structured storage of semantic information (Keil, 2013), the assumption that objects are categorised in a behaviourally useful fashion (Rosch, 1978), and category prototypes which allow for graded category membership (Rosch, 1975). These three different versions of the categorisation based approach to semantic cognition can explain many empirical phenomena, such as illusory correlations or progressive differentiation in childhood. However, as pointed out by Rogers and McClelland (2004) no single sub-theory can capture all empirical phenomena adequately. In addition, the categorisation based approach gives no adequate neural or mechanistic explanation through which semantic knowledge can be acquired and stored.

The theory-theory framework by Murphy and Medin (1985) presents a different perspective on the mechanisms underlying semantic cognition. The framework understands semantic knowledge not as a hierarchical system which is activated in response to cues or objects, but as the result of domain knowledge or theories. These theories in turn govern the properties ascribed to entities within a particular domain (Murphy & Medin, 1985; Rogers & McClelland, 2004). Such theories about the causal structure of the world allows agents to predict and explain observations and to make inductions about the semantic properties of encountered objects (Gopnik & Meltzoff, 1997; Murphy & Medin, 1985). In essence, theory-theory approaches remove the task of retrieving semantic knowledge from specific storage systems that are organised with respect to categories of objects and

assigns this task to causal theories. These theories are then activated when encountering a particular set of features (Murphy & Medin, 1985). With respect to the acquisition of semantic knowledge, the theory-theory framework emphasises that early proto-theories may guide judgements in early development and that knowledge grows in accordance with developing causal theories (Murphy & Medin, 1985; Rogers & McClelland, 2004). The idea of causal theories can find support in studies which documented that children display core knowledge in several domains early in life (Spelke & Kinzler, 2007). Additionally, research in artificial intelligence has emphasised the importance of causal theories for the emergence of intelligence (Lake et al., 2017; Pearl & Mackenzie, 2018). Criticisms of the theory-theory approach have emphasised that mechanisms of learning are not well defined and that the key idea of a conceptual theory lacks a formal definition (Rogers & McClelland, 2004). The theory-theory approach provides a conceptual compelling account of semantic cognition while failing to provide well-defined mechanisms of learning or storage that can be computationally implemented (Rogers & McClelland, 2004; Saxe et al., 2019).

In response to these theoretical criticisms, proponents of the parallel distributed processing approach (PDP) (Rogers & McClelland, 2004) developed a mechanistic account of semantic cognition that employs a connectionist approach. The theoretical framework by Rogers and McClelland (2004), demonstrates through model simulations that a multilayer neural network with non-linear activation functions can attain semantic knowledge. Such networks are trained via corrective adjustment of synaptic weights through the back-propagation of error signals in the network (Rogers & McClelland, 2004; Rumelhart, Todd, et al., 1993). In the network proposed by Rogers and McClelland (2004) each input unit to the network represents a particular class of items and output units in the final network layer represent object properties. Thus one-hot input vectors $\mathbf{x} \in \mathbb{R}^{N_1}$ represent individual objects. After a forward pass through adjustable weights and nonlinear activation function the network produces an output vector of properties $\hat{\mathbf{y}} \in \mathbb{R}^{N_3}$. The network then learns by comparing the produced output pattern $\hat{\mathbf{y}}^i$ for an example $\mathbf{x}^i$ to a target pattern $\mathbf{y}^i \in \mathbb{R}^{N_3}$ to reduce the squared error $\|\mathbf{y}^i - \hat{\mathbf{y}}^i\|^2 = (\mathbf{y}^i - \hat{\mathbf{y}}^i)^T(\mathbf{y}^i - \hat{\mathbf{y}}^i)$

or alternatively cross-entropy loss. As the squared error is a function of the network's synaptic weights, we can then compute the partial derivative with respect to each weight in the network and adjust synaptic weights to reduce discrepancies between output and target patterns (Rogers & McClelland, 2004; Rumelhart et al., 1985). As a consequence, learned representations in network weights will be able to retrieve object properties when presented with a particular object. This simple learning mechanism lies at the heart of modern machine learning (Bishop & Nasrabadi, 2006; Schmidhuber, 2015) and is capable to explain a surprising range of empirical phenomena in the literature on semantic cognition. Rogers and McClelland (2004) were able to reproduce phenomena such as progressive differentiation, coherent categories, or illusory correlations. Importantly, the PDP was able to demonstrate these capabilities while providing a concrete model architecture and a learning rule which attains semantic knowledge in end-to-end closed loop learning.

1.3 Semantic learning in deep linear networks

More recently, Saxe et al., 2019 studied the development of semantic cognition in deep linear networks. In contrast to the connectionist deep networks by Rogers and McClelland (2004), deep linear networks omit non-linear activation function in their network architecture. Importantly, deep linear networks allow for in depth analysis of theoretical principles that drive learning in deep neural networks while preserving many of the hallmarks of semantic cognition (Saxe et al., 2019; Saxe et al., 2021). Removing non-linear activation functions makes network learning mathematically tractable and allows for exact analytical solutions to learning dynamics that usually cannot be solved in the non-linear case.

Similar to Rogers and McClelland (2004), the deep linear network is required to map one-hot input items $\mathbf{x} \in \mathbb{R}^{N_1}$ to output properties $\mathbf{y} \in \mathbb{R}^{N_3}$ where the model produces a prediction vector $\hat{\mathbf{y}} \in \mathbb{R}^{N_3}$ after receiving a given input. A full set of P examples thus consists of $\{(\mathbf{x}^i, \mathbf{y}^i)\}_{i=1}^{P}$ input and output vectors where i denotes a particular example.

The network is parameterised by weight matrices $\mathbf{W}^1 \in \mathbb{R}^{N_2 \times N_1}$ and $\mathbf{W}^2 \in \mathbb{R}^{N_3 \times N_2}$ given the omission of nonlinearities outputs are simply computed as $\hat{\mathbf{y}} = \mathbf{W}^2\mathbf{W}^1\mathbf{x}$ yielding the desired output of size N_3. Here $\mathbf{h} = \mathbf{W}^1\mathbf{x}$ is a vector of hidden activations of size N_2 (Saxe et al., 2019). The loss function for each example thus becomes $J(\mathbf{W}^1, \mathbf{W}^2) = \frac{1}{2}\|\mathbf{y}^i - \hat{\mathbf{y}}^i\|^2$. The total error on our data set is given by $J_{total}(\mathbf{W}^1, \mathbf{W}^2) = \sum_{i=1}^{P} \|\mathbf{y}^i - \hat{\mathbf{y}}^i\|^2$. Saxe et al. (2019) then derive the learning rules by computing the derivative of the loss $J(\mathbf{W}^1, \mathbf{W}^2)$ with respect to the weight matrices as

$$\Delta\mathbf{W}^1 = \lambda\mathbf{W}^{2^T}(\mathbf{y}^i - \hat{\mathbf{y}}^i)\mathbf{x}^{iT},\ \Delta\mathbf{W}^2 = \lambda(\mathbf{y}^i - \hat{\mathbf{y}}^i)\mathbf{h}^{iT} \tag{1}$$

with λ signifying the learning rate. Under the assumption of a small, constant learning rate and small initial weights, the model learning dynamics then become represented in a time dependent singular value decomposition of the input-output correlation matrix

$$\hat{\Sigma}^{yx}(t) = \mathbf{U}\mathbf{A}(t)\mathbf{V}^T = \mathbf{W}^2(t)\mathbf{W}^1(t). \tag{2}$$

In this formulation, time-dependent singular values $a_1(t), ..., a_4(t)$ on the diagonal of $\mathbf{A}(t)$ obey exact analytical solutions reflective of the models learning trajectory and error patterns. Hallmarks of semantic cognition such as progressive differentiation, stage-like transitions, and illusory correlations were qualitatively reflected in the trajectories of singular values and in our time dependent input output correlation matrix $\hat{\Sigma}^{yx}(t)$. The analysis revealed that progressive differentiation proves inevitable in the learning of deep networks when tasked with learning hierarchically structured taxonomies (Cao et al., 2020; Saxe et al., 2019; Saxe et al., 2021). Stage-like transitions are driven by improvements in performance on particular levels in the hierarchy that are attained while overcoming saddle points in the non-convex loss landscape. This is reflected in effective singular values $a_1(t), ..., a_4(t)$. Lastly, of importance for the current project are illusory correlations. Saxe et al. (2019) note these to result from differently signed contributions of left and right singular vectors $\mathbf{u}_m^\alpha \mathbf{v}_i^\alpha$ for an item i and feature m which are scaled by evolving singular

values $a_\alpha(t)$ as $a_\alpha(t)\mathbf{u}_m^\alpha \mathbf{v}_i^\alpha$. The non-monotonic increase of singular values will inevitably result in illusory correlations in a hierarchically structured data set.

Beyond semantic learning, deep neural networks are now routinely employed as models of higher-level primate cognition. Networks are studied with respect to biological behaviour, neural representation, and neural computation (Hassabis et al., 2017; Khaligh-Razavi & Kriegeskorte, 2014; Lake et al., 2017; Saxe et al., 2021; Whittington & Bogacz, 2019). Importantly, work by Saxe et al. (2019) trained neural networks from small random weights. Recently, work by Flesch et al. (2021) revealed the importance of small initialisation in the comparison of deep neural networks with human data. Neural representations in networks trained from small random weights – termed the *rich* learning regime – displayed larger similarity to human neural geometry compared to networks trained from larger initial weights – termed the *lazy* learning regime. We thus suspect that an interplay of network architecture and initialisation is also important when comparing network and human behaviour. Crucially, much insight in semantic cognition focuses on long times-scales during which humans progressively acquire abstractions that form building blocks for their semantic abilities (Keil, 2013; Saxe et al., 2021). Similar to results in deep linear networks (Saxe et al., 2019), we focus on the acquisition of novel semantic knowledge in a learning task that spans a shorter time-scale. To this end, we use the analysis of learning dynamics in deep linear networks to derive predictions about learning dynamics in human participants.

1.4 Current predictions

Given the vast experimental literature on semantic cognition and the theoretical results obtained by Saxe et al. (2019), we seek to examine the learning dynamics of semantic information in human participants. Thereby key aspects of learning dynamics observed empirically and analytically in neural networks (Rogers & McClelland, 2004; Saxe et al., 2019) will serve as rich predictions for our behavioural paradigm.

While we focus on learning dynamics, two key research questions guide our analysis. First, we want to assess if we can empirically observe hierarchical differentiation in human behaviour over the course of learning. In particular, we ask if human learning respects the statistical regularities of a hierarchically structured data-set in a semantic learning task.

Secondly, we seek to examine if rapid stage-like transitions are reflected in human learning dynamics and if such changes stem from short bursts of improvements with respect to each category. Can we observe swift drops in error and do these drops stem from improvement on properties at particular levels of the semantic hierarchy?

2 Methods

2.1 Participants

We recruited a cohort of n = 49 adult participants (33 female, 13 male, and 2 other, self-reported) via the online platform Prolific (https://www.prolific.co/). Participant age ranged from 18 to 40 years with a mean age of 30.61 years. Participants were compensated with £7 plus a potential bonus of up to £3 given performance on their first attempts at solving each trial to incentivise learning. All participants provided informed consent before taking part in the experiment and the study was approved by the University of Oxford Central University Research Ethics Committee (reference number: R50750/RE004).

2.2 Stimuli

Visual stimuli used in the experiment consisted of a set of 8 classes of planets that were each defined by a set of features. Stimuli were generated with a custom planet generator developed by Sheahan (2021). Planet appearance was varied according to 7 distinct, continuous visual features such as the size of planetary rings, the height of planet mountains, or number of moons. Given our aim of teaching participants semantic properties associated with classes of stimuli, we defined stimulus categories as normal distributions in this feature space from which we drew exemplars for training. To maintain distinctiveness between planet categories, we optimised the distance of category means in this high dimensional space while fixing the distribution variance. More formally, given a planet category let $\mathbf{X} \in \mathbb{R}^{7\times n}$ denote the matrix of feature values for all n exemplars. Each column of this matrix $\mathbf{x}_i \in \mathbb{R}^7$ for $i = 1, ..., n$ then represents the specific features of a given exemplar planet i. The features of each planet in category p are then distributed according to a 7-dimensional normal distribution

$$\mathbf{x}_i \sim \mathcal{N}_7(\boldsymbol{\mu}^{(p)}, \boldsymbol{\Sigma}^{(p)}) \quad \text{with} \quad \boldsymbol{\Sigma}^{(p)}_{i,j} = 0, \text{ if } i \neq j \tag{3}$$

Here $\boldsymbol{\mu}^{(p)} \in \mathbb{R}^7$ denotes the vector of means for category p and $\boldsymbol{\Sigma}^{(p)} \in \mathbb{R}^{7\times 7}$ denotes

the respective covariance matrix. The diagonal values of $\mathbf{\Sigma}^{(p)}$ will then contain our fixed variance for each feature. This definition of categories is intended to facilitate learning of properties with respect to categories of stimuli and to prevent the memorisation of small idiosyncrasies in stimulus appearance during the learning task. Two example planets from different categories can be seen in Figure 1. A complete depiction of one example from each class can be found in Appendix A.

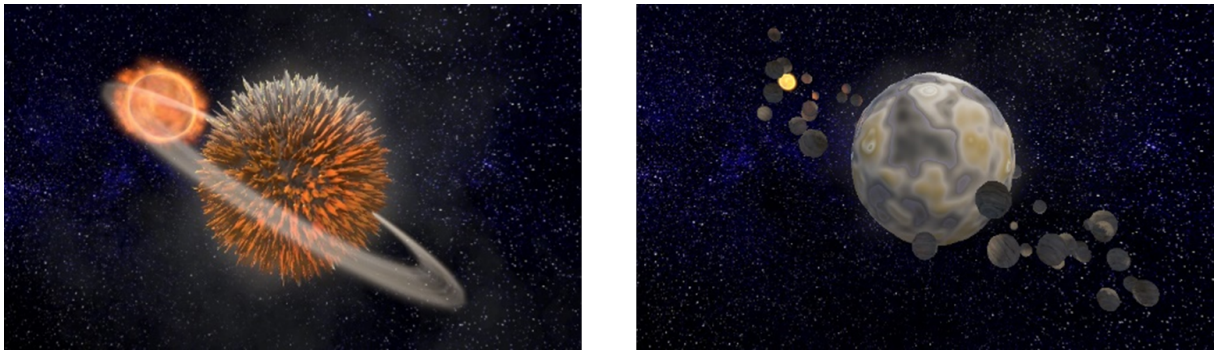


Figure 1: Two example planets generated with different mean vectors $\boldsymbol{\mu}^{(p)}$, each drawn from a normal distribution.

To find the appropriate values for our $\boldsymbol{\mu}^{(p)}$ for $p = 1, ..., 8$ we maximised the minimal distance between these mean vectors. This leads us to the following optimisation problem. Consider the means $\boldsymbol{\mu}^{(p)}$ for all our 8 categories $\boldsymbol{\mu}^{(1)}, ..., \boldsymbol{\mu}^{(8)}$. We can build the set of Euclidean distances of these vectors as $D = \{x \mid x = \|\boldsymbol{\mu}^{(i)} - \boldsymbol{\mu}^{(j)}\|, \text{ for } i \neq j \text{ and } i, j = 1, ..., 8\}$ then our optimisation problem becomes

$$\begin{gathered} \underset{\boldsymbol{\mu}^{(1)},...,\boldsymbol{\mu}^{(8)}}{\operatorname{argmax}} f(\boldsymbol{\mu}^{(1)}, ..., \boldsymbol{\mu}^{(8)}) \quad \text{where} \quad f(\boldsymbol{\mu}^{(1)}, ..., \boldsymbol{\mu}^{(8)}) = min(D) \\ \text{subject to} \quad \mathbf{b}_i^{(1)} \leq \boldsymbol{\mu}_i^{(p)} \leq \mathbf{b}_i^{(2)} \text{ for } p = 1, ..., 8. \end{gathered} \tag{4}$$

Here, the inequality constraints given by $\mathbf{b}^{(1)}, \mathbf{b}^{(2)} \in \mathbb{R}^7$ signify the range of values that each feature can take. By maximising the minimum of the set D we can find the mean vectors that are maximally distance in feature space. We chose to use SciPy 1.8.0 (Virtanen et al., 2020) to find $\boldsymbol{\mu}^{(1)}, ..., \boldsymbol{\mu}^{(8)}$ in the above optimisation problem. As values in the vectors were naturally bounded by the range of values engineered by Sheahan (2021),

we employed the L-BFGS-B algorithm to perform the procedure (Zhu et al., 1997). Given upper and lower bounds, we initialised our estimates as $\boldsymbol{\mu}_i^{(p)} \sim U(\mathbf{b}_i^{(1)}, \mathbf{b}_i^{(2)})$. We ran the optimisation for 1000 iterations and chose the vectors which converged on the largest value. The resulting mean vectors where then used for the generation of planet classes.

2.3 Semantic properties and hierarchical structure

We chose pseudo-words as the semantic properties that had to be learned by our participants. Pseudo-words were chosen from The Novel Object and Unusual Name Database (NOUN) (Horst & Hout, 2016). Word stimuli were matched for word length and number of vowels. For each participant, we randomly assigned words and classes of generated images to locations in the hierarchical structure. To remove the presence of a unique class identifier as usually present in symmetric tree structures (Figure 2A), we assigned two classes of visual stimuli to each walk in the graph (Figure 2B). Note that this permits two interpretations of the ensuing learning problem. Participants are learning two hierarchical structures in parallel, or participants learn a 4-level semantic graph in which the leaf properties are omitted (see Figure 2).

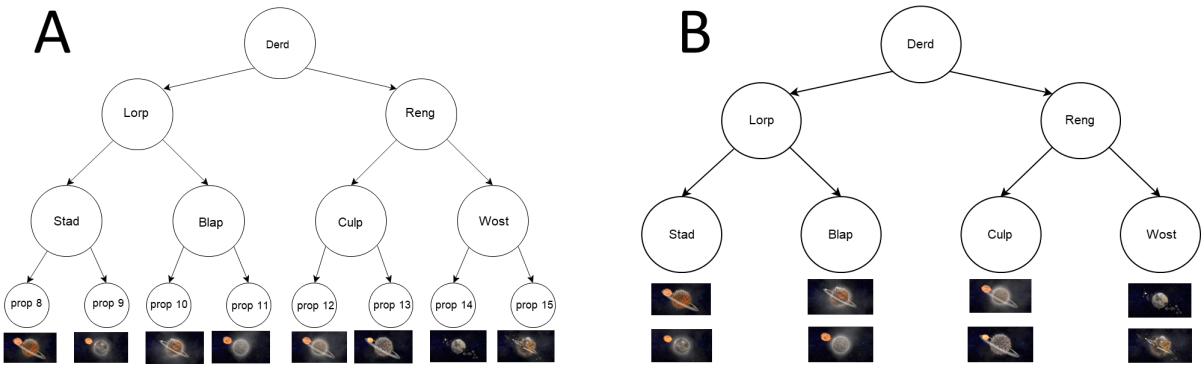


Figure 2: (A) A Hierarchical structure of semantic properties with the leaf level present. Note how each walk in the graph from top to bottom defines the properties assigned to each class. (B) The hierarchical structure used in the current task. Two classes of stimuli are underneath each walk in the graph giving them identical semantic properties. Labels in the hierarchy were randomised and are solely for illustration purposes.

2.4 Design

We set out with the intention to study learning dynamics of participants in a hierarchically structured semantic data-set. The key dependent variable in this context is participant accuracy on the task. The independent variables are the temporal trajectory of learning (i.e. over blocks or trials), as well as the different positions of properties in our semantic hierarchy. This somewhat unorthodox within-subjects design is not comparing different conditions per-se but examines the participant accuracy on semantic properties as a function of learning time and position in the hierarchical structure.

2.5 Task and procedure

The experiment was conducted online and the use of the web platform Prolific (https://www.prolific.co) allowed for completely remote participation. Experiments began with written instructions on screen. After an introduction about the general nature of the task participants provided informed consent. Subsequently participants were given specific instructions via screenshots from a representative trial with written explanations on screen. Full instructions can be found in Appendix B. Participants were able to read the instructions at their own pace and could start the experiment when ready. Importantly, participants were informed that they would have to repeat trials until correct, but that only performance on their first attempt would count towards their bonus payment. The experiment consisted of 8 blocks with each block containing 16 trials of learning. On each trial participants first viewed an image of a planet. Shortly after buttons would appear below the image. Participants then selected three buttons which were labelled with their respective semantic properties. After selection participants received feedback if their selection was correct. If their selection was incorrect the correct selection of button was highlighted and a red text would appear on screen informing them of their incorrect choice. The timing of an experimental trial can be found in Figure 3. Participants were then forced to repeat the trial until successful. The position of the buttons on the screen was reshuffled for each trial as well as for each new attempt. This prevented par-

ticipants to memorise spatial locations rather than button labels. An illustration of an experimental trial can be seen in Figure 3. We shuffled the order of trial unique stimuli across participants. Similarly, the assignment of labels to levels in the semantic hierarchy was randomised across participants. After each block we informed participants of their obtained bonus in this block which was calculated as the proportion of correct responses in each block.

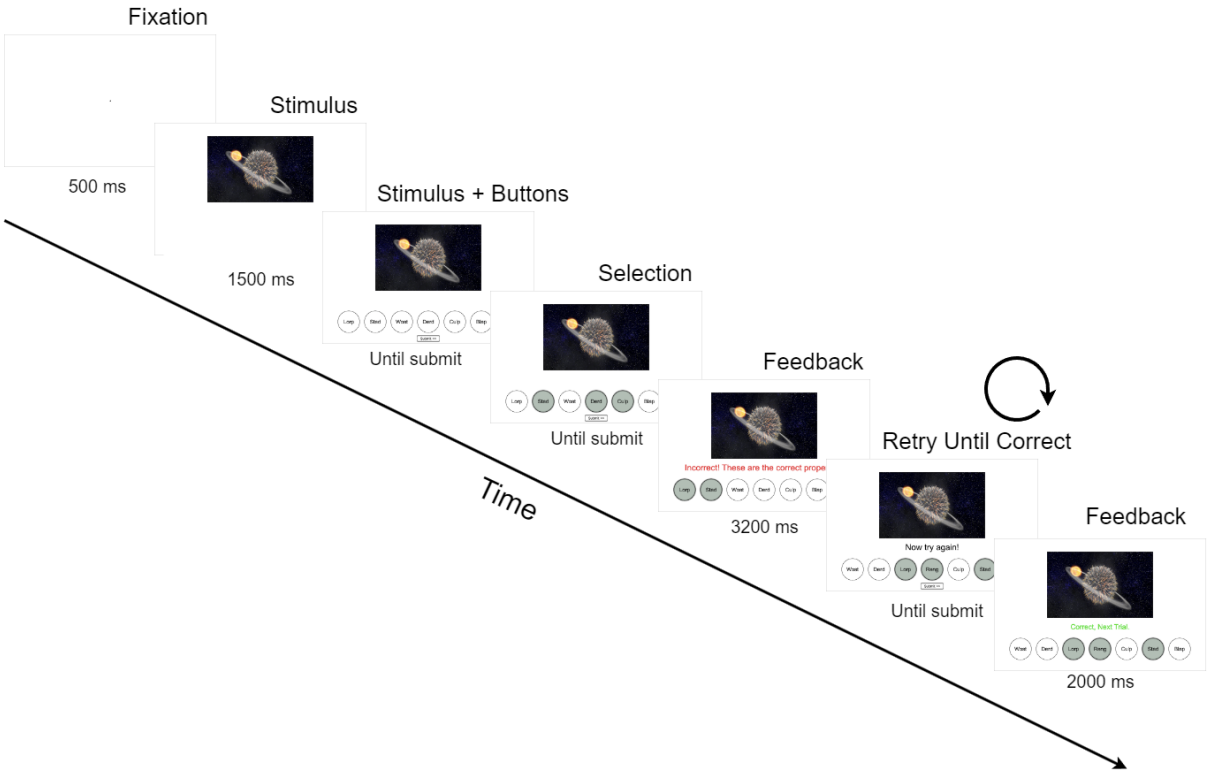


Figure 3: An example experimental trial the circular arrow indicates forced repetition.

2.6 Network simulations

To assess correspondence of our participant responses with results obtained in deep linear networks, we reproduced model simulations in deep linear networks. Results obtained by Saxe et al. (2019) assessed learning dynamics in linear networks trained from small initial weights. Given this constraint on analytical learning dynamics, we initialised linear and non-linear networks with a range of different weights to find best correspondence to human learning. Further motivation for these experiments stems from recent work on

differences in learning and internal network representations in networks initialised from different norm weights (Advani et al., 2020; Flesch et al., 2021; Woodworth et al., 2020). Learning dynamics of humans might in part be reflected in their correspondence to different initialisation schemes. In total, we performed experiments using four different network architectures and six different configuration of initial weights. All networks contained a single hidden layer with 16 units, 4 input units, and 7 output units. Our linear network was constructed as follows. Given the task of predicting semantic properties $\mathbf{y} \in \mathbb{R}^7$ from given input classes $\mathbf{x} \in \mathbb{R}^4$ the network computes $\hat{\mathbf{y}} = \mathbf{W}^2\mathbf{W}^1\mathbf{x}$ with $\mathbf{W}^1 \in \mathbb{R}^{16\times4}$ and $\mathbf{W}^2 \in \mathbb{R}^{7\times16}$. Secondly, we trained a network with ReLU activation functions in the hidden layer which are defined for each element in the input vector as $ReLU(x) = max(0, x)$. The network computes $\hat{\mathbf{y}} = \mathbf{W}^2 ReLU(\mathbf{W}^1\mathbf{x})$. The final two networks trained were identical to the previous networks but contained sigmoid non-linearities in the output layer, i.e. $\hat{\mathbf{y}} = \sigma(\mathbf{W}^2\mathbf{W}^1\mathbf{x})$ and $\hat{\mathbf{y}} = \sigma(\mathbf{W}^2 ReLU(\mathbf{W}^1\mathbf{x}))$.

Networks were initialised with weights drawn from a zero mean Gaussian distribution as

$$\begin{aligned} \mathbf{W}^1_{ij} &\sim \mathcal{N}(0, a_0^2/N_1) \\ \mathbf{W}^2_{ij} &\sim \mathcal{N}(0, a_0^2/N_3) \end{aligned} \tag{5}$$

We initialised these distributions with $a_0^2 \in \{1, 0.1, 0.01, 0.001, 0.0001, 0.00001\}$ for our six configuration of initial weights. Each network was trained with batch gradient descent. The number of epochs differed between networks trained with and without sigmoid activation functions. We found networks with activation functions to generally converge slower than networks without such non-linearity. This effect is in part driven by generally slower learning in networks with sigmoid activation functions. The derivative in sigmoid functions attains their maximum at $\max_x \frac{d}{dx}\sigma(x) = \max_x \sigma(x)(1 - \sigma(x)) = 0.25$ while linear activation functions have a constant derivative at $\frac{d}{dx}f(x) = 1$. To achieve approximate convergence on all networks, we trained the networks without sigmoid activation functions for a total of 800 epochs and networks with sigmoid output units were trained for a total of 4000 epochs. All networks were trained with a learning rate of

$\lambda = \frac{1}{P}$ with $P = N_1 = 4$. We ran each network with each initialisation scheme for a total of 49 runs to control for idiosyncratic learning trajectories stemming from a particular stochastic weight initialisation. In addition, this procedure gave us the same number of human participants and trained neural networks for analysis. Full training data can be found in Appendix C. Full software details can be found in Appendix D.

2.7 Trialwise function fitting

In order to adequately analyse participant learning dynamics, we fit a set of different functions to individual participants performance across trials. The reasoning behind this approach acknowledges that aggregate statistics of accuracy will often obscure stage-like transitions in learning. In other words, while participants may display learning trajectories with short bursts of improvement these may occur at different time points in training. Therefore group statistics may not reflect these individual trajectories. We decided to fit four different function types which we thought to plausibly describe learning trajectories to each participant's data across the three levels of the hierarchy and subsequently assessed their fit. Our chosen functions consisted of a simple linear function

$$f_{linear}(x, w, b) = wx + b. \tag{6}$$

Here slope is controlled by w and and the function off-set is controlled by the bias term b. We also fit a variety of a ReLU function as

$$f_{Relu}(x, w, b_1, b_2) = max(0, wx - b_1) \ + \ b_2. \tag{7}$$

Here the function slope is controlled by w. The parameters b_1 and b_2 control the horizontal and vertical off-set of the function from the origin. We further fitted a sigmoidal function with variable slope and offset as

$$f_{sigmoid}(x, b, k) = \frac{1}{1 + e^{-k(x-b)}}. \tag{8}$$

Here the parameter k controls the steepness of the function around its inflection point while the parameter b controls the horizontal offset of the inflection point around the origin. The discontinuous step function with linear region is defined as

$$f_{step}(x, b_1, b_2) = \begin{cases} 0, & x < b_1 \\ \frac{x}{b_2 - b_1} + \frac{b_1}{b_1 - b_2}, & b_1 \leq x \leq b_2 \\ 1, & b_2 < x \end{cases} \tag{9}$$

Here the offset and slope of the constant sections of this piece-wise function are controlled by the parameters b_1 and b_2. We fit these functions to individual participants for responses on each level of the hierarchy. Data for participant $p = 1, ..., n$ with $n = 49$ for each trial $i = 1, ..., m$ with $m = 128$ on level $j = 1, ..., k$ with $k = 3$ of the hierarchy is therefore in $x^p_{ji} \in \{0, 1\}$. To obtain fits we used the SciPy with the Dogleg algorithm for least squares minimisation (Voglis & Lagaris, 2004). The minimisation procedure was carried out with cross-validation. Functions were fit on even trials and evaluated on odd trials. Full software details can be found in Appendix D. To assess fit of the different models we interpreted model outputs as probabilities $P(x^p_{ji} = 1 \mid \theta_p) = f(x^p_{ji}, \theta_p)$ where θ_p indicates the best fitting parameters for a given function and participant.

For each participant, for each level of the hierarchy, and each trial we calculated

$$P(x^p_{ji} \mid \theta_p) = \; P(x^p_{ji} = 1 \mid \theta_p)^{x^p_{ji}} \, (1 - P(x^p_{ji} = 1 \mid \theta_p))^{(1 - x^p_{ji})} \tag{10}$$

as the likelihood of the response x^p_{ji}. The log-likelihood for a given sequence of choices averaged across the three level of the hierarchy then becomes

$$L(\theta_p) = \frac{1}{m} \frac{1}{k} \sum_{i=1}^{m} \sum_{j=1}^{k} log(P(x^p_{ji} \mid \theta_p)). \tag{11}$$

We subsequently compared the models using Bayesian model selection (Daunizeau et al., 2014; Stephan et al., 2009) to assess the fit with respect to the participant-wise data. Bayesian model selection is a powerful tool to determine the most likely model that might have generated observed data. Given the set-up of several hypothesis (function types) we decided to employ random-effects Bayesian model selection. Hereby, we treated models as potentially random-effects that can differ between participants (Stephan et al., 2009). To perform the analysis we used the Python implementation of the VBA toolbox (CPILab, 2021; Daunizeau et al., 2014).

3 Results

In our experiment, participants had to learn the association of classes of visual stimuli with hierarchically structured semantic properties. Each class of visual stimuli had to be mapped to three corresponding semantic properties. Learning in the task was facilitated by trial-wise feedback and forced repetitions of incorrect selections. In a subsequent step we ran several neural network models on an idealised version of the task to assess their fit to our human data. To understand stage-like transitions in models we fit several function classes to trial-wise data and analysed model evidence.

3.1 Exploratory analysis

In general, participants appear capable of performing the task and learned the required input-output mapping. We can see from Figure 4 that participants do perform above chance after a single block (16 trials) of training. In the below plot accuracy is calculated only as the correct selection of properties present in the hierarchical structure, but not as the correct non-selection of absent properties. Interestingly, not a single participant appears to perform numerically below chance-level at the end of training.

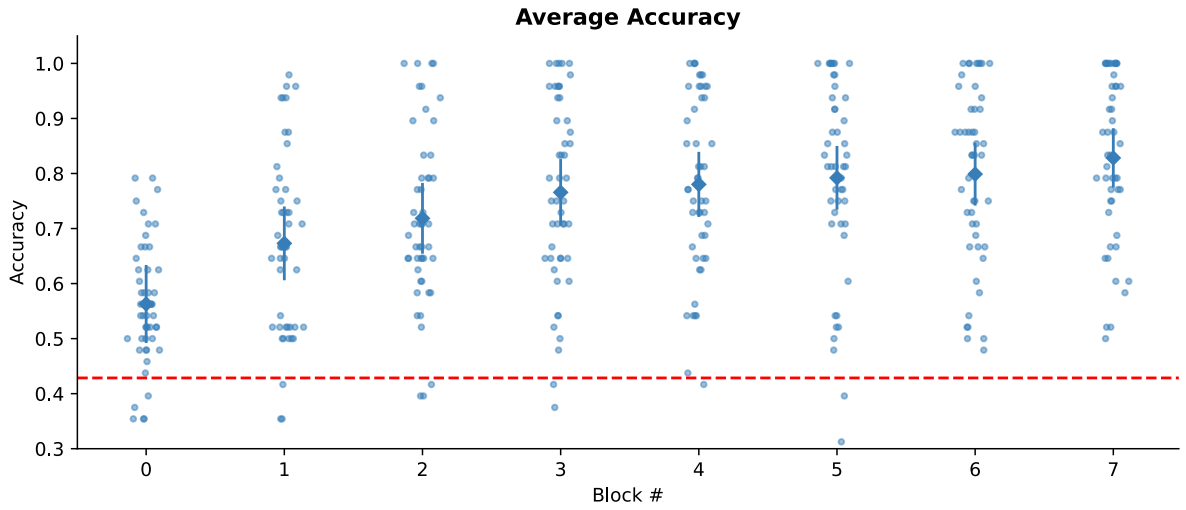


Figure 4: Accuracy of participants across all 3 property levels. Error bars indicate SEM. Light dots indicate individual participants performance. The red dotted line indicates chance performance.

We can also observe qualitatively that participant learning appears to respect the hierarchical structure which underlies the presented semantic properties. For example, we find that participants learn the top level in the hierarchy at a faster pace and can perform with nearly perfect accuracy in later blocks (Figure 5, orange line).

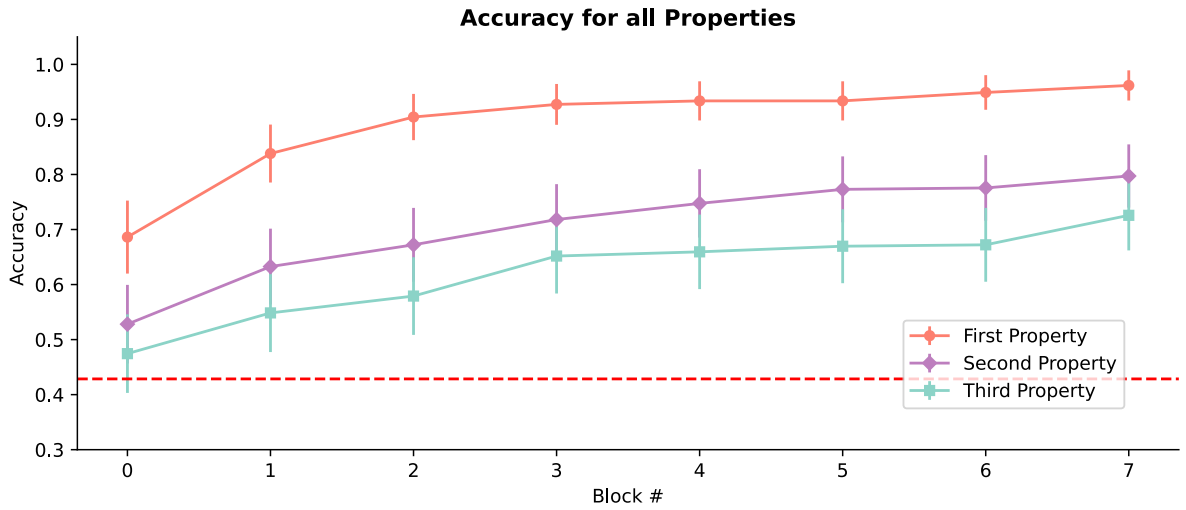


Figure 5: Accuracy of participants across all 3 property levels plotted as a function of blocks. Error bars indicate SEM. The red dotted line indicates chance performance.

This pattern holds for all three levels of the property hierarchy. Participants perform best for properties at the highest level while accuracy is lowest for properties at the lowest level in the hierarchy throughout training (Figure 5). However, participants are capable of learning lower-level distinctions as performance is above chance for all three levels. The difference in performance between the lowest and the mid-level of our hierarchy is especially remarkable as the lowest properties are most specific to individual classes of planets and are therefore a better identifier of respective categories.

3.2 Inferential test for hierarchical differentiation

To assess if participant learning follows a pattern of hierarchical differentiation, we performed a two-way repeated measures ANOVA. If participant learning indeed reflects the statistical regularities of the hierarchical structure we would see an effect of hierarchical level and an effect of block, as participants learn over time. We assessed this descriptive

pattern using a two-way repeated measures ANOVA to evaluate the effect of experimental block and hierarchical level on participants accuracy scores. The two-way repeated measures ANOVA revealed significant main effects of block $F(7, 336) = 41.635$, $p < .0001$, $\eta^2 = 0.151$ and level $F(2, 96) = 98.87$, $p =< .0001$, $\eta^2 = 0.26$ on participants accuracy in the semantic learning task. However, a term coding for the interaction of block and level was not significant $F(14, 672) = 0.251$, $p = 6.74e - 02$, $\eta^2 = 0.005$. Mauchly's test indicated that the assumption of sphericity had been violated for block $W = 0.14$, $p < .5$ and level $W = 0.55$, $p < .5$. Significance values for the main effect are therefore reported with Greenhouse-Geisser correction. The results indicate that participants' performance indeed respects the levels of the semantic hierarchy.

3.3 Assessing human input-output correlation matrices

To understand the patterns of progressive hierarchical differentiation and stage-like transitions, we assess the participants' time dependent input-output correlation matrices $\hat{\Sigma}^{yx}(t)$ in an analogous way to Saxe et al. (2019). Qualitatively, it appears that the highest level of the hierarchy is learned first. However, the dynamics of the two lower levels are not entirely clear from Figure 6. This patterns also persists when examining trial-wise input-output matrices. The highest level of the semantic hierarchy appears to be learned first followed by the acquisition of properties at lower points in the semantic hierarchy (Figure 7).

In accordance with the analysis by Saxe et al. (2019), we can decompose this input-output correlation matrix using a singular value decomposition (SVD) which generalises the spectral decomposition of symmetric matrices to the non-symmetric case. The singular value decomposition of our input-output matrix $\hat{\Sigma}^{yx}$ will decompose the given matrix of rank r into left and right orthogonal singular vectors and a diagonal matrix of singular values (the square root of eigenvalues of $(\hat{\Sigma}^{yx})^T \hat{\Sigma}^{yx}$) (Lay et al., 2016). The result of this decomposition in the final block of training is displayed in Figure 8. Note that left and right singular values were scaled for better comparability with Saxe et al. (2019). This,

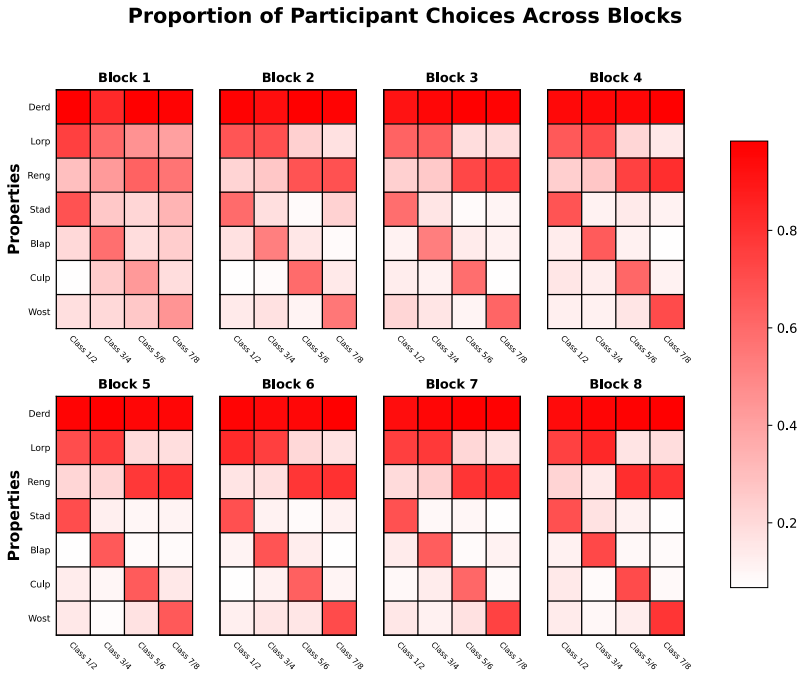


Figure 6: The input-output correlation matrix $\hat{\Sigma}^{yx}(t)$ as a function of blocks. Property and class labels are used for illustration purposes only as these are randomised across participants.

however, leaves the decomposition unchanged as scaling eigenvectors by a scalar does not change their corresponding singular values or eigenvalues.

The decomposition of the input-output matrix appears similar to the deep linear networks by Saxe et al. (2019) at convergence. Another important observation is that left, and right singular vectors are not constant throughout learning as in Saxe et al. (2019). Figure 10 illustrates the decomposition after 1 block of learning. Note how matrices U and V^T are different to those seen in Figure 9. A further important difference between the observed input-output correlation matrices and the ones observed stems from our experimental design. Participants are forced to make choices from their first trial onward. Hence, despite no initial knowledge our input-output correlation matrix does not start as a null matrix $\mathbf{0} \in \mathbb{R}^{8\times4}$, which is reflected in the structure of our singular values from the earliest stages of learning (Figure 9.).

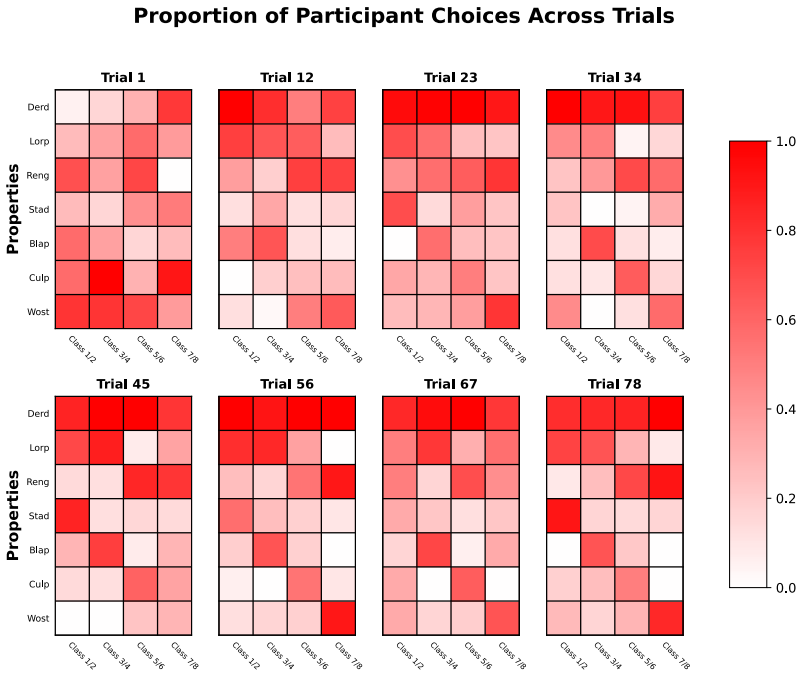


Figure 7: The input-output correlation matrix $\hat{\Sigma}^{yx}(t)$ as a function of trials. Property and class labels are used for illustration purposes only as these are randomised across participants.

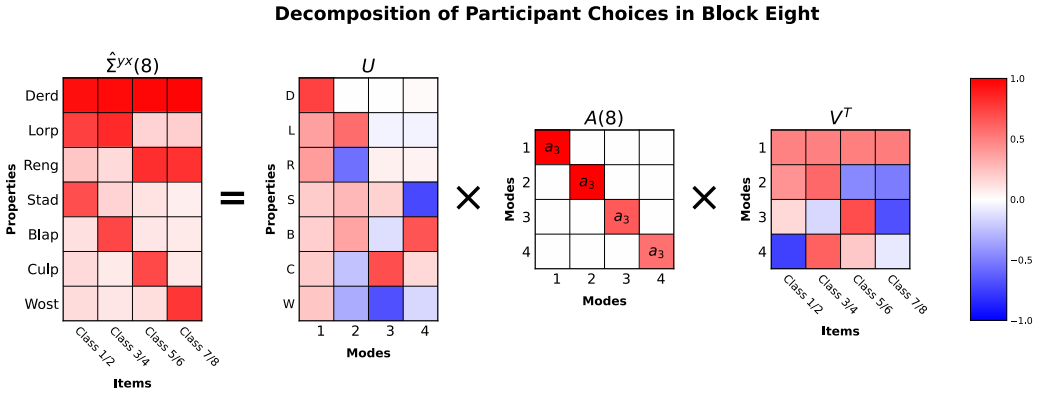


Figure 8: Singular value decomposition of our participant choices after learning. The magnitude of diagonal singular values recapitulates the structure learned. Larger singular values in the top left of $A(8)$ reflect the stronger performance on higher level properties.

3.4 Comparing human and network input-output correlation matrices

To further understand the pattern of progressive differentiation and stage-like transitions we will compare our obtained human input-output correlation matrices with neural

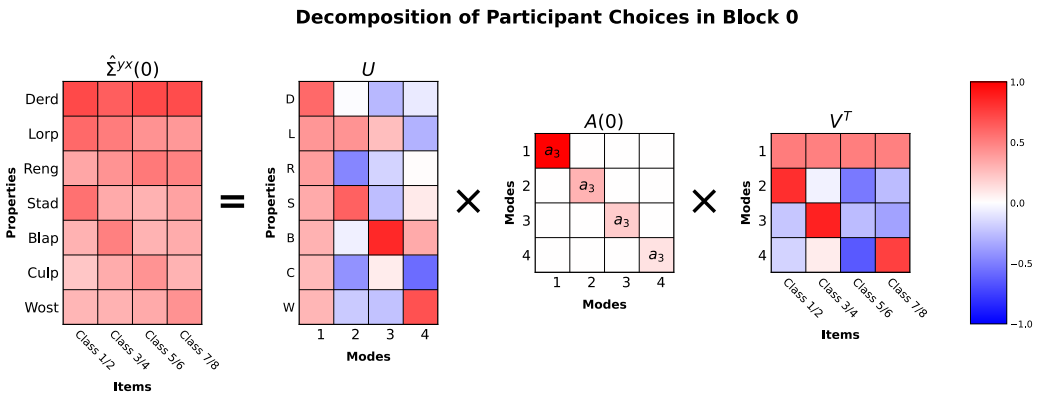


Figure 9: Singular value decomposition of our participant choices in the initial stages of learning. Singular values recapitulate the structure learned. The large singular value a_1 on the Diagonal of $A(0)$ reflects the stronger performance on the highest level of the hierarchy.

networks of different initialisation schemes and architectures. In particular, patterns of progressive differentiation and stage-like transitions are seen as a results of training networks from small initial weights. We calculated epoch-wise averages across the 49 runs of our networks in all initialisation regimes. Subsequently, we averaged over input-output matrices in 8 equal partitions to gain input-output correlation matrices comparable to those obtained from human participants across 8 blocks of learning. Given the different number of epochs in networks with and without sigmoid activation functions we averaged 500 and 100 epochs respectively into a single matrix.

To assess the similarity of these network input-output correlation matrices to our human data, we calculated the Euclidean distance between vectorised human and network matrices as

$$D(\hat{\Sigma}^{\mathbf{yx}}_{hum}(t), \hat{\Sigma}^{\mathbf{yx}}_{net}(t)) = \parallel vec(\hat{\Sigma}^{\mathbf{yx}}_{hum}(t)) - vec(\hat{\Sigma}^{\mathbf{yx}}_{net}(t)) \parallel \tag{12}$$

here $\hat{\Sigma}^{\mathbf{yx}}_{hum}(t)$ represents the average human input-output correlation matrix at block t and $\hat{\Sigma}^{\mathbf{yx}}_{net}(t)$ represents the average network input-output correlation matrix at partition t. This procedure was carried out for all initialisation schemes and networks to quantify potential correspondences. First, for networks without sigmoid activation functions, Euclidean distance was smallest for networks trained from small initial weights. We can see that the Euclidean distance for pure linear networks and for the network with ReLU activation

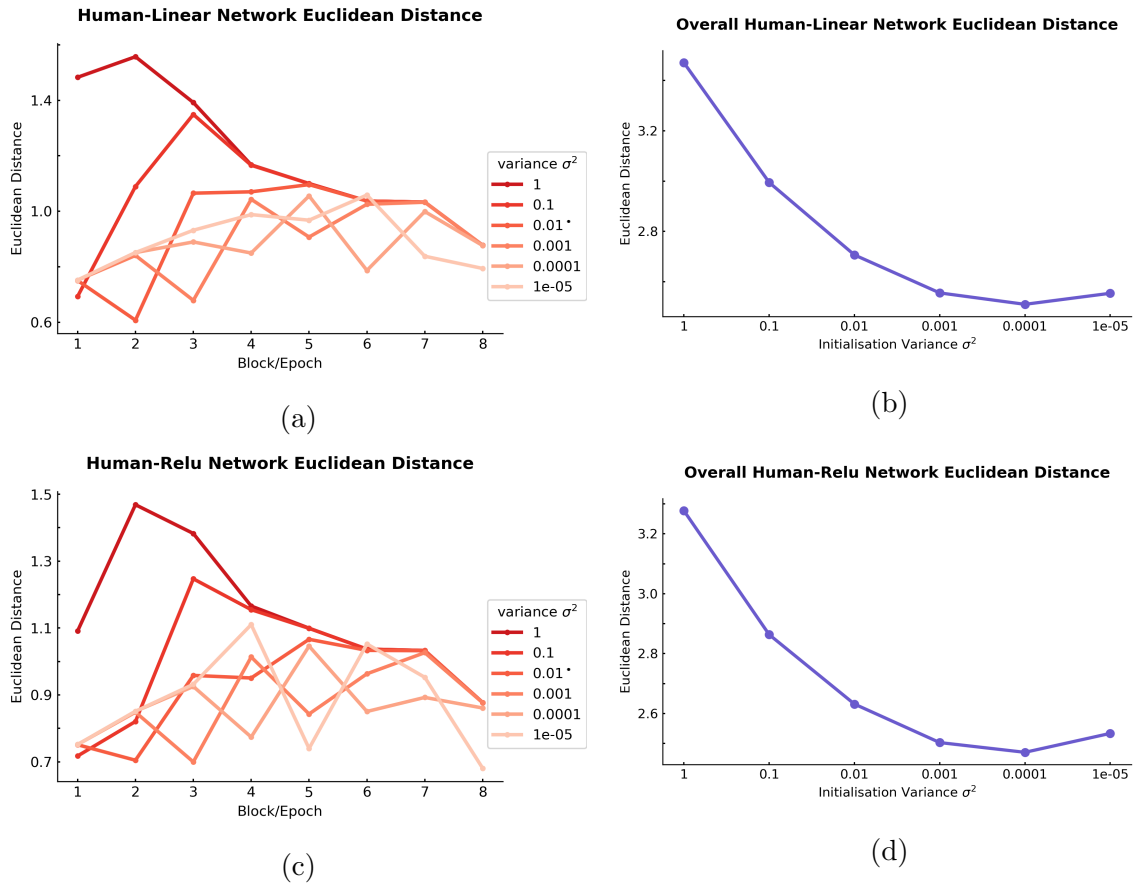


Figure 10: Euclidean distance between vectorised input-output-correlation matrices for humans and neural networks with linear outputs. (a) and (c) display the Euclidean distance as calculated by equation 12 as a function of block and initialisation variance. (b) and (d) display the overall Euclidean distance between humans and neural network matrices as calculated by equation 13 as a function of initialisation variance.

function in the hidden layer tends to be smaller across blocks as seen in Figure 10a and 10c. For large initialisations Euclidean distance is larger. This pattern becomes even more evident when calculating the general Euclidean distance of all blocks for each initialisation as

$$D_{all}(\hat{\Sigma}^{\mathbf{yx}}_{hum}(1),...,\hat{\Sigma}^{\mathbf{yx}}_{hum}(8),\hat{\Sigma}^{\mathbf{yx}}_{net}(1),...,\hat{\Sigma}^{\mathbf{yx}}_{net}(8)) = \\ \|vec(\left[\hat{\Sigma}^{\mathbf{yx}}_{hum}(1),...,\hat{\Sigma}^{\mathbf{yx}}_{hum}(8)\right]) - vec(\left[\hat{\Sigma}^{\mathbf{yx}}_{net}(1),...,\hat{\Sigma}^{\mathbf{yx}}_{net}(8)\right])\| \quad (13)$$

We can see from Figure 10b and 10d that the Euclidean distances between networks and humans are smallest when training networks from small initial weights. Given pat-

terns of progressive differentiation and stage-like transitions in networks trained from small initial weights, the result supports our prediction that human learning in response to hierarchically structured semantic data displays progressive differentiation and stage-like transitions. For illustration we plot the input-output correlation matrices for the linear model with best correspondence to our human data in Figure 11.

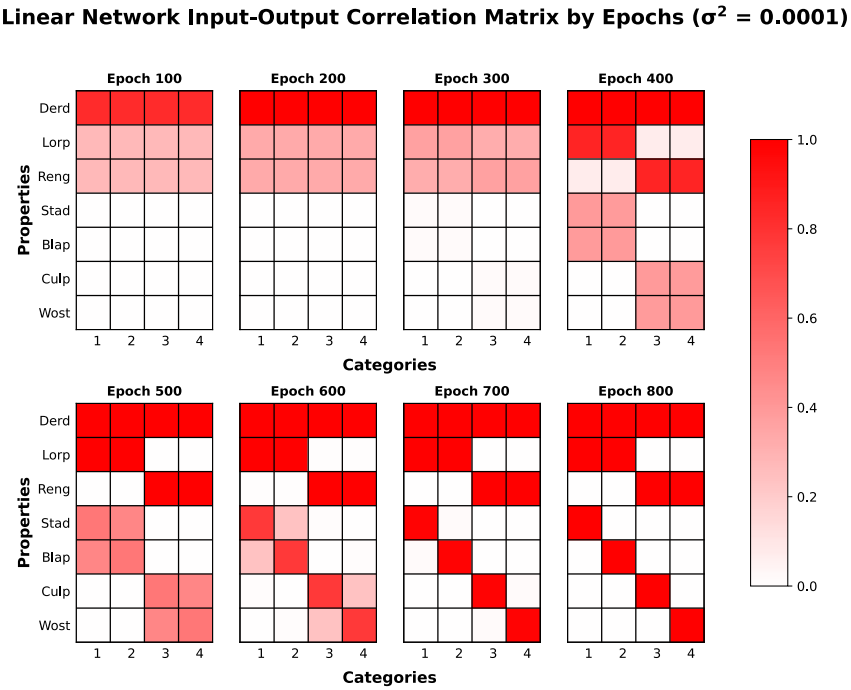


Figure 11: The average input-output correlation matrix $\hat{\Sigma}^{yx}(t)$ of the linear networks with weights in the first and second layer initialised as $\mathcal{N}(0, 0.0001/N_1)$ and $\mathcal{N}(0, 0.0001/N_3)$. Each matrix is the average 49 different runs of the Networks averaged across 100 epochs. Labels for illustration purposes as network inputs and outputs are idealised vectors.

Results are comparable when running equivalent networks with sigmoid activation functions in the output layer. The relationship between weight initialisation and fit to human data appears generally similar. That is, larger initial weights correspond to larger Euclidean distances of human and neural network input-output matrices. Figure 12a and 12c show that this pattern holds across blocks. We can see that the closest distance between human and neural network models is obtained for initial weights drawn from $\mathcal{N}(0, 0.001/N_1)$ and $\mathcal{N}(0, 0.001/N_3)$ for the first and second layer respectively. This result is slightly different from those obtained in network with linear output units. The smallest

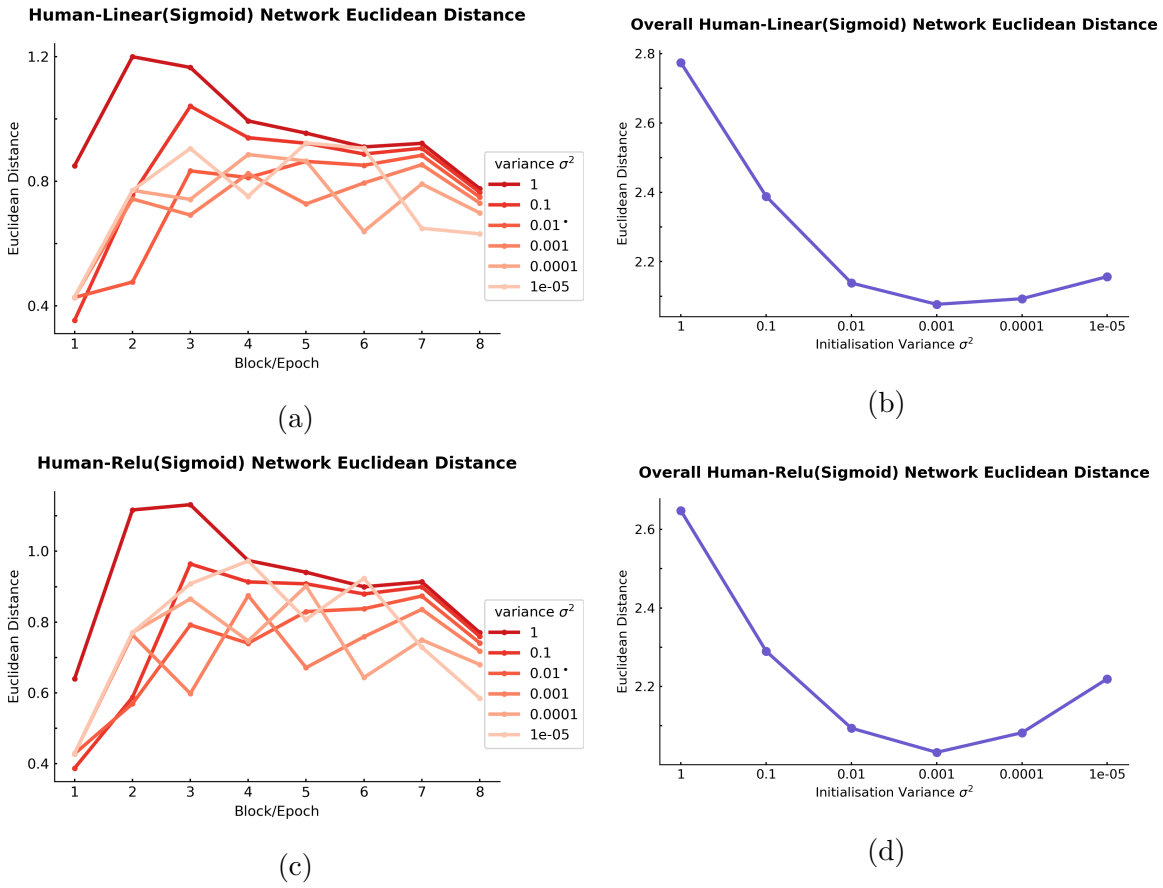


Figure 12: Euclidean distance between vectorised input-output-correlation matrices for humans and neural networks with sigmoid output units. (a) and (c) display the Euclidean distance as calculated by equation 12 as a function of block and initialisation variance. (b) and (d) display the overall Euclidean distance between humans and neural network matrices as calculated by equation 13 as a function of initialisation variance.

Euclidean distances are attained when initial weights are drawn from a normal distribution with variance one order of magnitude larger. We show the input-output matrices for the best corresponding network with linear units in the hidden layer and sigmoid activation functions in the output layer in Figure 13.

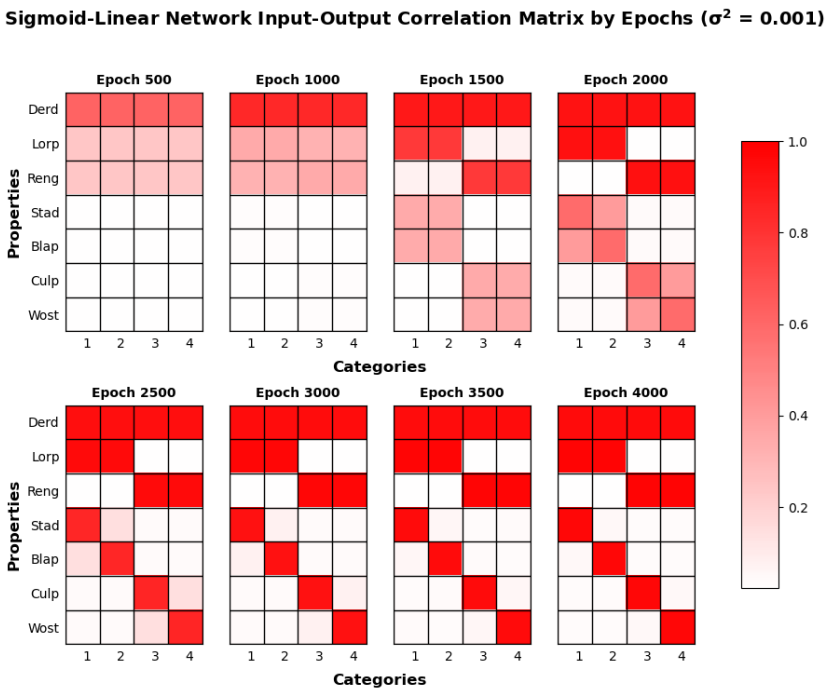


Figure 13: The average input-output correlation matrix $\hat{\Sigma}^{yx}(t)$ of a linear network with sigmoid activation functions with weights in the first and second layer initialised as $\mathcal{N}(0, .001/N_1)$ and $\mathcal{N}(0, .001/N_3)$. Each matrix is the average 49 different runs of the Networks averaged across 100 epochs. Labels for illustration purposes as network inputs and outputs are idealised vectors.

3.5 Participant-wise function fitting

To quantify learning trajectories beyond group aggregates, we fit functions to each participant's trial-wise accuracy scores for all three levels in the hierarchy. Firstly, given the average log-likelihood in Equation 11 we compared these scores for our four candidate functions (see Figure 14).

To examine if participant-wise log-likelihoods differed between function types we compared models using Bayesian model selection (Daunizeau et al., 2014; Stephan et al., 2009). This procedure enabled us to compute the Exceedance Probabilities for our 4 hypothesis functions (see Figure 15). Exceedance Probabilities here represent the probability that a given model generated the observed data. The results indicates strongest support for the linear model.

Furthermore, we computed the estimated model posterior frequencies. These quanti-

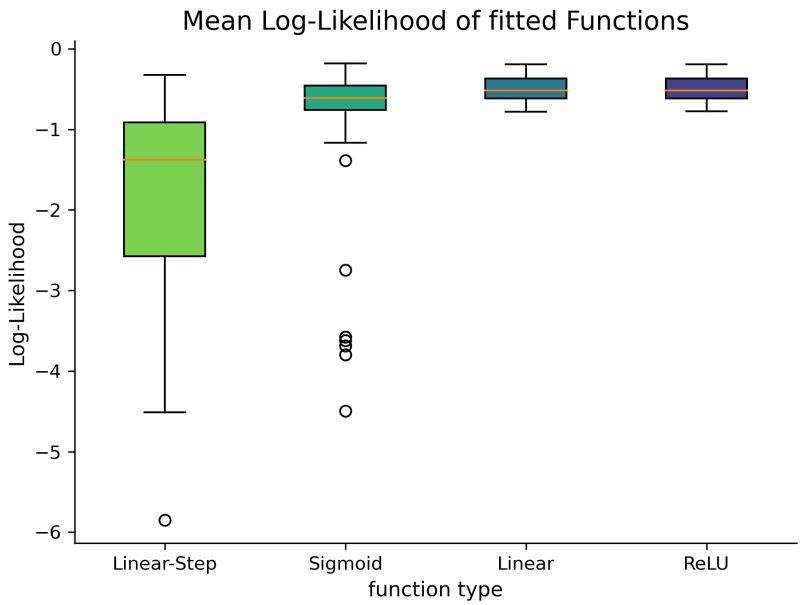


Figure 14: Log-likelihood of each function for the fitted function for all $n = 49$ participant for each. Boxes represent the interquartile range. Whiskers are positioned at 1.5 times interquartile range. Yellow lines indicate Median.

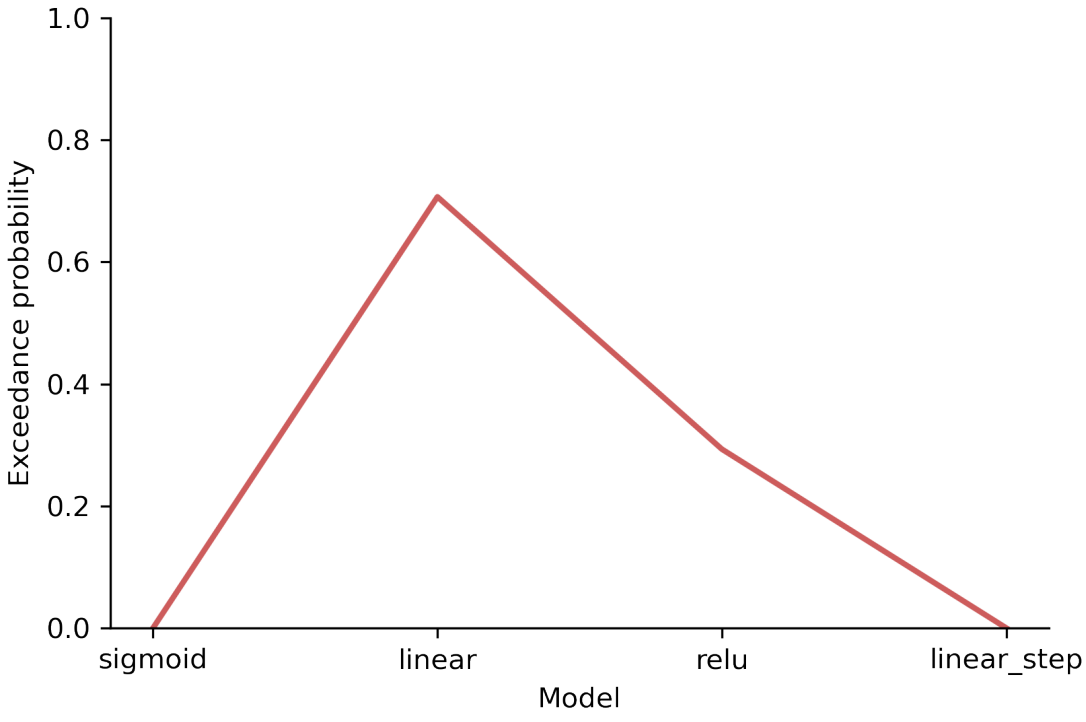


Figure 15: Exceedance probabilities of our four functions.

ties represent the probability that a particular model prevails for a given random participant (see Figure 16). The results recapitulate those of our Exceedance Probabilities. For a given random participant the linear model is most likely to prevail. The fact that the linear model appears as the most plausible hypothesis potentially indicates that individual participants learning trajectories follow a more gradual, rather than stage-like, pattern.

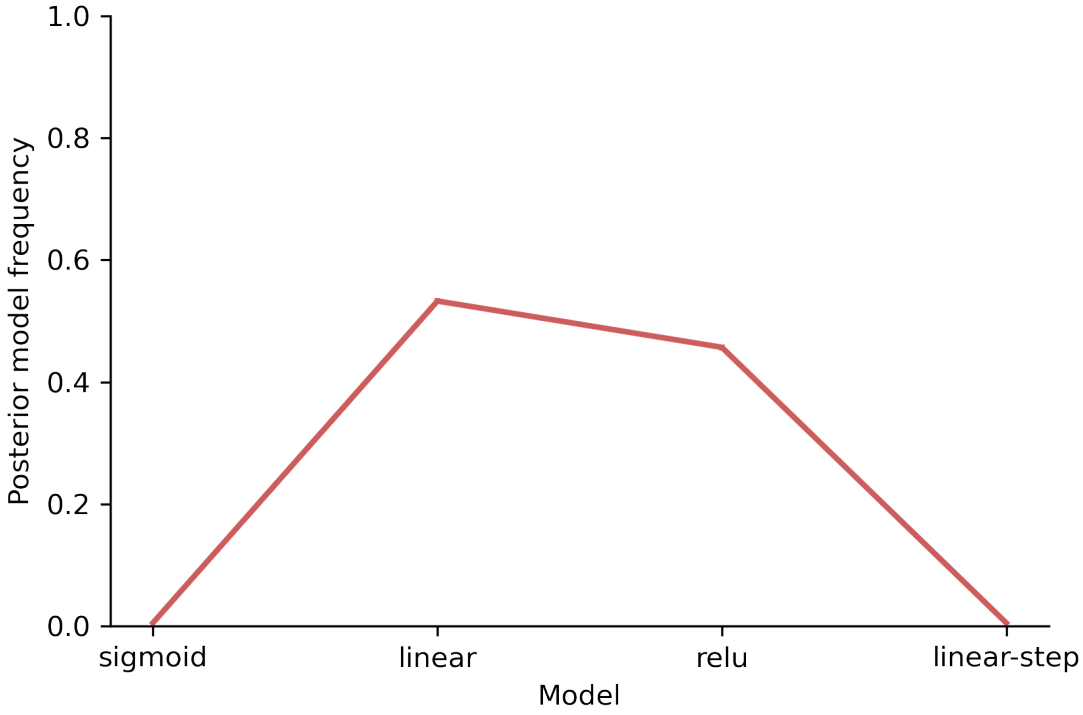
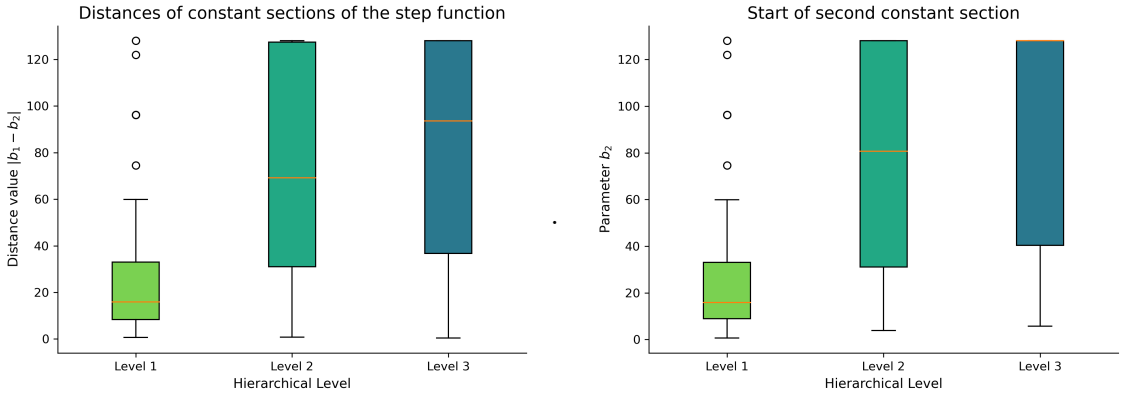


Figure 16: Estimated model frequency of our four functions for a given participant.

Despite this result, we proceeded to analyse the participant-wise parameters in the linear step function. That is, we tried to understand learning patterns in a type of function that *a priori* assumes stage-like learning. For our linear step function, the slope of the function's non-constant section is determined by the difference in offsets $|b_1 - b_2|$. We assessed stage-like transitions via the steepness of this slope across levels of the semantic hierarchy. A visualisation of this can be found in Figure 17a. A smaller distance between the parameters in this context indicates faster acquisition of knowledge. That is, a more stage-like transition. Given the non-normality of participant wise $|b_1 - b_2|$ we conducted a Friedman test to assess if steepness of learning trajectories differs between levels of



(a) (b)

Figure 17: Parameter comparisons between levels. (a) Distribution of participant-wise slopes of the linear step-function $|b_1 - b_2|$, smaller difference indicates larger slope and (b) Distribution of parameters b_2, which indicates the attainment of peak performance. Boxes represent the interquartile range. Whiskers are positioned at 1.5 times interquartile range. Lines indicate Median.

semantic knowledge. The results revealed that the differences in offsets $|b_1 - b_2|$ were significantly different between the three levels of the hierarchy $X_F^2 = 51.05, p < .0001$. Post-hoc analyses with the Bonferroni-corrected Wilcoxon signed rank test revealed that slopes of learning curves where steeper on the first level in the hierarchy ($median = 15.89$) than the on the second ($median = 69.19$) and third level ($median = 93.66$) both $p < .0001$. In addition, learning curves were steeper on the second than on the third level $p < 0.05$.

Secondly, we also assessed the differences in distribution of parameters b_2 between different levels as seen in Figure 17b. The parameter indicates the start of the second linear section of the graph and therefore can serve as an indicator for order of acquisition. A Friedman test revealed that the parameter b_2 differed significantly between the three levels of the hierarchy $X_F^2 = 67.06, p < .0001$. Post-hoc analyses revealed that using a Bonferroni-corrected Wilcoxon signed rank test indicated that participant-wise parameters b_2 for the first level ($median = 15.90$) were significantly lower than parameters on the second ($median = 80.72$) and third level ($median = 128.0$) both $p < .0001$). Furthermore, the difference between parameters between the second and third level was significant ($p < .001$). This indicates that lower-level properties in the semantic hierarchy

were generally acquired later in learning.

4 Discussion

4.1 Key contributions

The current work makes several key contributions to the experimental and theoretical literature on semantic cognition in humans and deep neural networks. We generated predictions for the structure and trajectory of semantic learning from empirical (Rogers & McClelland, 2004) and analytical (Saxe et al., 2019) results obtained in deep neural networks.

Our first key contribution is concerned with the hierarchical differentiation of semantic knowledge. In the context of a semantic learning task with hierarchically structured properties analogous to model simulations by Saxe et al. (2019) and predictability trees used in developmental studies by Keil (2013), we found results that at first glance resemble those obtained in deep linear networks and humans. Human performance throughout learning appears to be sensitive to the hierarchical structure of semantic knowledge. Hereby participants learned semantic properties at higher level nodes in the hierarchy with more accuracy than properties located at lower points in the structure. We show that similar to results in deep neural networks (Saxe et al., 2019) these patterns can be quantified in input-output correlation matrices and their decomposition. In decomposed matrices, the strength of singular values quantify the acquisition of semantic knowledge. These results extend previous findings in semantic learning that often focused on changes over much longer developmental time scales (Carey, 1985; Keil, 2013; Mandler & McDonough, 1993). Our results highlight the utility of connectionist theories (Rogers & McClelland, 2004) and modern analytical theories of deep neural networks (Saxe et al., 2019) as predictive tools in semantic cognition. While learning mechanisms (such as backpropagation) in the employed models are frequently regarded as biologically implausible (Crick, 1989; Grossberg, 1987; Whittington & Bogacz, 2019), it appears that there is merit to the precise predictions made by these theories. Despite model parsimony and alleged biological implausibility, deep neural networks and human participants may both exploit regularities

in the learning task in a progressive, hierarchical fashion.

Our second contribution is concerned with the comparison of human and network input-output correlation matrices under different architectures and initialisation. By calculating the Euclidean distance of average network and human input-output correlation matrices we find intriguing patterns of similarity. For neural networks with linear output activation functions, a gradient of similarity describes their relationship to human data. With larger initial weights, similarity to human data decreases. This result is interesting as it provides quantitative support for the qualitative similarity of deep linear network learning trajectories to those of human learners noted by Saxe et al. (2019). Hallmarks of progressive differentiation and stage-like transitions only appear in deep linear networks trained from small norm weights and disappear at larger weight initialisations (Saxe et al., 2019; Saxe et al., 2021). The larger similarity may translate to learning patterns more resembling those seen in deep linear networks. These results not only strengthen the case for neural networks as predictive models of human semantic learning but also relate to recent work noting larger similarity of behaviour and neural representations between humans and neural networks trained from small norm initial weights (Flesch et al., 2021). Despite not focusing on neural data we find larger correspondence between humans and neural networks when training from smaller initial weights.

Our final contribution is the analysis of individual participant's data via the fitting of functions to trial-wise performance across the three levels of the semantic hierarchy. We find that contrary to our expectations, functions which have an inherently more stage-like form such as sigmoidal or linear-step functions fit participant-wise data less closely for all three levels of the semantic hierarchy. We interpret this results as contrasting with the idea of stage-like transitions reported previously in neural networks (Rogers & McClelland, 2004; Saxe et al., 2019) and humans (Inhelder & Piaget, 1958; Keil, 2013; Siegler, 1976). However, we suspect that these results could be an artefact of our particular technique used to fit and evaluate functions. We fit functions to individual par-

ticipant data using least squares optimisation, however, we evaluated performance using log-likelihoods. While theoretically we might expect models to converge on similar solutions, log-likelihoods will strongly penalise large differences between human and model performance. When fitting functions that asymptote at or close to 0 and 1, such as sigmoid or linear step functions, inaccurate predictions are penalised more strongly than is the case for functions that are mostly linear in their domain. Despite these limitations the analysis of model parameters allows for insights into patterns of stage-like transitions and progressive differentiation. When comparing the steepness of the slope in our linear step-function we find that properties at higher levels in the semantic hierarchy appear to be learned at a faster pace than those at lower levels in the hierarchy. This observation recapitulates findings by Saxe et al. (2019) and Cao et al. (2020) in neural networks trained with gradient descent. The derivative of mean squared error displays larger norm peaks during the learning of higher level semantic properties – indicating a faster acquisition (Cao et al., 2020). Furthermore, the start of the second linear interval of the step function b_2 displays patterns that recapitulate progressive learning. The highest level of performance is attained earlier in training for higher level properties. The result on the analysis of fitted function parameters appear to confirm progressive differentiation and stage-like transitions in the human learning of semantic properties.

We presented three strands of evidence from analysing human behavioural data on a population level, comparing humans and deep neural networks, and the modeling of individual participants' learning trajectories. Principally our results align with predictions derived from deep linear (Saxe et al., 2019) and connectionist models (Rogers & McClelland, 2004). While linear functions appeared to approximate individual learning trajectories more closely most of our results recapitulate the progressive and stage-like development of hierarchical semantic taxonomies in humans (Carey, 1985; Keil, 2013) and neural networks (Rogers & McClelland, 2004; Saxe et al., 2019). Even specific signatures such as steepness of slopes during stage-like transitions (Cao et al., 2020) appear retained when analysing participant wise data. More broadly, we find our results to be in line

with observations on *rich* and *lazy* learning in deep neural networks (Flesch et al., 2021) whereby we find closer resemblance of network performance to human data at smaller initialisation. Against the backdrop of a larger literature comparing highly complex deep learning architectures against human behaviour and representations (Khaligh-Razavi & Kriegeskorte, 2014; Lake et al., 2017), we show that comparably simple and in some cases mathematically tractable models are capable of sensible predictions with respect to human semantic cognition.

4.2 Limitations

Several limitations apply to the current work. While we attempted to set-up our human experiment in a fashion that closely resembles modeling experiments by Saxe et al. (2019), it can be difficult to design tasks in a fully analogous fashion. For instance, inputs in deep linear networks are fully orthogonal. While we attempted to make classes of stimuli visually distinct, we can only speculate if representations of experimental stimuli after higher-level visual processing are indeed represented in an orthogonal fashion. Furthermore, our way of forcing participant responses to facilitate learning is clearly not analogous to neural networks. The range of possible outputs in deep linear networks spans all real numbers while human responses are constrained to $\{0, 1\}$ on any given property. We compare neural networks trained from *tabula rasa* with human participants with a rich reservoir of prior experience and previous knowledge (Lake et al., 2017). Therefore, networks require substantially more training instances to adequately learn the semantic task. The network architectures we employed for predictions and evaluation are limited. While model parsimony enables clear-cut predictions the dynamics of semantic cognition are to our knowledge still unexplored in large models. This constraint limits our claims of analogous learning to a relatively small set of neural network architectures. Finally, we believe that our procedure for fitting functions to individual participant data could be improved. A future study should fit participant-wise functions using the same metric employed for evaluation.

4.3 Future directions

The are a range of possible research directions which follow from our work. One option would be to assess semantic learning in humans over longer time scales. As mentioned previously, the acquisition of semantic knowledge is often considered in the context of long-term human development (Carey, 1985; Keil, 2013; Mandler & McDonough, 1993). Such a task with a more complex semantic hierarchy could allow for a more detailed analysis of learning patterns in semantic cognition. In a prolonged learning task, we may also reasonably expect changes in neural representation storing learned semantic properties. Neuroimaging of temporal and prefrontal regions involved in semantic cognition (Martin & Chao, 2001) could then be used to compare emerging representations as seen in deep linear networks (Cao et al., 2020; Flesch et al., 2021; Saxe et al., 2019) to human patterns of activation. Representational similarity analysis may constitute a promising methodological avenue for this type of analysis as the approach is frequently used to compare representations between humans and neural networks (Khaligh-Razavi & Kriegeskorte, 2014; Kriegeskorte & Kievit, 2013).

A further promising avenue is to assess illusory correlations (Carey, 1985) when making inferences in a comparable learning task. In such a paradigm we could assess inductive performance in response to new objects when participants are presented with only partial semantic information. The result of a task with carefully designed tests of generalisation ability might shed further light on the intricate connections between progressive differentiation and inductive failures as observed in neural networks (Saxe et al., 2019) and developing children (Carey, 1985).

Finally, more theoretical work could advance our work by providing a more principled understanding of semantic learning in more complex neural network architectures. Given the ever broadening range of applications of deep neural networks in research and technology (Goodfellow et al., 2016; Schmidhuber, 2015), a principled understanding of

semantic learning in complex model architectures might further elucidate computational principles to sharpen experimental predictions.

4.4 Conclusion

In conclusion, we have presented results from a semantic learning task that trained human participants in a fashion analogous to recent findings in deep linear networks. Our analysis implies general patterns of progressive differentiation and stage-like transitions through the comparison of human input-output matrices to those of deep neural networks. While results of participant-wise fitting of learning trajectories contradicts the idea of stage-like learning, we find an analysis of model parameters to agree with our experimental predictions. Our results provide empirical support for the utility of simple neural networks as tools for the study of semantic cognition.

References

Advani, M. S., Saxe, A. M., & Sompolinsky, H. (2020). High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, *132*, 428–446.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.

Cao, Y., Summerfield, C., & Saxe, A. (2020). Characterizing emergent representations in a space of candidate learning rules for deep networks. *Advances in Neural Information Processing Systems*, *33*, 8660–8670.

Carey, S. (1985). *Conceptual change in childhood*. MIT press.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, *8*(2), 240–247.

CPILab. (2021). Group-bayesian-model-comparison. *Online: https://github.com/cpilab/group-bayesian-model-comparison*.

Crick, F. (1989). The recent excitement about neural networks. *Nature*, *337*(6203), 129–132.

Daunizeau, J., Adam, V., & Rigoux, L. (2014). Vba: A probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS computational biology*, *10*(1), e1003441.

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A. M., & Summerfield, C. (2021). Rich and lazy learning of task representations in brains and neural networks. *BioRxiv*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Mit Press.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive science*, *11*(1), 23–63.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245–258.

Horst, J. S., & Hout, M. C. (2016). The novel object and unusual name (noun) database: A collection of novel images for use in experimental research. *Behavior research methods*, *48*(4), 1393–1409.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures* (Vol. 22). Psychology Press.

Keil, F. C. (2013). Semantic and conceptual development. *Semantic and conceptual development*. Harvard University Press.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, *10*(11), e1003915.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in cognitive sciences*, *17*(8), 401–412.

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

Lay, D. C., Lay, S. R., & McDonald, J. J. (2016). *Linear algebra and its applications*. Pearson.

Mandler, J. M. (2000). Perceptual and conceptual processes in infancy. *Journal of cognition and development*, *1*(1), 3–36.

Mandler, J. M., & McDonough, L. (1993). Concept formation in infancy. *Cognitive development*, *8*(3), 291–318.

Martin, A., & Chao, L. L. (2001). Semantic memory and the brain: Structure and processes. *Current opinion in neurobiology*, *11*(2), 194–201.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological review*, *92*(3), 289.

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

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of experimental psychology: General*, *104*(3), 192.

Rosch, E. (1978). Principles of categorization.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation* (tech. rep.). California Univ San Diego La Jolla Inst for Cognitive Science.

Rumelhart, D. E., Todd, P. M. et al. (1993). Learning and connectionist representations. *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience*, *2*, 3–30.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537–11546.

Saxe, A. M., Nelli, S., & Summerfield, C. (2021). If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, *22*(1), 55–67.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85–117.

Sheahan, H. (2021). Hcategorylearn. *Online: https://github.com/hannahsheahan/HCategoryLearn*.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive psychology*, *8*(4), 481–520.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental science*, *10*(1), 89–96.

Squire, L. R. (1986). Mechanisms of memory. *Science*, *232*(4758), 1612–1619.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage*, *46*(4), 1004–1017.

Tulving, E. (1984). Precis of elements of episodic memory. *Behavioral and Brain Sciences*, *7*(2), 223–238.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R.,

Larson, E., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Voglis, C., & Lagaris, I. (2004). A rectangular trust region dogleg approach for unconstrained and bound constrained nonlinear optimization. *WSEAS International Conference on Applied Mathematics*, *7*.

Whittington, J. C., & Bogacz, R. (2019). Theories of error back-propagation in the brain. *Trends in cognitive sciences*, *23*(3), 235–250.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., & Srebro, N. (2020). Kernel and rich regimes in overparametrized models. *Conference on Learning Theory*, 3635–3673.

Zhu, C., Byrd, R. H., Lu, P., & Nocedal, J. (1997). Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on mathematical software (TOMS)*, *23*(4), 550–560.

Acknowledgements

I have many people to thank for the completion of this thesis who provided invaluable support to me in many ways. This thesis would never have been possible without my supervisor Professor Christopher Summerfield and his invaluable advice on all aspects of the project. I am especially grateful for Chris's sheer, endless stream of ideas and suggestions that improved the project way beyond its initial form. I am also deeply indebted to my secondary supervisor, Dr. Andrew Saxe, for his thorough advise on all theoretical aspects of this project and for allowing me to take a glimpse into the world of theoretical neuroscience and machine learning.

I am also deeply grateful to the ESRC Grand Union DTP for providing me with funding during this MSc Project and for my upcoming DPhil. I would never have been able to take up my studies without this support.

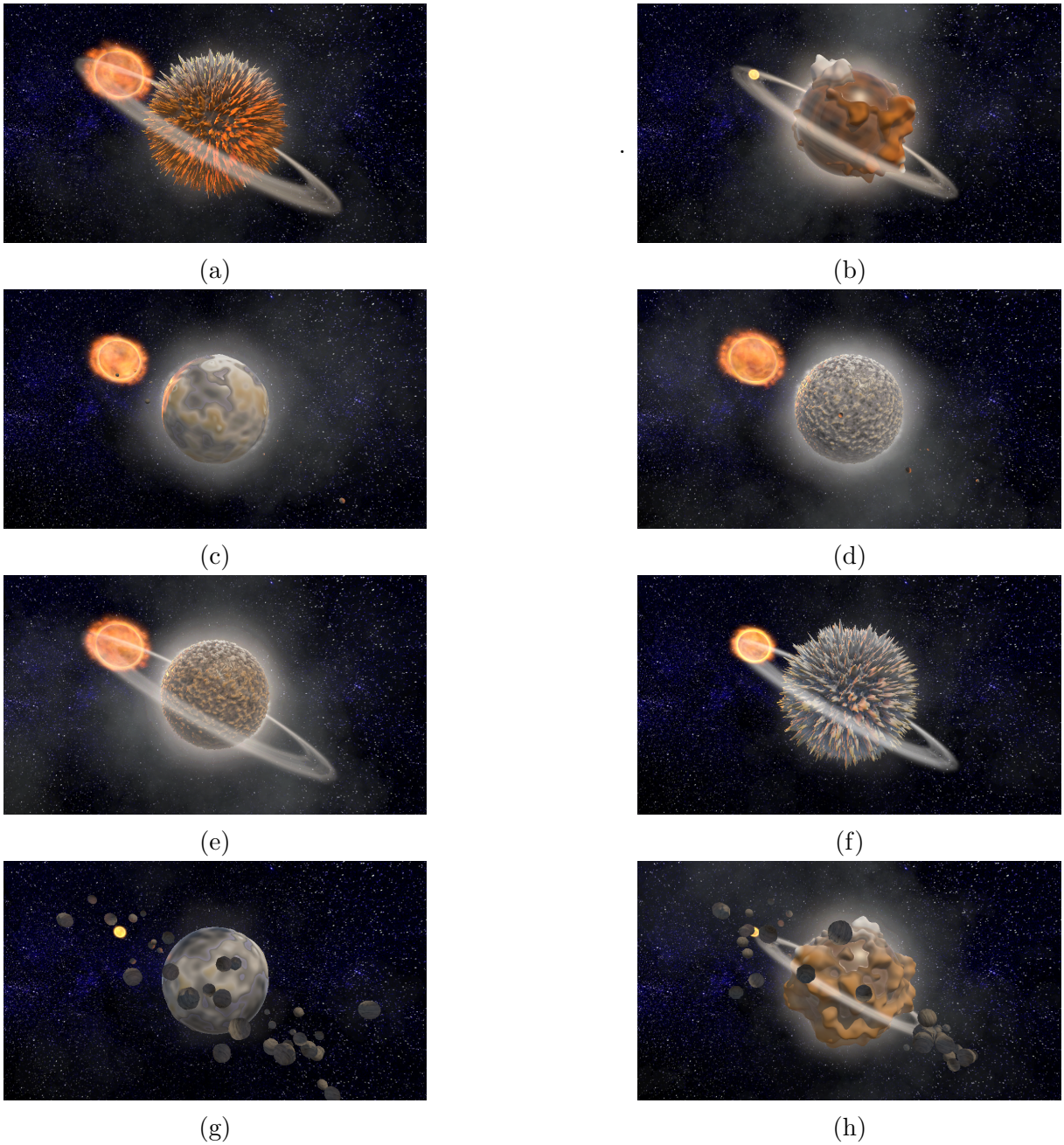
Beyond this, I am extremely grateful to all members of the Human Information Processing Lab and the UCL Theory of Learning Lab. There are too many people to list here but each and every one of you has provided me with either encouragement, advice, targeted skills training, or interesting intellectual challenges. This has allowed me to grow and learn immensely in the last year.

I would also like to express special thanks to all anonymous participants who kindly volunteered their data for this study.

I am also very thankful for my parents who always support and encourage me on my endeavours even when I am far away. I am grateful for my friends Mo, Satwik, and Graham for their encouragement and support. Finally, I would like express deep gratitude to my partner, Nurah, for her love and support at all times and for the opportunity to share this journey with her.

Appendices

A Full stimuli examples



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 18: Examples of all 8 classes of used stimuli.

B Experimental instructions

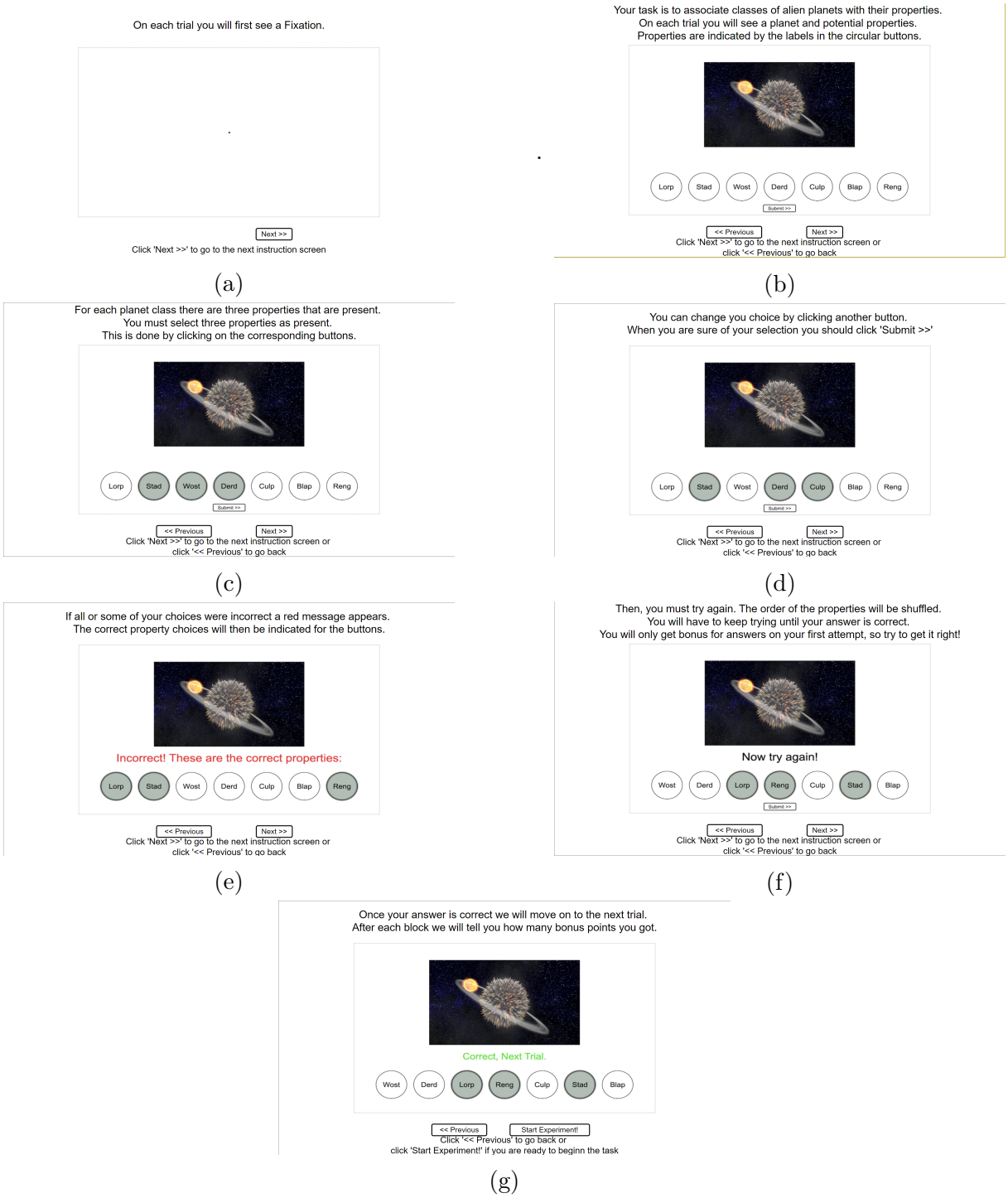


Figure 19: Experimental instructions in order of presentation.

C Network inputs and outputs for training

Similar to Saxe et al. (2019) we define our input-output correlation matrix of training examples as

$$\mathbf{\Sigma}^{yx} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Similarly network inputs are

$$\mathbf{\Sigma}^{x} = \mathbf{I} \in \mathbb{R}^4$$

D Software

Neural network simulations and trial by trial function fitting were implemented in Python 3.9.12. To run our neural network simulations PyTorch 1.2.1 and NumPy 1.22.3. Function fits were obtained with SciPy 1.7.3. For statistical tests we employed python packages Pandas 1.4.2, Pingouin 0.5.2, NumPy 1.22.3, and rpy2 3.5.2. Figures were generated with Matplotlib 3.5.1. Bayesian model selection was carried out using the Python implementation of the VBA toolbox (CPILab, 2021; Daunizeau et al., 2014).