VISUAL CATEGORY LEARNING AND CURRICULUM EFFECTS IN HUMAN AND MACHINE AGENTS

BY

Ronald B. Dekker, BSc.[*]

MASTER THESIS (36 EC) [**]

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Brain & Cognitive Sciences (track: Cognitive Science) at the University of Amsterdam

Supervisor:
Christopher Summerfield, PhD

Co-assessor & UvA representative:
Ilja G. Sligte, PhD

01 September 2016 – 28 February 2017

[*] Student number: 10006419

[**] Carried out at University of Oxford, Medical Sciences Division, Department of Experimental Psychology, Human Information Processing lab

VISUAL CATEGORY LEARNING AND CURRICULUM EFFECTS IN HUMAN AND MACHINE AGENTS

Ronald Dekker

KEYWORDS

Category learning
Curriculum learning
Artificial neural networks
Convolutional neural networks
Naturalistic stimuli

ABSTRACT

Whilst much attention has been paid to the computational mechanisms underlying category learning with simple stimuli, instructed tasks, and binary responses, few studies have explored how humans learn to categorize complex, naturalistic stimuli with minimal prior knowledge of the task rules. In particular, little is known about how the training regimes ('curricula') under which the task is performed facilitate the rate of learning. In the current study, we used curricula in which category-relevant variance and variance in extraneous dimensions were modified independently. With this design, we explored category learning at different levels, using both artificial neural network simulations and real participants. Using complex, naturalistic stimuli, we found that learning is impaired when training exclusively near the decision boundary, both in humans and in neural networks. This diametrically opposes findings for simple artificial stimuli and suggests that identification of relevant dimensions and/or hypothesis testing on dimension space is integral to category learning of real-world stimuli. In addition, we explored how relevant and irrelevant dimensions interact. Here, we observed that category-irrelevant variance impairs learning, but only when category-relevant variance is sufficiently low. We conclude with recommendations for future research, including use of similar designs to study transfer effects in category learning.

INTRODUCTION

Clustering our sensory inputs into meaningful categories allows us to make sense of an almost infinitely complex world. The categories we employ are constantly modified and refined – for example, a radiologist must learn over time to distinguish tumors from scanning artefacts, while a native speaker of Japanese will have to learn the phonetic contrast between the phonemes /r/ and /l/ to later become proficient in English.

Artificial vs. Naturalistic stimuli

Categorization is a fundamental building block of cognition, and so it is perhaps not surprising that there is a vast literature describing the neural and computational mechanisms that mediate this process. What is striking, however, is that extant work on category learning has largely limited itself to stylized, artificial stimuli, such as gratings or simple shapes. While these stimuli are tightly controlled and have led to important advances in psychology and neuroscience, experiments employing these stimuli differ fundamentally from learning in non-laboratory environments, where inputs are characterized by rich, complex and many-dimensional variation. As a result, real-world stimuli likely have different demands, such as identifying relevant dimensions among a large set of candidates (Pashler & Mozer, 2013). Another drawback of the focus on learning novel categories of artificial stimuli is that this has done little to account for the fact that real-world learning is seldom 'from scratch', and instead builds on prior experience and information. To illustrate, the native speaker of Japanese from our example, now having mastered English, would subsequently find it much easier to learn a closely related language, such as Frisian. The ability to flexibly re-use past knowledge in novel contexts is a hallmark of humans, in who this capacity is in several important respects more developed than in any other species (Penn & Holyoak, 2008). Given these differences, it is likely that extant work on learning of artificial categories has limited utility in developing optimal learning curricula for real-life applications, such as learning a new language or training to become a radiologist.

Curriculum effects

Although the examples used to learn a category can be crafted in many ways and presented in many orders (together called a "curriculum"), there has only been a limited number of studies investigating effective training methods for visual category learning (Ashby & Maddox, 2005; Pashler & Mozer, 2013). In extant work on differential sampling of training examples, several studies have reported that exclusively sampling far from the decision boundary can improve performance, especially when outcomes are probabilistic (Giguère & Love, 2013). Notably, this method performs better than random sampling for some real-world applications, such as predicting baseball game outcomes (Giguère & Love, 2013) and distinguishing tumorous from non-tumorous mammograms (Hornsby & Love, 2014). Indeed, it may be the case that "fading", i.e. introducing prototypical exemplars early in training, has a beneficial impact on performance, by alerting participants to key relevant features for discrimination. However, in other settings, training on difficult examples close to threshold yields better performance at test than training farther from the boundary (Pashler & Mozer, 2013). This latter finding matches the intuition that examples close to the boundary yield more information about the location of the threshold, and matches findings in machine learning algorithms for classification (e.g. SVMs: Vapnik, 2013). These contradictory findings may be attributable to a number of factors, including the use of one-dimensional vs. multidimensional stimuli, naturalistic vs. artificial stimuli, or use of probabilistic vs. deterministic categories. Of particular interest is the finding that fading is effective when stimuli contain additional, task-irrelevant dimensions (Pashler & Mozer, 2013). That is, for multidimensional stimuli, easy examples facilitate learning, at least early on. Furthermore, this learning of multidimensional stimuli is more effective when variation is introduced gradually, rather than in large steps (Medin & Bettger, 1994). While these findings provide important clues as to the end results of these curricula, their effects on the dynamics of learning remain unknown. Thus, it remains uncertain whether the observed differences are attributable to different rates of learning, or to more fundamental differences, such as different category representations induced by task demands. In addition, this work suggests that curriculum effects may be far more pronounced in naturalistic stimuli, which have many potentially meaningful dimensions, in contrast to commonly studied artificial stimuli.

In light of these gaps of knowledge, the goal of the current study is to better understand the training regimes that facilitate category learning of naturalistic stimuli in healthy human adults. This avenue of research is expected to have translational merit in informing optimized learning methods in human adults in real-world contexts. Furthermore, developing an understanding of the effects of extraneous dimensions on category learning sets the stage for investigating how previous knowledge can be re-used in new tasks, for instance by flipping relevant and irrelevant dimensions. In turn, this could facilitate comparisons with and future research involving machine agents in domains such as continual and transfer learning, where large differences between human and machine performance currently remain (Canini, Shashkov, & Griffiths, 2010; Torrey & Shavlik, 2009)

The current study

In the present study, we investigate curricula using interspersed test trials, to chart the temporal dynamics of learning. In addition, we use naturalistic images of algorithmically generated trees, which vary structurally only in their degrees of 'branchiness' and 'leafiness', but which from the participant´s perspective may feature a large number of candidate dimensions. Since only one of the manipulated dimensions determines the category, the other dimension is task-irrelevant. The effects of task-irrelevant dimensions have, to our knowledge, never been investigated directly, despite being ubiquitous in real-life classification and having important implications for models of flexible hypothesis selection (e.g. Khan, Mutlu, & Zhu, 2011)

We expect participants to learn fastest when initial presentations feature large variance in the relevant dimension, guiding attention to the relevant visual features. In addition, there could be differences in terminal performance, which may suggest construction of a complex decision criterion in lieu of accurately identifying the univariate decision criterion. For the task-irrelevant dimension, we predict that higher variance will attenuate performance, owing to increased distractor saliency.

In what follows, three experiments will be described. The main experiment is a behavioral study with human participants, which will be preceded by simulation experiments of two artificial neural network (ANN) architectures trained to classify the same stimuli. The aim of these simulations is twofold: (i) to discover any latent structures in the curricula and (ii) to provide a baseline performance which informs a suitable parameter space for human participants. Concretely, the first experiment simulates learning by taking the raw parameter values as inputs for a multi-layer perceptron (MLP). Thus, this experiment reflects learning differences inherent to the curricula and the network properties, but does not take effects at the stimulus feature level into account. Conversely, the second experiment employs a convolutional neural network

(CNN) architecture, which is trained and tested on real images. Finally, the third experiment will be performed on human participants. After the individual descriptions of these experiments, the paper will conclude with a general discussion in which key findings and limitations are examined and recommendations for future research are made.

GENERAL METHODS

Stimuli

Used visual stimuli were naturalistic tree images which were procedurally generated using a python-based tree generator developed by the author*. This approach affords the possibility of large-scale stimulus generation, such as the 25.000 images used in the CNN experiment in the current study.

The stimuli are experimentally controlled in two dimensions: the number of branches and the number of leaves per branch. Example stimuli are displayed below in Figure 1.

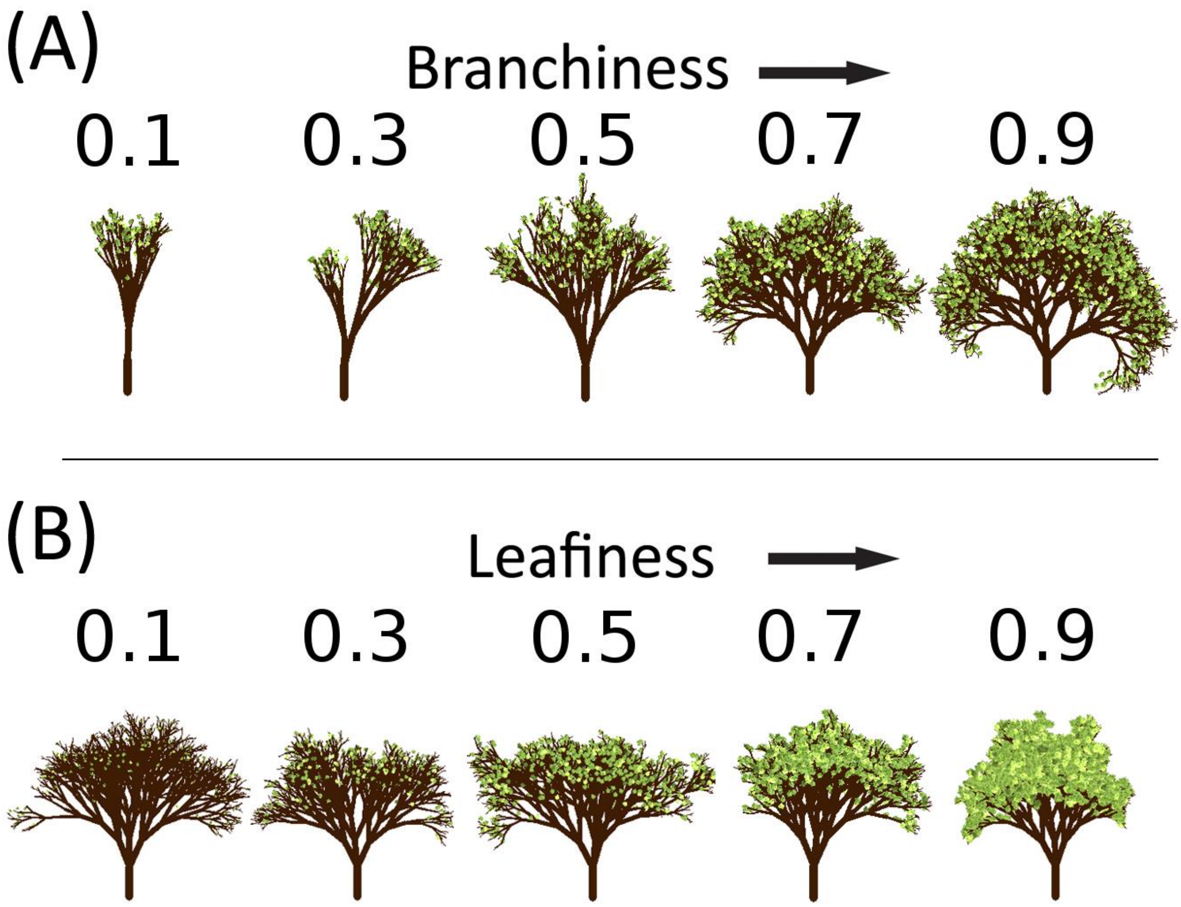

Figure 1: Example stimuli. (A): Branchiness dimension. (B): Leafiness dimension. All other factors are stochastic.

Critically, the stimuli feature several candidate dimensions, such as overall curvature and tree height, allowing us to control variability along both task-relevant and task-irrelevant dimensions. Participants perform binary category assignments on these stimuli. Although they receive no further instructions and have to rely on feedback to learn, the optimal (experimenter-imposed) decision criterion is always univariate and at the midpoint of that dimension.

*: Publicly available at https://github.com/ronald976/RonTree

Curricula

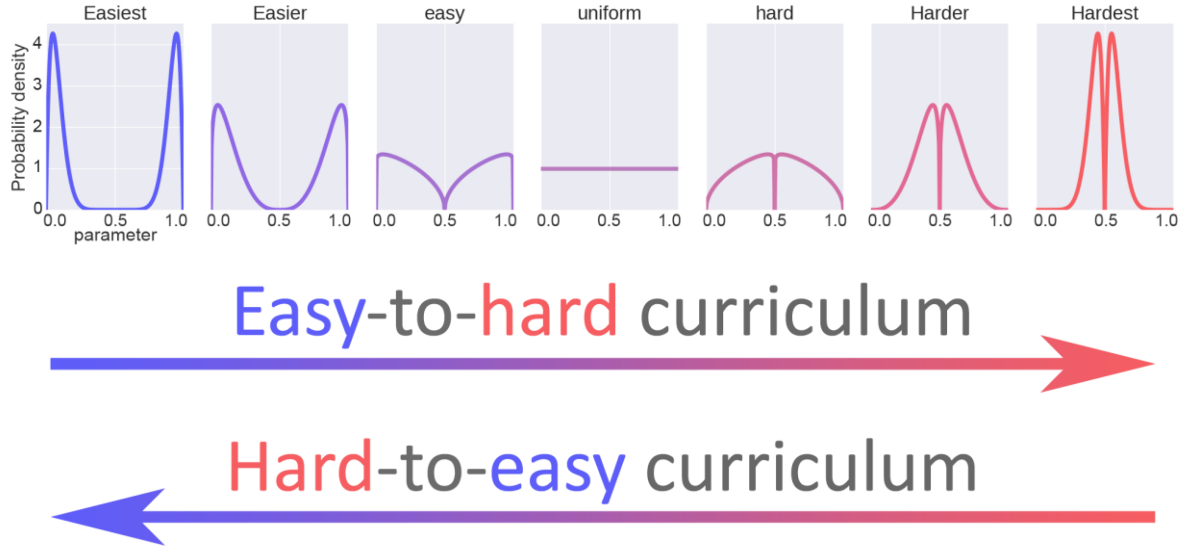

Figure 2: Curriculum design. Curricula were designed such that over the course of 7 experimental blocks, dimensions gradually became more difficult to discern (easy-hard) or vice versa (hard-easy).

The curriculum design is visualized in Figure 2. Curricula were designed such that dimensions gradually became more difficult to discern (easy-hard) or vice versa (hard-easy). Each experiment consisted of 7 blocks, and after each block, the distribution from which parameters were drawn would change, either decreasing (easy-hard) or increasing (hard-easy) variance within that dimension. Curricula were selected independently for the relevant and irrelevant dimension, resulting in a total of 4 curricula, in addition to a uniform control condition. An overview of this design is given in Table 1. In the remainder of this document, the resulting conditions will be referred to as indicated in this table.
To implement these curricula, parameters were drawn from a concatenated beta distribution, which was defined as follows (1):

Formula 1: Concatenated beta distribution.

For $0 \leq x \leq 0.5$: $\quad \text{pdf} = 2x^{\alpha-1}(1-2x)^{\beta-1}$

For $0.5 < x \leq 1$: $\quad \text{pdf} = 2(x-0.5)^{\beta-1}(2-2x)^{\alpha-1}$

Where pdf stands for probability density function and {α, β} parameters which define the shape of the distribution.

Crucially, this design ensures that at any given moment during the experiment, the portion of parameter space that has been observed by participants is on average equal between experimental conditions. Thus, differences between curricula cannot be attributed to participants seeing a wider range of stimuli in some conditions over others.

Statistical analysis

In each experiment, statistical analysis was done using a multivariate general linear model. This model took two independent variables: relevant dimension curriculum and irrelevant dimension curriculum, and assessed their effects on two dependent variables: early learning and terminal accuracy. To determine how curriculum affects early learning, a powerlaw function was fitted to test accuracy in the first two out of seven experimental blocks on a per-subject basis. The end points of this fit were then used as inputs for the general linear model. Second, to investigate differences in terminal accuracy, per-subject mean accuracies in the last two blocks were computed. In addition, the Tukey HSD test was used for post-hoc analyses.

Table 1: Experimental curricula

Condition	Relevant dimension curriculum	Irrelevant dimension curriculum
EH-EH	Easy-hard	Easy-hard
HE-HE	Hard-easy	Hard-easy
EH-HE	Easy-hard	Hard-easy
HE-EH	Hard-easy	Easy-hard
Uniform (control)	Uniform only	Uniform only

EXPERIMENT 1

In Experiment 1, we trained a multi-layer perceptron on raw parameter values. Consequently, this experiment is informative about learning patterns inherent to the curriculum design and the properties of the network architecture, but does not yet take any features of the specific stimulus into account.

Method

The MLP architecture is visualized in Figure 3. The network consisted of 2 input nodes, 2 hidden layers containing 256 nodes each and a single-value output. Inputs were bidimensional and drawn from the [0, 1] interval, with probabilities determined by their respective curricula and the current block. In addition, some random noise was added to the inputs. Noise values were drawn from a Gaussian distribution with mean 0 and standard deviation 0.1 and were drawn independently for both dimensions. Hidden layers used a rectified linear activation function. The output layer used a step function with threshold 0.5 as an activation function, converting output to a single binary number, indicating predicted category membership. Weights and biases were initialized randomly.

The network learned by minimizing its cost function, which was the cross entropy between predicted and true category labels. This was done using the Adam algorithm, a gradient-descent based optimization method (Kingma & Lei Ba, 2015). During each cycle, the network was trained on a single epoch ("full run") of 2499 training examples in minibatches of 100 trials. After each minibatch, the network was evaluated on the full test set of 1001 examples to monitor the learning progress. During these evaluations, weights were not updated. After completing all trials, weights and biases were reset, a new test and training set were generated and the next cycle began. In total, 1000 cycles were run per condition.

Results

Figure 4 displays test set accuracy as a function of training trials and condition. In this experiment, curriculum influenced learning (F(2,594) = 234854.15, p < 0.001). Post hoc comparisons revealed that both for early learning and for terminal accuracy, there were significant differences between all levels of relevant and irrelevant dimension (p < 0.002 for all contrasts). In the relevant dimension, early learning was slightly improved in EH (*M* = 0.789, *SD* = 0.003) compared to control (0.765+-0.005), but more markedly attenuated in HE (0.637+-0.003). A similar pattern persisted for terminal accuracy, with EH (0.919+-0.001) outperforming control (0.909+-0.001) and accuracy being lowest in HE (0.870+-0.001). In the irrelevant dimension, early learning was greater in the uniform condition (0.765+-0.005) than in EH conditions (0.721+-0.003) with HE performing worst (0.705+-0.003). Again, this pattern remained consistent for terminal accuracy, albeit with smaller mean differences than for early learning. Terminal accuracy was highest for uniform (0.909+-0.001), intermediate for EH (0.896+-0.001) and lowest for HE (0.893+-0.001).

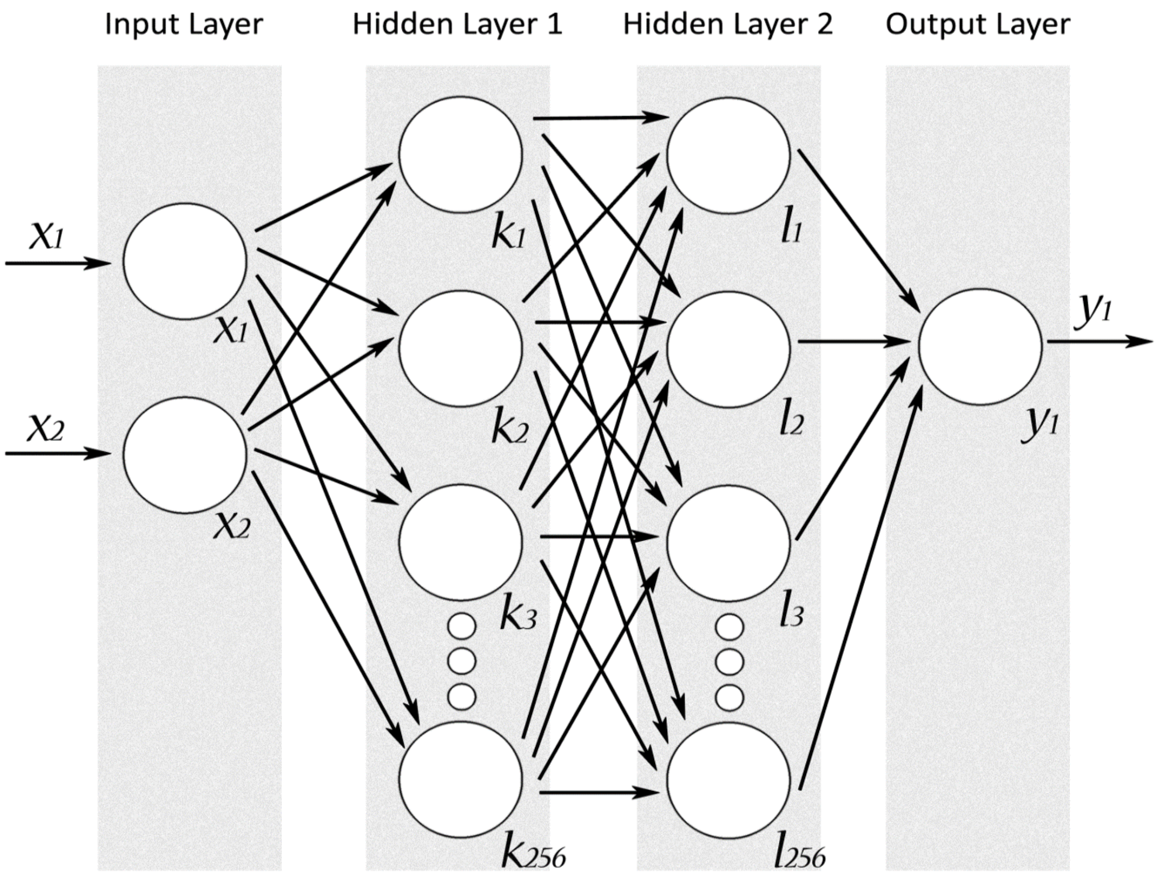

Figure 3: Multi-layer perceptron network architecture for Experiment 1. Inputs (x) go through hidden layer 1 (k), then through hidden layer 2 (l), in order to finally compute output (y). Gray rectangles indicate layers, while circles indicate nodes and arrows indicate weights.

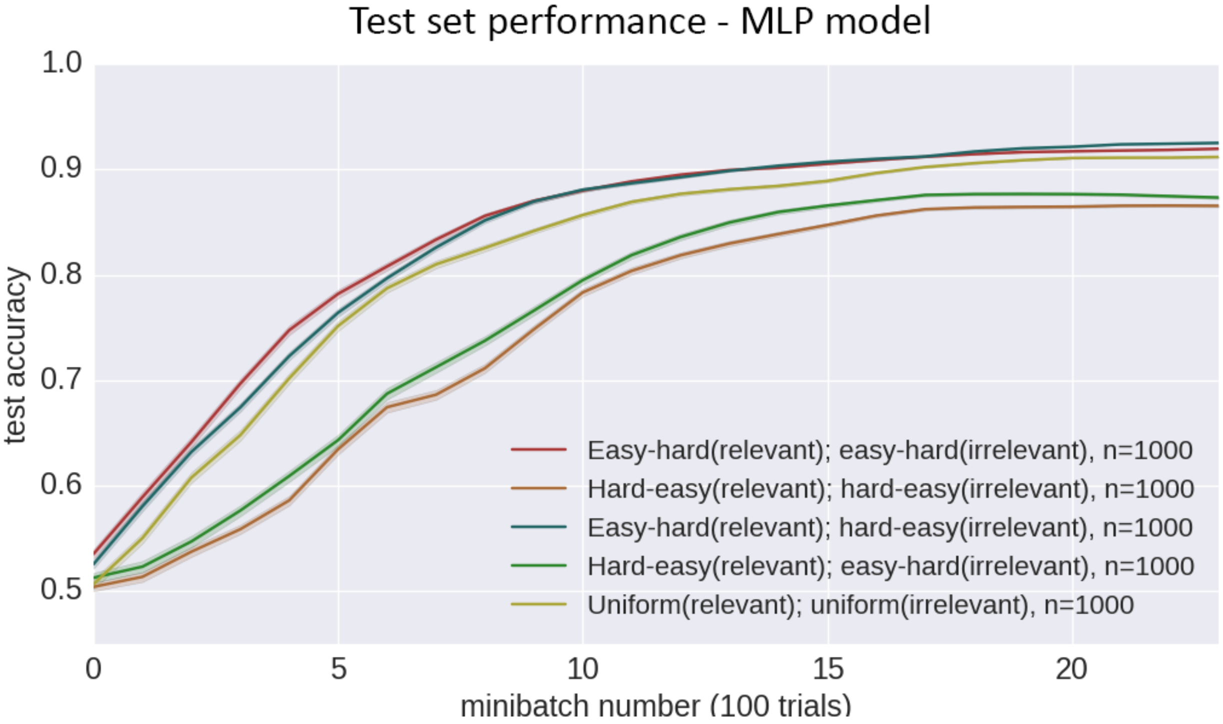

Figure 4: Test accuracy for multi-layer perceptron. Test set accuracy was computed after every minibatch, which consisted of 100 training trials. Here, the mean test accuracy after each minibatch is visualized per condition. Note: in this and all further plots, minibatch and block numbers are counted starting from zero. In all plots which have error bars, error bars represent standard errors.

EXPERIMENT 2

In Experiment 2, we trained a convolutional neural network on real images. These networks operate by sliding a kernel over the input image, dividing it into smaller sections for which the network then computes diagnostic features, which become increasingly complex in deeper layers of the network (LeCun, Bottou, & Bengio, 1998). Hence, they bear a striking resemblance to visual signal transduction in the mammalian brain, on which they are based (Pinto et al., 2009). In visual cortex, processing relies on receptive fields, which are spatially segregated sections of retinal input. Initially, these are represented in basic features, but representations become increasingly complex along the visual hierarchy. It should be noted that CNNs remain vastly different from real brain processing in many respects, including their method of weight updating or "learning", which is based on gradient back-propagation, which faces several difficulties with biological plausibility (Bengio, Lee, Bornschein, & Lin, 2015)

Method

The used CNN architecture is outlined in Figure 5. The network consisted of a 60x60x3 input layer, then two convolutional layers , the first being 60x60x32 and the second being 30x30x64, both using 2x2 max-pooling, then a fully-connected layer with 1024 nodes, ultimately producing a single value binary output. Weights and biases were initialized randomly, and a dropout rate of 25% (i.e. keep rate of 75%) was used in the fully-connected layer. Kernel size was 3x3 pixels, with a stride of 1 in both x and y dimensions. As in the MLP, the cost function was the cross entropy between the predicted and true category labels, which was minimized using the Adam method.

The CNN was trained and tested on 60x60 RGB images. We ran 60 cycles per condition per relevant dimension (leafiness or branchiness). In each cycle, the network was trained on a single epoch, consisting of 8000 training examples in minibatches of 100 trials. After each minibatch, the network was evaluated on the full test set of 1001 examples, during which weights were not updated. After completing all trials, weights and biases were reset, and the next cycle began.

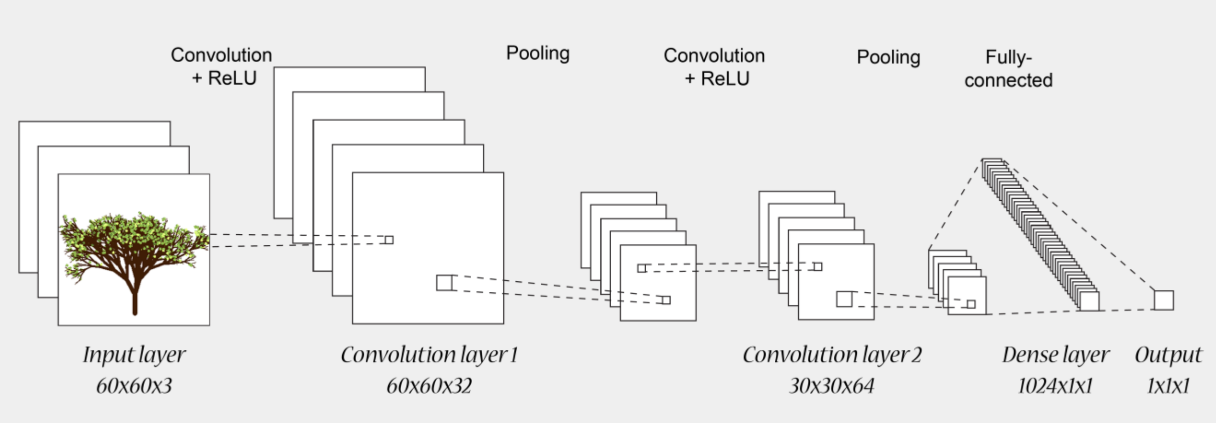

Figure 5: Convolutional neural network architecture for Experiment 2. Inputs were 60x60 RGB images. These were processed by two convolutional layers, each using a rectified linear unit (ReLU) activation function and 2x2 max pooling. Then, the signal passed through a 1024-node dense layer, which was similar to hidden layers in the MLP, but received many more (14400) inputs. Finally, the network produced a single output, which was put through a stepwise activation function to produce a single binary value.

Results

Figure 6 displays training set performance as a function of training trials and condition, and Figure 7 shows performance at test. At test, there was an effect of curriculum on learning (F(2,594) = 234854.15, p < 0.001). Post hoc comparisons using the Tukey HSD test revealed that early learning is affected by relevant dimension (p < 0.001 for all contrasts). Early learning was improved in EH (mean accuracy 0.646 +- 0.004) compared to control (0.597 +- 0.004), but was attenuated in HE (0.543 +- 0.004). Differences persisted for terminal accuracy, but with reduced magnitude (p < 0.007 for all contrasts). Terminal accuracy was greatest for EH (0.852 +- 0.002), intermediate for control (0.842 +- 0.003), and lowest for HE (0.809 +- 0.003).

Irrelevant dimension curriculum did not affect early learning (p > 0.923 for all contrasts), but there was a small but significant effect of irrelevant effect on terminal accuracy. This was driven by control (0.842 +- 0.003) and EH (0.835 +- 0.002) curricula outperforming HE (0.826 +- 0.002) (p < 0.008 for both contrasts).

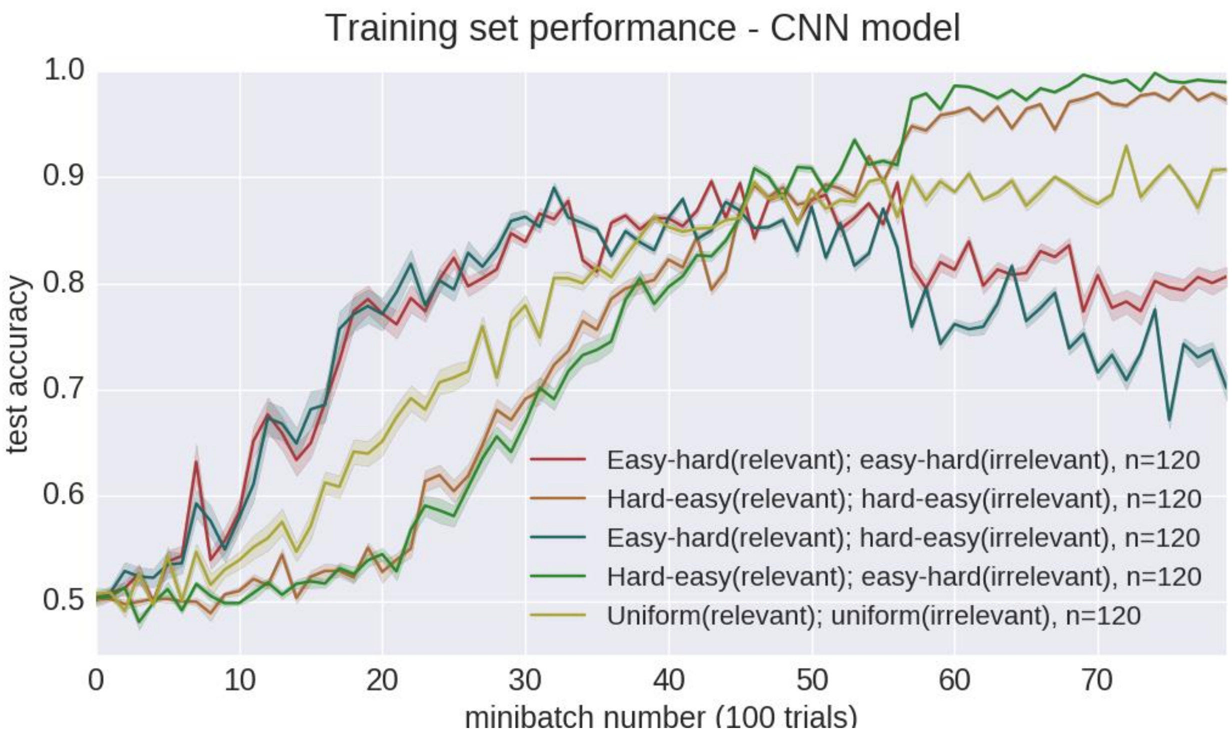

Figure 6: Training accuracy for Convolutional Neural Network. Training accuracy was computed for every minibatch, which consisted of 100 training trials. Here, the training accuracy in each minibatch is visualized for each condition separately.

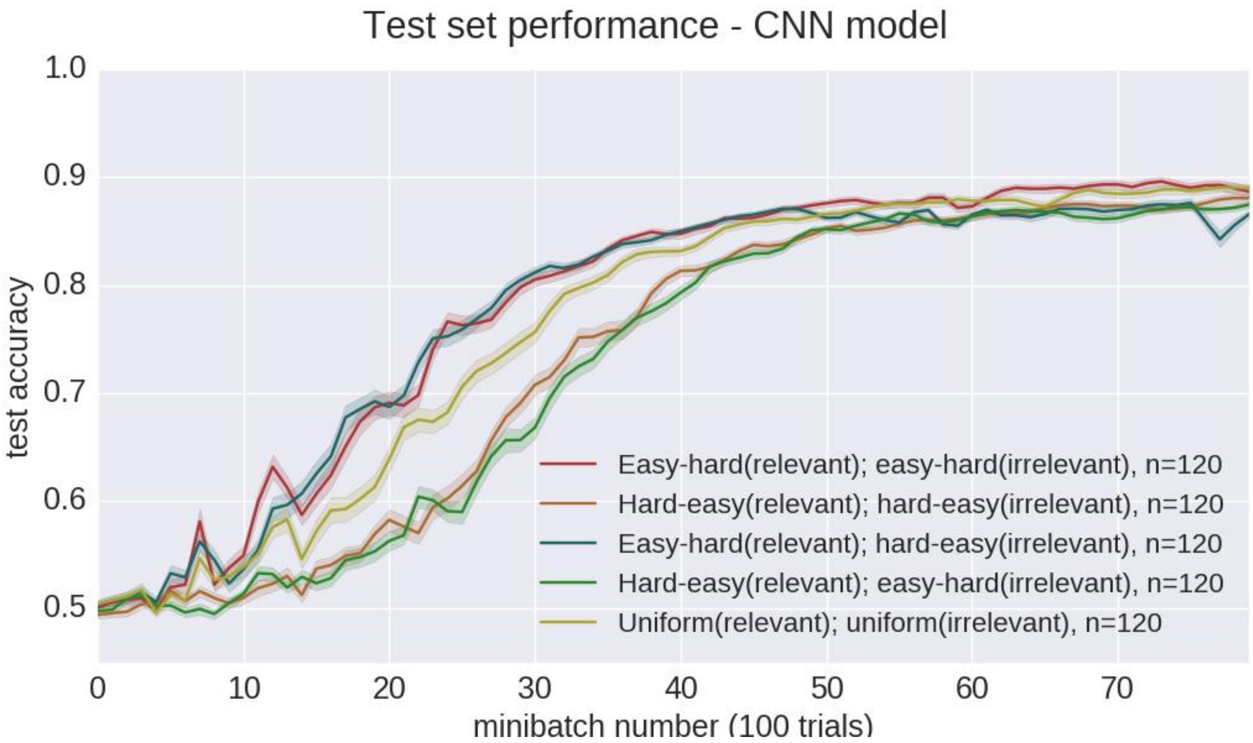

Figure 7: Test set performance in Convolutional Neural Network. Test accuracy was computed on a separate test set after every minibatch. Here, the test accuracy after each minibatch is visualized per condition.

EXPERIMENT 3

In the third and final experiment, we investigated how real participants performed on the task. Participants were tested online using Amazon Mechanical Turk, a commercial online participant recruitment platform, which has been shown to be effective in recruiting demographically diverse participants (Buhrmester & Kwang, 2011) and to be consistent with laboratory results (Crump, McDonnell, & Gureckis, 2013)

Method

In total, 80 participants participated in this experiment (16 per condition). 67 of these were used in the final analyses, with the remaining 13 being rejected for not performing above chance (<55% accuracy) or for only using a single key to respond.

Participants received the following instructions:

> *“Thank you for participating in our experiment. In this experiment, you will be shown several trees. It is your task to find out which of these belong to Quercus fictifolia, a newly discovered subspecies of oak. While you have no idea what these trees look like now, we will provide you with feedback, telling you when you were correct or incorrect. Note that sometimes, you receive no feedback - this is intended. You can use the following keys to indicate your belief:*
> *Left arrow = not a fictifolia tree*
> *Right arrow = fictifolia tree*
> *You can earn £2 + up to £5 bonus. The task will be over faster if you perform well and the payout will be higher, so please do your best!“*

The experiment consisted of 7 blocks, each having the following structure: 21 training trials, then 10 test trials, then 21 training trials, then 10 test trials. After each block, participants could take a small break and continue when ready.

Trials had the following structure: first, a cue appeared, indicating that a new stimulus would appear soon. This cue remained on screen for 500ms. Then, the stimulus would appear and remain visible for 2500ms. After stimulus offset, a loading bar appeared for 5200ms, which grew smaller over time, indicating how much time participants had left to respond. Participants could respond at any time between stimulus onset and loading bar offset, prompting immediate feedback. If they responded correctly, the word "correct" appeared in green for 600ms. If they responded incorrectly or not at all, the word "incorrect" appeared in red for 800ms. In no-feedback blocks (i.e. for test trials), the feedback text was always "no feedback", which appeared in gray for 700ms. All stimuli were presented on a light gray background at the screen center. Experimental stimuli were 302x302 pixel trees.

Key assignment was randomized, ensuring that performance started at 50%, regardless of any initial preferences participants might have. Stimulus order was shuffled within blocks (separately for training & test), causing each participant to have a unique trial order.

Results

Figure 8 displays training set performance as a function of training trials and condition, and Figure 9 shows performance at test. Curriculum affected learning (F(2,61) = 2382.05, $p < 0.001$). Post hoc comparisons reveal that early learning was influenced by relevant dimension. Specifically, it was impaired in HE (0.630 +- 0.024) compared to EH(0.772 +- 0.021) and control (0.792 +- 0.034) ($p < 0.001$ for both contrasts). There were no main effects of irrelevant dimension and there were no effects on terminal accuracy. Notably, however the mean amount of early learning in the HE-HE condition was in-between the low rate of HE-EH and the higher rate of all other conditions. Illustrating this, possible interactions between relevant and irrelevant curricula are plotted in Figure 10. This constitutes a significant interaction, whereby higher initial variance in the irrelevant dimension attenuates early learning, but only when the relevant dimension was initially difficult ($p = 0.047$).

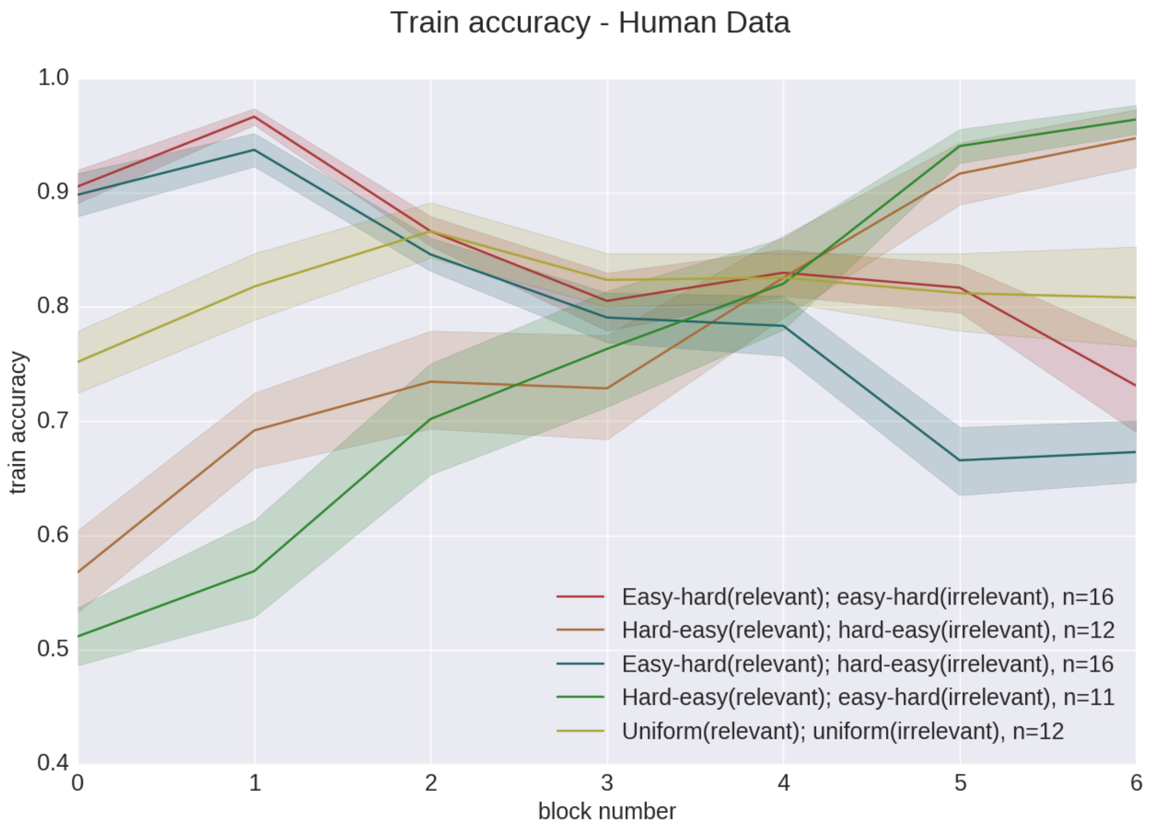

Figure 8: Training data – Human participants. Train accuracy per experimental block per condition.

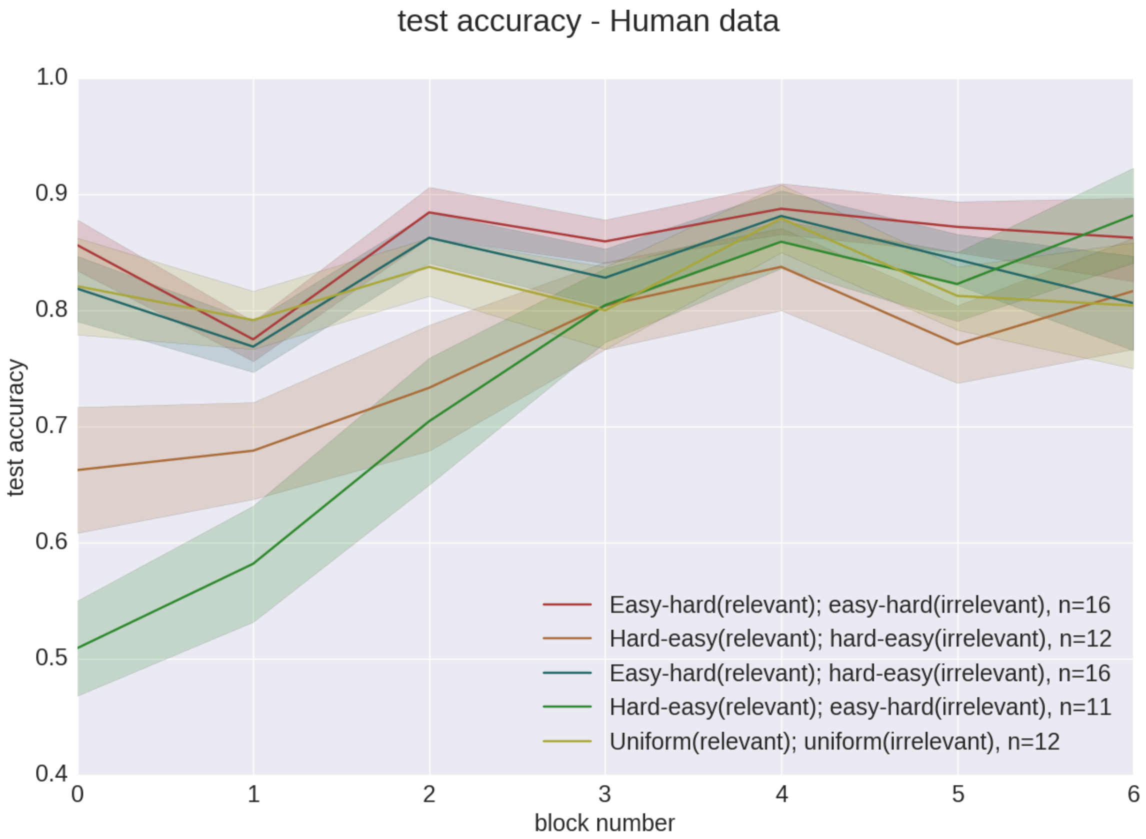

Figure 9: Test performance of human participants. Mean test accuracy in each experimental block, separated by condition.

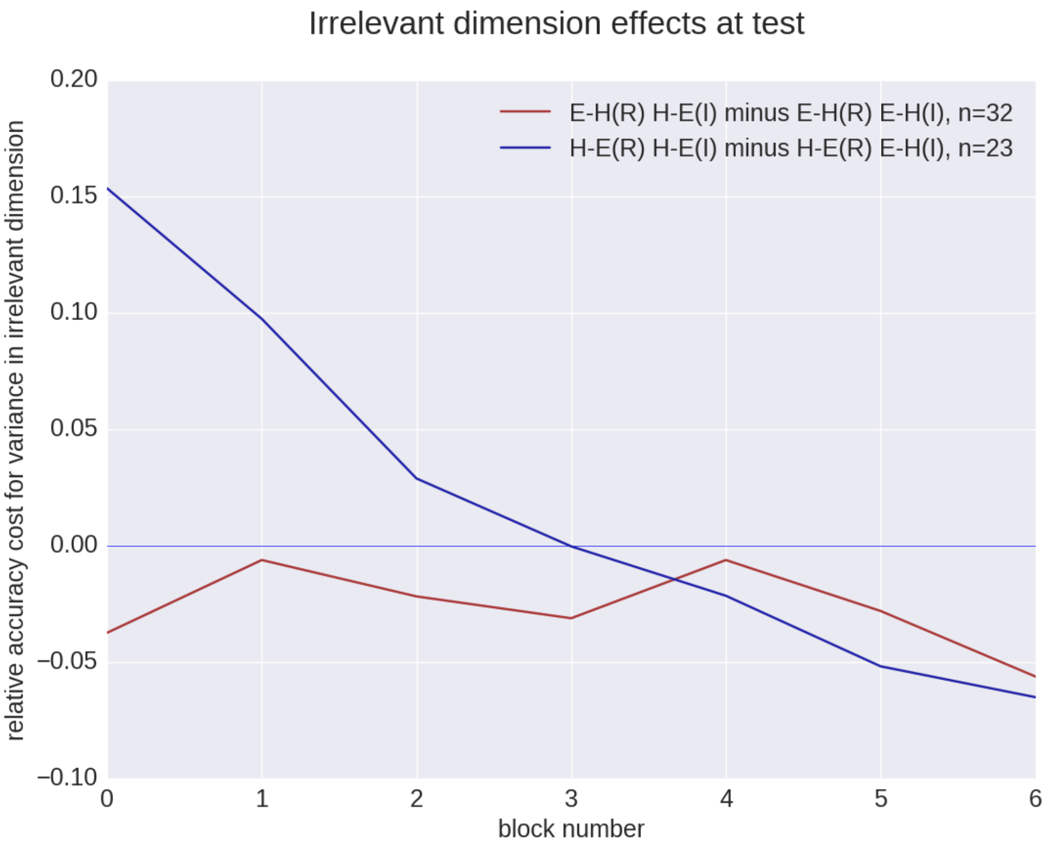

Figure 10: Relative accuracy cost for variance in irrelevant dimension. X-axis represents block number, while y-axis represents the accuracy cost of higher initial variance in the irrelevant dimension, for curricula in which the relevant dimension is easy-hard (red) and for curricula in which the relevant dimension is hard-easy (blue). Accuracy costs represent fractions: an accuracy cost of 0.15 is synonymous with a 15% accuracy drop relative to optimal performance.

GENERAL DISCUSSION

Conclusions

In this series of experiments, we sought to investigate how choice of learning curriculum can facilitate or impair learning of naturalistic visual categories with minimal task instruction. We hypothesized that showing easy examples first would bootstrap learning by guiding attention to diagnostic features. Conversely, beginning with the difficult examples was expected to hinder learning by obscuring which variance among stimuli was category-relevant. Our human data confirmed that early learning was attenuated in HE conditions in the relevant dimension. No evidence was found for a benefit conferred to early learning in EH conditions, although it is difficult to make conclusive statements regarding this, given that EH and uniform conditions already reached plateau performance in the first block. Thus, potential differences could be obscured by a ceiling effect. To investigate the potential facilitative function of EH curricula, future studies should either increase task difficulty, or further limit the number of consecutive training trials.

Second, we were interested in how variance in a task-irrelevant dimension affected learning. In our human data, there were no main effects, but there was evidence for an interaction whereby task-irrelevant variance impairs early learning, but only when the relevant dimension is relatively difficult. An intuitive explanation for this interaction would be that extraneous dimensions, which serve as distractors, are most effective when they are relatively salient compared to task-relevant dimensions.

Comparison to ANNs

With regard to our predictions on the effects of relevant dimension on learning, it is interesting to note that the MLP model already adheres to these predictions, despite the fact that this model lacks both attention and concrete stimulus features. Since the learning rule for backpropagation ensures that weight updates are directly proportional to the magnitude of the input, we chose the interval [0, 1] as our parameter space, to ensure that regardless of current beta distribution, the mean input value was always identical (always 0.5). One possible explanation that predicts this pattern of responses would be that the added noise in the model only leads to incorrect training examples (i.e. examples where the input value was incongruent with the assigned category label) when the input value was close to the decision boundary before the addition of noise. Then, the number of examples close to the boundary is directly proportional to the number of incorrect training examples, impairing the performance of difficult-first curricula and improving the performance of easy-first curricula. However, in a pilot study which is not reported here, we ran simulations which were completely identical, but without the addition of noise. Regardless, the qualitative pattern of responses was still the same, rendering this hypothesis unlikely to drive the observed results. Instead, a more plausible explanation is that higher input variance contributes to faster convergence. This interpretation is supported by the finding that in our MLP simulations, early learning was higher for conditions where the irrelevant dimension was EH, which had higher initial input variance, but was completely orthogonal to the stimulus category. In future studies, it would be interesting to determine the conditions under which this effect occurs, and which computational mechanism underlies it.

In both MLP and CNN simulations, having an HE curriculum in the relevant dimension did not only impair early learning, but also lowered terminal accuracy. Given the very low slopes at the end of learning curves in both simulations, this is likely attributable to a different convergence point rather than a different convergence speed. An intuitive explanation for this would be that in HE curricula, examples in the final blocks are far from the decision boundary. Thus, if an agent has already learned an approximately accurate heuristic, new examples will not contribute to an improvement in their accuracy. To illustrate, if an agent has already learned some representation of 'click right when input leafiness is above 0.7', easy examples where input leafiness is always <0.3 or >0.7 will not aid them in further correcting their representation of the decision boundary. Thus, aside from impairing early learning by obscuring the relevant dimension, the HE curriculum may also hinder late learning by not providing accurate category boundary information in later stages of learning. This suggests that while training on an 'idealized training set', which only contains extreme examples, can confer benefits to real-world category learning compared to random sampling (Giguère & Love, 2013; Hornsby & Love, 2014), an aptly designed fading procedure could be more effective than idealized training by facilitating continuation of learning in later stages, at least when category labels are sufficiently deterministic.

One key difference between the ANN simulations and real participants is that participants learn extremely quickly, in some conditions even reaching ceiling performance after within the first block. In all likelihood, this difference is attributable to more comprehensive priors in humans, who can relate new stimuli to a lifetime of experience, while the neural networks start from scratch. With this in mind, it would be interesting to investigate how unsupervised pre-training relates to this difference, both to as a model for human priors, and to investigate its role in facilitating transfer by constructing purely data-driven, rather than task-driven stimulus features (Erhan, Courville, & Vincent, 2010). Alternatively, it could be interesting to relate this to mere exposure effects in human category learning (Folstein, Gauthier, & Palmeri, 2010), which is conceptually similar to, but has not directly been compared to unsupervised pre-training.

Critical notes

One thing that should be kept in mind in interpreting the results of this study is that due to the different computational and practical costs of the three conducted experiments, the sample sizes between experiments are vastly different. Because of this, some minute differences in the MLP model (N=1000 per condition) are highly significant, whereas potentially much more meaningful effects in the human data (N=16 per condition before dropout) are not. For example, the 0.3% difference in MLP terminal accuracy between irrelevant EH and HE is significant with $p < 0.001$, whereas support for the potential interaction of irrelevant and relevant dimensions described in the conclusions, which peaks at around 15% difference, barely reaches significance. While it is not the author's intent to discourage scientific scrutiny, the current study forms a cautionary tale against the sole reliance on p-values and emphasizes the importance of effect sizes and exploratory analyses, and perhaps of considering alternatives to frequentist inference.

Furthermore, one of our results was that in the MLP and CNN models, the uniform irrelevant condition performs well both at early learning and for terminal accuracy, in some cases significantly better than EH and HE. In interpreting this result, it is important to consider that the uniform irrelevant condition was always paired with the uniform relevant condition, and that this effect is likely driven by the absence of relevant permutations in the design rather than any properties of the uniform irrelevant condition itself.

One other critical note should be made with regard to the stimuli used in the CNN experiment. For reasons of computability, images were down-sampled to 60x60 pixels, compared to the 302x302 used for human participants. However, in previous machine vision research, more complicated tasks have been performed with lower-resolution images. For example, the CIFAR100 dataset is used for 100-class classification using 32x32 images (Krizhevsky, 2009). Thus, 60x60 images should be sufficient for a single-class classification problem, such as that in the current study.

Recommendations for future research

One important goal of the current study was to enable future extensions of this experiment. Foremost among these is the question how learning curricula affect transfer, rather than initial learning. Specifically, after training a participant on either the leafiness or branchiness of stimuli as in the current study, it would be interesting to see how this learning might facilitate or impair learning in a second stage, where the relevant and irrelevant dimensions would be reversed. In such an experiment, enhanced performance at transfer when variance in the task-irrelevant dimension is high would suggest that saliency facilitates re-use, even when participants are trained to ignore this salient dimension. Alternatively, effects in the opposite direction could be attributable to greater need to suppress the irrelevant dimension in the source task amounting to more effective training to ignore this feature. Either result would be informative for models of flexible hypothesis selection in human agents, which currently do not account for the selectivity of attention (Khan et al., 2011). In addition, such studies have the potential benefit of informing the field of machine learning, where the superior performance of humans compared to machines on transfer learning tasks is a major open problem (Canini, Shashkov, & Griffiths, 2010; Torrey & Shavlik, 2009). Second, the current study made use of a decision rule which was relatively easy to verbalize – trees could have many or few branches, and they could have many or few leaves. This raises the question if the observed curriculum effects are just a consequence of increased saliency of an explicit rule, and to what extent curricula could assist in learning difficult-to-verbalize decision rules.

Concluding remarks

Taken together, the results presented in this study suggest that identification of relevant dimensions and/or hypothesis testing on dimension space is integral to category learning of realistic stimuli. They also show both striking similarities (such as the qualitative effects of learning curricula) and marked differences (such as the speed and shape of learning) between humans and successful machine learning algorithms such as deep neural networks, prompting the key question of what is unique about human category learning.

Acknowledgements

Special thanks go to Christopher Summerfield for supervising the project and delivering valuable feedback. Finally, the author would like to thank everyone from the Summerfield lab for creating a positive, welcoming atmosphere during the project and for the many thought-provoking discussions.

Literature

Ashby, F., & Maddox, W. (2005). Human category learning. *Annu. Rev. Psychol.* Retrieved from http://www.annualreviews.org/doi/abs/10.1146/annurev.psych.56.091103.070217

Bengio, Y., Lee, D.-H., Bornschein, J., & Lin, Z. (2015). Towards Biologically Plausible Deep Learning. *arXiv Preprint arxiv:1502.0415*, 18. https://doi.org/10.1007/s13398-014-0173-7.2

Buhrmester, M., & Kwang, T. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *On Psychological Science*. Retrieved from http://pps.sagepub.com/content/6/1/3.short

Canini, K., Shashkov, M., & Griffiths, T. (2010). Modeling Transfer Learning in Human Categorization with the Hierarchical Dirichlet Process. *ICML*. Retrieved from https://pdfs.semanticscholar.org/6a8b/f4ccae0c229089ea7f5ea2527fcea434c07d.pdf

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE*, *8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

Erhan, D., Courville, A., & Vincent, P. (2010). Why Does Unsupervised Pre-training Help Deep Learning ? *Journal of Machine Learning Research*, *11*, 625–660. https://doi.org/10.1145/1756006.1756025

Folstein, J., Gauthier, I., & Palmeri, T. (2010). Mere exposure alters category learning of novel objects. *Frontiers in Psychology*. Retrieved from http://www.frontiersin.org/Journal/DownloadFile/1/2216/1959/1/21/fpsyg-01-00040_pdf

Giguère, G., & Love, B. C. (2013). Limits in decision making arise from limits in memory retrieval. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(19), 7613–8. https://doi.org/10.1073/pnas.1219674110

Hornsby, A. N., & Love, B. C. (2014). Improved Classification of Mammograms Following Idealized Training. *Journal of*

Applied Research in Memory and Cognition, *3*(2), 72–76. https://doi.org/10.1016/j.jarmac.2014.04.009

Khan, F., Mutlu, B., & Zhu, X. (2011). How do humans teach: On curriculum learning and teaching dimension. *Advances in Neural Information Processing Systems*, 1449–1457. Retrieved from http://papers.nips.cc/paper/4466-how-do-humans-teach-on-curriculum-learning-and-teaching-dimension

Kingma, D. P., & Lei Ba, J. (2015). Adam: A Method of Stochastic Optimization. *arXiv Preprint arXiv:1412.6980*, 1–15. Retrieved from https://arxiv.org/abs/1412.6980

Krizhevsky, A. (2009). Learning Multiple Layers of Features from Tiny Images. Retrieved from https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

LeCun, Y., Bottou, L., & Bengio, Y. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*. Retrieved from http://ieeexplore.ieee.org/abstract/document/726791/

Medin, D. L., & Bettger, J. G. (1994). Presentation order and recognition of categorically related examples. *Psychonomic Bulletin & Review*, *1*(2), 250–254. https://doi.org/10.3758/BF03200776

Pashler, H., & Mozer, M. (2013). When does fading enhance perceptual category learning? *Journal of Experimental Psychology:* Retrieved from http://psycnet.apa.org/journals/xlm/39/4/1162/

Penn, D., & Holyoak, K. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain*. Retrieved from http://journals.cambridge.org/abstract_S0140525X08003543

Pinto, N., Doukhan, D., DiCarlo, J. J., Cox, D. D., Fukushima, K., Bishop, C., … Eberhart, R. (2009). A High-Throughput Screening Approach to Discovering Good Forms of Biologically Inspired Visual Representation. *PLoS Computational Biology*, *5*(11), e1000579. https://doi.org/10.1371/journal.pcbi.1000579

Torrey, L., & Shavlik, J. (2009). Transfer learning. *Handbook of Research on Machine Learning*. Retrieved from https://books.google.com/books?hl=nl&lr=&id=gFpKXO8H_6YC&oi=fnd&pg=PA242&dq=Lisa+Torrey+and+Jude+Shavlik+transfer+learning&ots=6LScQPkvjv&sig=vQm6JbT2IqHLYJWwVe_o36843kY