COMPARING SUPERVISED LEARNING DYNAMICS: DEEP NEURAL NETWORKS MATCH HUMAN DATA EFFICIENCY BUT SHOW A GENERALISATION LAG

Lukas S. Huber
University of Bern & University of Tübingen

Fred W. Mast
University of Bern

Felix A. Wichmann
University of Tübingen

ABSTRACT

Recent research has seen many behavioral comparisons between humans and deep neural networks (DNNs) in the domain of image classification. Often, comparison studies focus on the end-result of the learning process by measuring and comparing the similarities in the representations of object categories once they have been formed. However, the process of how these representations emerge—that is, the behavioral changes and intermediate stages observed during the acquisition—is less often directly and empirically compared.

Here we report a detailed investigation of the learning dynamics in human observers and various classic and state-of-the-art DNNs. We develop a constrained supervised learning environment in which we align learning-relevant conditions such as starting point, input modality, available input data and the feedback provided. Across the whole learning process we evaluate and compare how well learned representations can be generalized to previously unseen test data.

Our findings indicate that DNNs demonstrate a level of data efficiency comparable to—and sometimes even exceeding that—of human learners, challenging some prevailing assumptions in the field. However, comparisons across the entire learning process also reveal representational differences: while DNNs' learning is characterized by a pronounced generalisation lag, humans appear to immediately acquire generalizable representations without a preliminary phase of learning training set-specific information that is only later transferred to novel data.

1 INTRODUCTION

Representational alignment is the process of comparing and harmonizing the internal representations of different information processing systems such as deep neural networks (DNNs) and humans (Sucholutsky et al., 2023). Two aspects are crucial for this endeavour: First, representations need to be measured across systems to enable comparison.[1] Second, when disparities are uncovered, the focus shifts to increasing alignment by updating the representation of at least one of the compared systems. To do the latter effectively, the sources of the observed representational differences have to be identified first. Which aspects of the system or the system's input data have to be altered to increase representational alignment? Finding the right aspects is arguably a non-trivial task since the search space of differences between systems like DNNs and humans is vast, and not all of these variations are pertinent to the representations themselves.

Here we argue that to better pinpoint the origin of representational differences—and thus to provide the basis for increasing alignment—it may be crucial to align and compare the process in which representations are acquired. For each learning condition aligned during representation acquisition, the

[1] Either this measurement is directly comparable or needs to be bridged into a shared space to allow for comparison.

search space for possible sources of differences becomes more constrained: Eliminating potential relevant causes for representational divergence during the learning process enables attempts to increase alignment to be more data-driven and directed, and thus more effective.

1.1 Background and related work

One way of identifying representational differences (or similarities) between humans and DNNs is by measuring and comparing their behavior in image classification tasks. [2] For an overview, see Wichmann & Geirhos (2023) and Sucholutsky et al. (2023), section 4.3.4. However, DNN-to-human comparisons are often fraught with difficulty (Funke et al., 2021; Firestone, 2020): Unlike DNNs, which typically learn from scratch using static, uni-modal data, humans process continuous, multi-modal information and leverage prior knowledge. Additionally, while DNNs are predominantly trained in a supervised manner, human learning heavily relies on interactions with unlabeled data[3]. A further central caveat is that comparative studies often focus only on the end-result of the learning process once representations have been formed (e.g., see Geirhos et al., 2018b; Muttenthaler et al., 2022; Dujmović et al., 2020; Zhou & Firestone, 2019; Kubilius et al., 2016). Only a few studies try to directly align the process in which representations are acquired in humans and DNNs.

In one extensive study investigating adult learning, human participants had to learn to differentiate pairs of novel objects (Lee & DiCarlo, 2023). Concurrently, they compared human learning with DNN models consisting of a fixed encoding stage and a tuneable decision stage. The decision stage was adapted based on feedback similar to that received by human subjects. While certain models produced relatively accurate predictions of human learning behavior, they showed weaker one- and few-shot generalisation capabilities compared to human learners. In another study, Orhan (2021) conducted a scaling experiment and found that current self-supervised visual representation learning algorithms require orders of magnitude more training data than humans to achieve similar performance in complex and realistic object recognition tasks. Another way to investigate and compare the process in which representations are acquired is by directly contrasting children's learning with DNNs. Huber et al. (2023) reported that children show robustness to image distortions with significantly less data exposure than DNNs. All these findings emphasize the efficiency of human learning processes over current DNNs. Here we aim to contribute to this emerging body of research by providing a side-by-side comparison of learning dynamics in supervised visual representation learning and assessing if these discrepancies persist when relevant aspects of the learning process are carefully controlled.

1.2 Present study

In the present study, we introduce a psychophysical paradigm that allows for the direct comparison of not only the end-results of visual representation learning but also its dynamics, from the initial absence of representation to the availability of generalizable representations. More specifically, we measure and compare how humans and various DNNs acquire representations for novel objects in a constrained environment with aligned learning conditions. We track behavioural changes across six epochs of a non-trivial image classification task—reflecting the train-test iteration process common in machine learning. At each epoch we measure and compare how well learned representations can be generalized to previously unseen data. During representation acquisition we align the following aspects of the learning process:

Learning goal. Both humans and DNNs are tasked with learning representations of 3D objects that were previously unknown to them.

Starting point & prior knowledge. In contrast to DNNs, which often begin learning from scratch, adult humans are equipped with a wealth of representations of natural and artificial objects. To level the playing field at least somewhat, we utilize DNNs pre-trained on ImageNet, hopefully mimicking at least partially the extensive representational knowledge adult humans possess already. Given that human and DNN performance on original ImageNet images is usually near-ceiling (Geirhos et al., 2018b; Russakovsky et al., 2015; Shankar et al., 2020; Dodge & Karam, 2017), it suggests that these

[2] This study focuses on supervised learning of visual representations. However, note that there is substantial research on human-to-machine comparison in few-shot learning (e.g., Lake et al., 2013; Morgenstern et al., 2019) and reinforcement learning (e.g., see Eckstein & Collins, 2020; Mnih et al., 2015; Silver et al., 2016).

[3] For a comprehensive discussion of the issues arising when using DNNs as models of human vision see Bowers et al. (2023) and the accompanying peer commentaries.

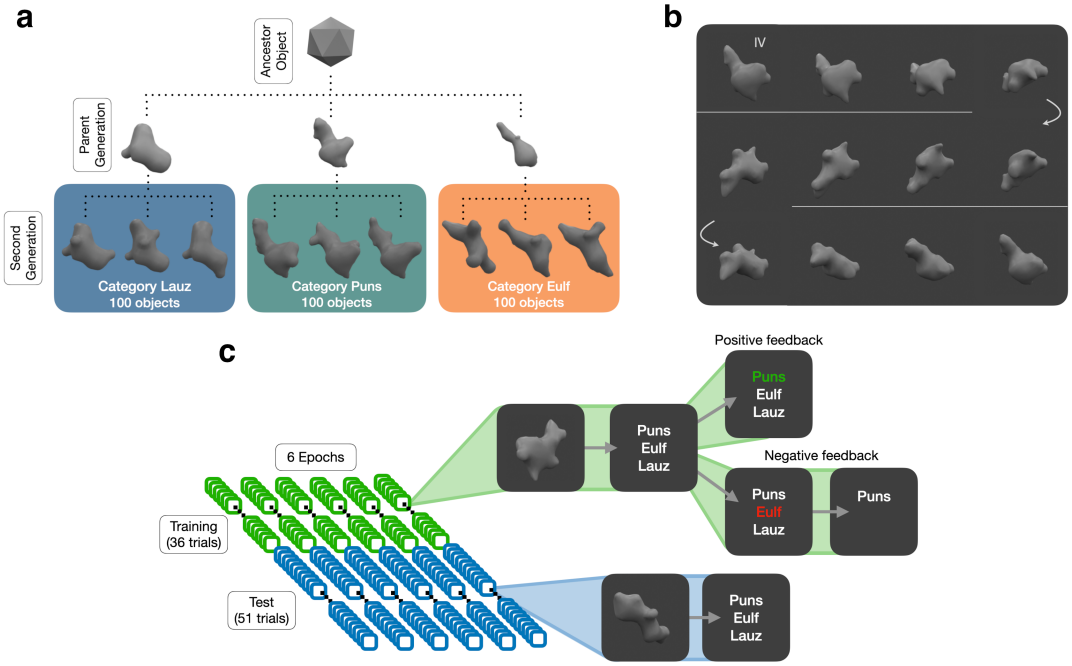

Figure 1: **Naturalistic novel objects were generated to serve as the basis for the creation of stimuli employed in the learning task**. (**a**) Taxonomy diagram illustrating the genesis of novel objects, originating from an icosahedron and spanning two generations, with the second generation serving as the basis for our stimuli. Objects sharing the same parent in the generation process are assigned to the same category. (**b**) Displayed is a set of twelve distinct renderings of an object from the category "Puns", showcasing 30-degree pitch rotations (x-axis) from its initial position (indicated by "IV"). Panel (**c**) shows a blueprint of the learning task. There are six epochs each consisting of 36 training trials (green) and 51 test trials (blue). All test sets contain different images to ensure novelty. During training, participants received corrective feedback mimicking supervised learning in DNN training. Depending on whether the given response was correct or not, the chosen category would either light up in green or red for one second. If the response was incorrect, this would be followed by a screen displaying the correct response for one second. In test trials, no feedback is provided, aiming to restrict supervised learning to the training phase.

pre-trained models and humans are starting from a similar level of existing image classification capacity for i.i.d data.

Input data. In typical learning scenarios, humans learn from continuous, multi-modal data (e.g., see Zaadnoordijk et al., 2022), whereas many DNNs are trained on static and uni-modal input. Here we eliminate this difference in learning conditions and match the input data such that both systems are only provided with the exact same static 2D images.

Learning modality. It has been argued, that humans do not typically learn from labelled data, but rather in a semi- or unsupervised manner (e.g., see Gibson et al., 2013). Consequently, it has been put forward that self-supervised models are a better proxy for human learning (e.g., see Lotter et al., 2020; Konkle & Alvarez, 2022). However, for behavioral outcomes in image classification tasks Geirhos et al. (2020) has not found a better alignment of self-supervised models with human observers as compared with supervised models (but see Storrs et al., 2021; Zhuang et al., 2021). Reflecting on these findings, our paradigm adopts a supervised learning environment for both human participants and DNNs. This approach facilitates the alignment of both systems within a consistent and controlled framework.

The resulting side-by-side comparison of learning dynamics enables us not only to ensure that observed differences or similarities in the way that information is represented are not mere artifacts of unaligned learning conditions, but also to systematically investigate presumed differences between humans and DNNs, previously hypothesized to account for the disparities in measured representations.

2 METHODS

2.1 STIMULI

To track the acquisition of object representations, a suitable dataset of stimuli is needed: Novel objects for which neither humans nor DNNs have representations available yet. However, for maintaining external validity and ensuring the generalizability of results, it is crucial that these objects and categories, despite being novel, still bear relevance to real-world conditions. In other words, objects should be naturalistic and the formation of categories within the dataset should not strictly adhere to arbitrary rule-based definitions but rather reflect the concept of family resemblance (Wittgenstein, 1968, 65ff).[4]

Hauffen et al. (2012) provided a method to generate 3D objects that satisfy these conditions (see also Brady & Kersten, 2003; Pusch et al., 2022). They describe a set of algorithms simulating the biological process of embryogenesis and the evolutionary process of phylogenesis to create novel objects. Here we employed these algorithms to create a dataset suitable to study the acquisition of novel representations in humans and DNNs.[5] Beginning with an icosahedron as an ancestor object, we initially generated a primary set of three distinct objects, forming the parent generation. Subsequently, from each of these parent objects, we created a second generation comprising 100 unique objects per parent: the Lauz, Puns and Eulf. Categories were established based on lineage, defining each category by the objects that shared a common parent in the initial generation (see Figure 1, Panel **a**).

For each of the 3D objects we created a series of 23 images (224×224 pixels) rendered using `Blender's` Python API (version 3.5.0). This rendering involved rotating each object along the pitch (x) and yaw (z) axis at intervals of 30 degrees, capturing various perspectives. An example of some rotations can be seen in Figure 1, Panel **b**. The process resulted in a stimulus set comprising 6,900 images in total (3 categories $\times$ 100 objects $\times$ 23 rotations). In a pretest we found that object categories were too similar and participants struggled with learning. Thus, to enhance intraclass coherence and interclass separability, we restricted the dataset to a subset of 3,450 images, focusing on the 50% most similar objects within each category. This selection was obtained using the structural similarity index measure (Wang et al., 2004, known for its typically reasonable alignment with human visual perception of similarity), to calculate pairwise similarities of the objects' initial position renderings within each category.

2.2 LEARNING ENVIRONMENT

For behavioral comparisons to be meaningful, we aimed to match and maintain consistency in the learning environment across both humans and DNNs. To this end we adapted the standard train-test protocol—commonly employed to train DNNs—also for human observers. The basis for training and test phases is a *forced-choice, time-limited image classification task*. Each trial of this task is built in the following way: After a fixation stimuli (combination of bulls eye and cross hair, see Thaler et al., 2013) (400 ms), participants are presented with a target-image in the centre of a computer screen for 300 ms, which is followed by a pink noise mask with $1/f$ spectral shape (300 ms). On a subsequent response-screen participants have to choose one of three categories, which they think corresponds to the image just seen (this is a standard procedure to collect psychophysical data, e.g., Huber et al., 2023; Geirhos et al., 2018a).[6] For the training phases, participants receive feedback on whether the given response was correct or not—mimicking supervised learning from labeled data. After being presented with all images of the training set, performance is evaluated on a test set to assess the progress in terms of transfer learning (for an overview see Pan & Yang, 2009). To align with the condition of DNNs having their parameters frozen during evaluation, effectively restricting learning to the training phase, we implemented two adjustments. Firstly, to prevent supervised learning during evaluation, human observers did not receive any feedback for test trials. Secondly, to guarantee that both humans and DNNs always encountered previously unseen images, we ensured each test set was composed of hitherto unseen images. By contrasting the development of training and test errors across several intervals

[4] According to Wittgenstein, category boundaries are often not sharp but fuzzy, characterized by a series of overlapping similarities rather than a set of necessary and sufficient features shared by all members—a concept he encapsulates as family resemblance.

[5] See Appendix A for further details.

[6] The short presentation time and the noise mask are employed to ensure the involvement of mainly feedforward mechanisms and thus to measure core object recognition as described in DiCarlo et al. (2012).

of training and testing, it becomes possible to track and compare the acquisition and generalizability of object representations. For an overview of the learning task, see Figure 1, Panel **c**.

2.3 Dataset composition

We randomly selected six objects per category for the training set, choosing two views for each object (the initial position and a 90-degree rotation on the yaw-axis), resulting in a total of 36 images. To ensure the test sets contained previously unseen images, we included novel perspectives of training objects as well as images of objects not encountered during training. Thus, we included novel perspectives of training objects as well as images of objects not encountered during training. Specifically, each of the six *different* test sets comprised eight novel perspectives (both minus and plus 30 and 60 degrees along the pitch and yaw axes) of one training object, and identical rotations (including the initial viewpoint) of a previously unencountered object from each class, resulting in 51 images per test set, resulting in 51 images per test set ($3 \times 8 + 3 \times 9$). This approach ensured that each test set included unique, previously unseen images and maintained a consistent level of difficulty across test sets by using identical rotations. Additionally, this allowed us to evaluate the ease of transferring learned representations to novel perspectives of familiar objects compared to recognizing familiar perspectives of new objects.

2.4 Observers

Humans. We collected data from 12 human participants recruited from the subject-pool of the University of Bern, in a controlled lab setting. In addition to the pool-intern course credit, they also had the chance to win a cafeteria voucher worth $ 15 if they exceeded a classification performance of 70%. If they did not, they received a can of mate tea instead. Images were presented at the center of the screen corresponding to $5° \times 5°$ of visual angle at a viewing distance of approximately 69 cm. A constant viewing distance was maintained by employing a chin-rest and adjusting the monitor's position to ensure participants did not have to look either up or down. Before iterating through all training and test sets, participants completed ten practice trials where they were familiarized with the task using images of sea lions, whales, and dolphins.

Models. We fine-tuned and evaluated two groups of DNN models. The first group consisted of classic convolutional neural networks (CNNs) used for image classification and included: AlexNet (Krizhevsky et al., 2012), VGG-16 (Simonyan & Zisserman, 2014), and ResNet-50 (He et al., 2016). To account for recent developments in computer vision research, we also included several state-of-the-art (SOTA) models with refined architectures and training protocols: a vision transformer(ViT Dosovitskiy et al., 2021), an attention-based network, relaxing the translation-invariance constraint of CNNs; a ConvNeXt (Liu et al., 2022), which is model that incorporates design principles and architectural elements from both CNNs and ViTs; and a second generation EfficientNet (Tan & Le, 2021), a model designed to be more efficient and effective, particularly in terms of speed and accuracy for both training and inference tasks. All models were pre-trained on ImageNet-1K and obtained either from the `PyTorch` model zoo (AlexNet, VGG-16, ResNet-50, ConvNeXt and EfficientNet) or from the `Hugging Face` model library (ViT). Prior to fine-tuning, the output layer of each model was modified to enable predictions across three distinct classes. Input data was not preprocessed and no data augmentation was applied—to keep learning as similar as possible between DNNs and humans. For all models we used a Adam optimizer with a constant learning rate of 0.001, and a batch size of 4.[7] For each model, we conducted twenty fine-tuning runs, spanning six epochs, using the identical training data set as used for the human observers. During training, predictions were continuously logged for each image, and after each epoch, predictions for the respective test set were also logged. Additional details about the employed models can be found in Appendix D.

3 Analysis & Results

We aimed to measure and compare the behavioral changes and intermediate stages observed during the acquisition of novel representations in humans and DNNs. To this end, we not only report and calculate learning progress for training data but evaluate for each epoch how well the learned representations can be generalized to previously unseen data.

[7]Since the test sets contained 51 images, the last batch of test images consisted only of 3 images.

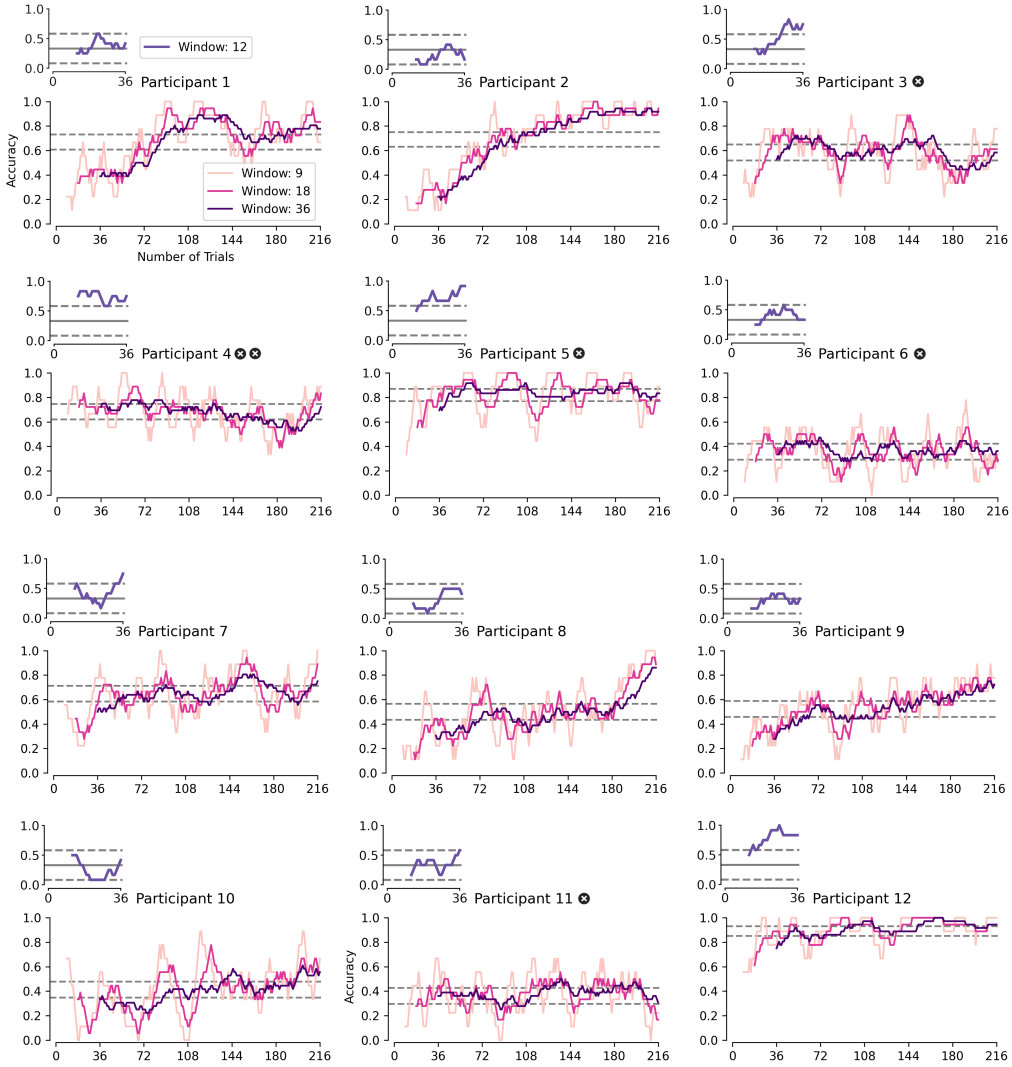

Figure 2: **Inter-individual differences in human learning dynamics.** For each participant we show moving averages of classification accuracy across the whole training (different trajectories mark different window sizes). Early learning is separately displayed in the small plots (purple trajectories). In these plots dashed lines mark binomial (Clopper-Pearson) confidence intervals around chance performance. We reason that if the trajectory does not exceed the upper bound of this confidence interval, this indicates that the observer does start training without any prior knowledge. The larger plots illustrate the evolution of training performance throughout the entire training period, as depicted by moving averages of varying sizes (different shades of pink). Here, dashed lines mark a confidence interval around the mean performance of individual observers. If the performance was significantly below the mean at the start, and is significantly higher at the end of training, this indicates that learning has occurred. Black circles with white crosses mark excluded participants because they either did not start learning from scratch (Participant 4), or did not show any indication of learning (Participants 3–6, and 11).

3.1 Ensuring equal starting conditions and occurrence of learning

To ensure a level playing field, we first confirmed that humans did not possess pre-existing representations of the novel objects. This was assessed by observing the classification performance in the initial training epoch, under the assumption that starting from zero knowledge should result in performance levels *not* exceeding chance at the onset of training. We evaluated each participant's classification accuracy using a moving average over 12 trials, which corresponds to one-third of the training set (see tiny plots in Figure 2, purple colored trajectories). When we assessed whether the initial accuracy

surpassed the upper bound of a confidence interval around chance level performance, only Participant 4 started with a classification accuracy significantly above chance level and was thus excluded from further analysis. All other participants initially demonstrated classification accuracies not significantly higher than chance, suggesting no pre-existing familiarity with the categories to be learned. However, during the first epoch, some showed signs of early learning, as indicated by their progression from chance-level accuracy to significantly higher scores by the end of the first epoch.

To evaluate whether the human observers actually learned during the training, we examined the change in training performance across all epochs (see larger plots in Figure 2). For each participants we calculated their mean classification performance across all training epochs. We considered a participant to have learned if they began the training with an accuracy significantly lower than their own mean performance, and then achieved a performance significantly higher than this average by the end of the training. This approach revealed that, of the remaining 11 participants, 7 exhibited clear learning based on this (strict) criterion. Consequently, our subsequent analyses are exclusively focused on this subset of 7 learners who demonstrated tangible improvement over the course of the training without prior knowledge at the beginning of the experiment.

3.2 Observing and comparing learning dynamics in humans and DNNs

Across humans and all models we consistently observe not only progressive learning but also effective transfer to previously unseen instances in the test datasets. In the subsequent paragraphs, we look closer at the learning dynamics of both humans and DNNs. In Figure 3, Panel **a** we show training and test accuracy as a function of epochs for all models and human observers. Further details can be found in Appendix B & D.

Trainig performance. All observers, spanning human participants and various model architectures, effectively learned to categorize the training data. By the end of the sixth epoch, humans demonstrated substantial learning, correctly identifying 78.96% of all training images. Classic CNNs achieved a mean accuracy of 93.10% by the best training epoch, a performance that SOTA models slightly improved upon, achieving an average accuracy of 94.91%. This close-to-ceiling performance on training data contrasts with the 78.96% accuracy attained by humans, indicating an advantage in data efficiency for models. However, this seems not surprising since DNNs tend to memorize training data, particularly when there are only few training samples available (Arpit et al., 2017). Such overfitting to the training data would result in poor generalization to the test sets, but, as detailed in the next paragraph, that was not the case in the present study.

Generalisation. Increasing test performance indicates that both humans and DNNs successfully transferred the acquired knowledge to previously unseen images, suggesting the emergence of generalizable representations and not mere memorization of the training data. Throughout the training, human participants exhibited a consistent increase in test performance, reaching 75.35% at epoch 6. While all models exhibited steadily increasing test performance until epoch 4, some plateaued at this point and others showed a decrease in test performance towards the end of training, indicating signs of overfitting. Consequently, only the peak test performance was considered as a measure of transfer learning in models, akin to early stopping. By this metric, classic CNNs reached a test performance of 76.14%, with SOTA models outperforming at 81.57%, placing humans in a similar range to that of classic CNNs. These results indicate that—at least in our aligned learning environment—DNNs show a level of data efficiency that is on par with, and occasionally surpasses, that of humans.

We also investigated how well humans and DNNs generalized to previously unseen perspectives of training objects vs. to familiar perspectives of previously unseen objects. The results suggest that there is no significant distinction in performance between these two kinds of test images (see Appendix C).

Early learning. To analyse early learning in the initial epochs, we computed a binomial (Clopper-Pearson) confidence interval centered around chance performance assuming $p = 33.\overline{33}\%$ and $n = 51$ (size of the test sets). Both humans and DNNs consistently outperformed the upper bound of the confidence interval of 47.05% by the second epoch, underscoring the rapid acquisition of generalizable representations. This early proficiency in learning shared across humans and diverse model architectures highlights the inherent potential for transfer learning in both biological and artificial systems.

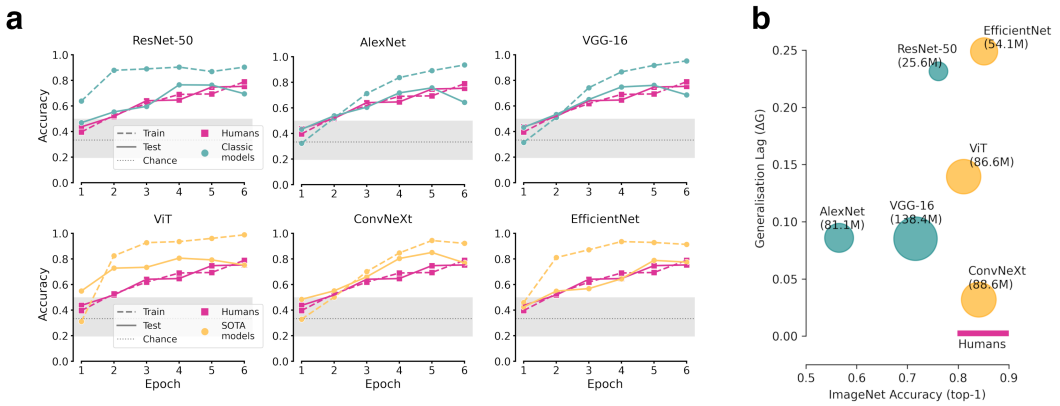

Figure 3: **Observed learning dynamics indicate that both, humans and DNNs, learn novel generalizable representations from limited amounts of training data to a degree that is comparably effective.** However, while humans immediately form generalizable representations, DNNs show a pronounced generalisation lag. Panel (**a**) shows training (dashed lines) and test (solid lines) performance trajectories in terms of classification accuracy for all classic CNNs (teal) and SOTA models (yellow). The performance of models is averaged across 20 fine-tuning runs (see Appendix D for individual runs). In each plot, model performance is contrasted to the mean performance of human observers (pink). Detailed information on the learning dynamics of individual participants can be found in Figure 2 (training) and in Appendix B (test). Shaded areas designate binomial confidence intervals (Clopper-Pearson) around chance performance for training sets—accuracies exceeding the upper bound suggest performance significantly above chance. In Panel (**b**) we plot Generalisation lag (ΔG; see Appendix D for details) as a function of ImageNet accuracy. The circle area for models reflects the number of parameters optimised during training. Top-1 ImageNet accuracy for humans is somewhat difficult to estimate from available studies since different measures are reported: 92.9–99% top-1 accuracy for entry-level categories (Dodge & Karam, 2017; Geirhos et al., 2018a), 84.9% top-5 accuracy on ImageNet-1k (Russakovsky et al., 2015), and up to 97.3% multilabel-accuracy (Shankar et al., 2020). Therefore, we designate human top-1 ImageNet accuracy with a rather conservative lower and upper bound of 80–90%.

Generalisation lag. When analyzing the learning trajectories of training versus test datasets within observers, distinct patterns emerge between humans and DNNs. Humans succeed in immediately generalizing from the training data to previously unseen images, as indicated by the alignment of their training and test performance trajectories. Models, however, reveal a marked discrepancy between training and test performance, with test performance consistently lagging behind. To quantitatively assess this generalisation lag (ΔG), we computed the mean difference between training and test accuracy, selectively focusing on epochs where learning occurred—excluding epochs where training performance did not significantly exceed chance level and epochs where models showed signs of overfitting. Table 1 summarizes generalisation lag for humans and different DNNs. Additional information on the calculation of ΔG is given in Appendix D.

Table 1: **Immediate generalisation in humans and generalisation lag in DNNs.** For models, color reflects group (teal: classic CNNs, yellow: SOTA models. The *Epoch* column specifies which epochs are considered when computing ΔG—the first integer indicates at which epoch training performance significantly exceeds chance level and the second until which epoch test performance increases. If the second integer does not equal 6, this signals overfitting.

Observer	Δ G	Epochs
Humans	0.002	2–6
ConvNeXt	0.032	3–5
VGG-16	0.085	2–5
AlexNet	0.086	2–5
ViT	0.139	2–4
ResNet-50	0.232	2–4
EfficientNet	0.249	2–6

Comparing the acquisition of representations in terms of ΔG suggests divergent learning dynamics between humans and DNNs. Humans demonstrate a near-instantaneous ability to generalize, evidenced by a minimal generalisation lag ($\Delta G = 0.002$). On the contrary, models display a substantial gener-

alisation lag (except from ConvNeXt; $\Delta G = 0.033$), illustrating distinct learning strategies between humans and DNNs. Subsequent examination was directed at the potential relationship between generalisation lag and variables such as ImageNet accuracy or model size, as detailed in Figure 3, Panel **b**. Intuitively, one might anticipate an inverse relationship where higher ImageNet accuracy corresponds to a lower generalisation lag. However, our findings revealed no such pattern, indicating that neither ImageNet accuracy nor model size serves as a reliable predictor of generalisation lag.

4 Discussion

In the present work we measured and compared the acquisition of novel representations in human observers and DNNs. To ensure the validity of our behavioral comparisons, we carefully controlled the learning environment and parallelized the learning conditions for both participants and models: Both groups were exposed to similar initial conditions and learned exclusively from static, labeled, unimodal input data and similar supervised learning signals. The tailored learning task required the recognition and categorization of images showing novel naturalistic 3D objects. Objects belonged to one of three classes, each demarcated by fuzzy, rather than distinct boundaries—mimicking real-world categories. Periodic evaluation phases during training examined the ability to transfer learned representations to previously unseen inputs. Our findings reveal that both humans and DNNs have the capacity to learn generalizable representations from a remarkably limited dataset—merely 36 images repeatedly presented in 6 epochs.

These findings challenge certain prevailing assumptions in the field. It has been frequently posited that human learning is exceptionally data-efficient, showcasing robust one- and few-shot learning capabilities (Lake et al., 2013; Lee & DiCarlo, 2023; Morgenstern et al., 2019), while DNNs, and CNNs in particular, are believed to be less efficient, requiring significantly more training data to reach human-level performance in object recognition tasks (Huber et al., 2023; Orhan, 2021). Contrary to these assertions, our results indicate that when learning conditions are equalized and within a supervised learning context, this disparity diminishes.

While our findings suggest that humans may not inherently possess an advantage in basic data efficiency, we identified notable representational differences, particularly in terms of generalisation lag. Our analysis revealed a two-phase learning pattern in models: an initial phase where specific features of the training data are learned, followed by a later phase where these representations are refined and abstracted for generalisation. In contrast, humans seem to bypass this staged process, directly forming generalizable representations from the onset of training, allowing for robustness against non-diagnostic features while retaining diagnostic attributes, and enabling recognition across identity-preserving transformations. Remarkably, the hurdle of generalisation lag persists not just within classic CNN frameworks but also in more recent architectures such as ViTs, suggesting a fundamental challenge in the alignment of generalisation behavior.

While shedding light on representational differences and learning dynamics, the presented work is accompanied by certain limitations. The scope of models analyzed represents only a small subset of available models and all employed models were pre-trained on ImageNet. Future studies might test and assess generalisation lag in models pre-trained on other, more human-like datasets such as SAYCam (Sullivan et al., 2021), or ecoset (Mehrer et al., 2021). Furthermore, also only a single dataset of novel objects was employed in the learning task. While we believe our stimuli choice should ensure external validity, it would be important to replicate our findings with other novel objects such Fribbles (Barry et al., 2014) or Greebles (Gauthier et al., 1998).

5 Conclusion

Synchronizing learning conditions and comparing the full learning trajectory revealed representational differences between humans and DNNs in terms of how quickly newly acquired knowledge can be generalised. We believe this insight to contribute to a deeper understanding of human cognition and learning, adding yet another desideratum to the list of properties artificial systems should possess to be behaviorally aligned with us: Increasing representational alignment requires more than enhancing data efficiency—instead our results indicate that we need to also focus on teaching DNNs to acquire immediately generalizable representations.

ACKNOWLEDGMENTS

The authors appreciate Nik Meili's valuable assistance in collecting data from human participants. Special thanks go to the members of the Wichmann Lab and the Brüders Retreat for insightful discussions and their constructive input.

Lukas S. Huber was funded by the Swiss National Science Foundation (project number: P000PS_214659). Felix A. Wichmann is a member of the Machine Learning Cluster of Excellence, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645.

REFERENCES

Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pp. 233–242. PMLR, 2017.

Tom J Barry, James W Griffith, Stephanie De Rossi, and Dirk Hermans. Meet the fribbles: novel stimuli for use within behavioural research. *Frontiers in Psychology*, 5:103, 2014.

Jeffrey S Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico Adolfi, John E Hummel, Rachel F Heaton, et al. Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 46:e385, 2023.

Mark J Brady. *Psychophysical investigations of incomplete forms and forms with background*. University of Minnesota, 1999.

Mark J Brady and Daniel Kersten. Bootstrapped learning of novel objects. *Journal of Vision*, 3(6):2–2, 2003.

James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

Samuel Dodge and Lina Karam. A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, pp. 1–7. IEEE, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR*, 2021.

Marin Dujmović, Gaurav Malhotra, and Jeffrey S Bowers. What do adversarial images tell us about human vision? *Elife*, 9:e55978, 2020.

Maria K Eckstein and Anne GE Collins. Computational evidence for hierarchically structured reinforcement learning in humans. *Proceedings of the National Academy of Sciences*, 117(47):29381–29389, 2020.

Chaz Firestone. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43):26562–26571, 2020.

Christina M Funke, Judy Borowski, Karolina Stosio, Wieland Brendel, Thomas SA Wallis, and Matthias Bethge. Five points to check when comparing visual perception in humans and machines. *Journal of Vision*, 21(3):16–16, 2021.

Isabel Gauthier, Pepper Williams, Michael J Tarr, and James Tanaka. Training 'greeble' experts: a framework for studying expert object recognition processes. *Vision Research*, 38(15-16):2401–2428, 1998.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2018a.

Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems*, 31, 2018b.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. On the surprising similarities between supervised and self-supervised models. In *NeurIPS 2020 Workshop SVRHM*, 2020.

Bryan R Gibson, Timothy T Rogers, and Xiaojin Zhu. Human semi-supervised learning. *Topics in Cognitive Science*, 5(1):132–172, 2013.

Karin Hauffen, Eugene Bart, Mark Brady, Daniel Kersten, and Jay Hegdé. Creating objects and object categories for studying perception and perceptual learning. *JoVE (Journal of Visualized Experiments)*, (69):e3358, 2012.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Lukas S Huber, Robert Geirhos, and Felix A Wichmann. The developmental trajectory of object recognition robustness: Children are like small adults but unlike big deep neural networks. *Journal of Vision*, 23(7):4–4, 2023.

Talia Konkle and George A Alvarez. A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, 13(1):491, 2022.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

Jonas Kubilius, Stefania Bracci, and Hans P Op de Beeck. Deep neural networks as a computational model for human shape sensitivity. *PLoS Computational Biology*, 12(4):e1004896, 2016.

Brenden M Lake, Russ R Salakhutdinov, and Josh Tenenbaum. One-shot learning by inverting a compositional causal process. *Advances in Neural Information Processing Systems*, 26, 2013.

Michael J Lee and James J DiCarlo. How well do rudimentary plasticity rules predict adult visual object learning? *PLOS Computational Biology*, 19(12):e1011713, 2023.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

William Lotter, Gabriel Kreiman, and David Cox. A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nature Machine Intelligence*, 2(4):210–219, 2020.

Johannes Mehrer, Courtney J Spoerer, Emer C Jones, Nikolaus Kriegeskorte, and Tim C Kietzmann. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8):e2011417118, 2021.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Yaniv Morgenstern, Filipp Schmidt, and Roland W Fleming. One-shot categorization of novel object classes in humans. *Vision Research*, 165:98–108, 2019.

Lukas Muttenthaler, Lorenz Linhardt, Jonas Dippel, Robert A Vandermeulen, and Simon Kornblith. Human alignment of neural network representations. In *SVRHM 2022 Workshop@ NeurIPS*, 2022.

Emin A Orhan. How much human-like visual experience do current self-supervised learning algorithms need in order to achieve human-level object recognition? *arXiv preprint arXiv:2109.11523*, 2021.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2009.

Roland Pusch, Julian Packheiser, Charlotte Koenen, Fabrizio Iovine, and Onur Güntürkün. Digital embryos: a novel technical approach to investigate perceptual categorization in pigeons (columba livia) using machine learning. *Animal Cognition*, 25(4):793–805, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning*, pp. 8634–8644. PMLR, 2020.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Katherine R Storrs, Barton L Anderson, and Roland W Fleming. Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour*, 5(10):1402–1417, 2021.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Jascha Achterberg, Joshua B Tenenbaum, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

Jessica Sullivan, Michelle Mei, Andrew Perfors, Erica Wojcik, and Michael C Frank. Saycam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open Mind*, 5:20–29, 2021.

Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, pp. 10096–10106. PMLR, 2021.

Lore Thaler, Alexander C Schütz, Melvyn A Goodale, and Karl R Gegenfurtner. What is the best fixation target? the effect of target shape on stability of fixational eye movements. *Vision Research*, 76:31–42, 2013.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Felix A Wichmann and Robert Geirhos. Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science*, 9, 2023.

Ludwig Wittgenstein. *Philosophical investigations. Wittgenstein*. Macmillan, 1968.

Lorijn Zaadnoordijk, Tarek R Besold, and Rhodri Cusack. Lessons from infant learning for unsupervised machine learning. *Nature Machine Intelligence*, 4(6):510–520, 2022.

Zhenglong Zhou and Chaz Firestone. Humans can decipher adversarial images. *Nature Communications*, 10(1):1334, 2019.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.

A Appendix: Details on the generation of novel objects

All novel 3D objects, used as a basis for the stimuli in this study, were created with the `Digital Embryo Workshop (DEW)` written by Mark Brady and Dan Gu. Some details on the employed algorithms can be found in Mark Brady's PhD Thesis (Brady, 1999, Appendix A, p.211). Table 2 shows the parameter settings for the generation of the parents and the second generation. Note that the Hydro PCD and Hydro Interaction PCD parameters were always kept at zero and for each batch of objects we used a random seed.

Table 2: Parameter settings for 3D object generation with DEW. "PCD" is short for "programmed cell death".

Parameter	**Parent Gen.**	**Second Gen.**
Threshold Growth	6	3
Interaction Growth	6	0
Shrinkage PCD	4	0
Shrinkage PCD Interaction	6	2

B Appendix: Test data of individual observers

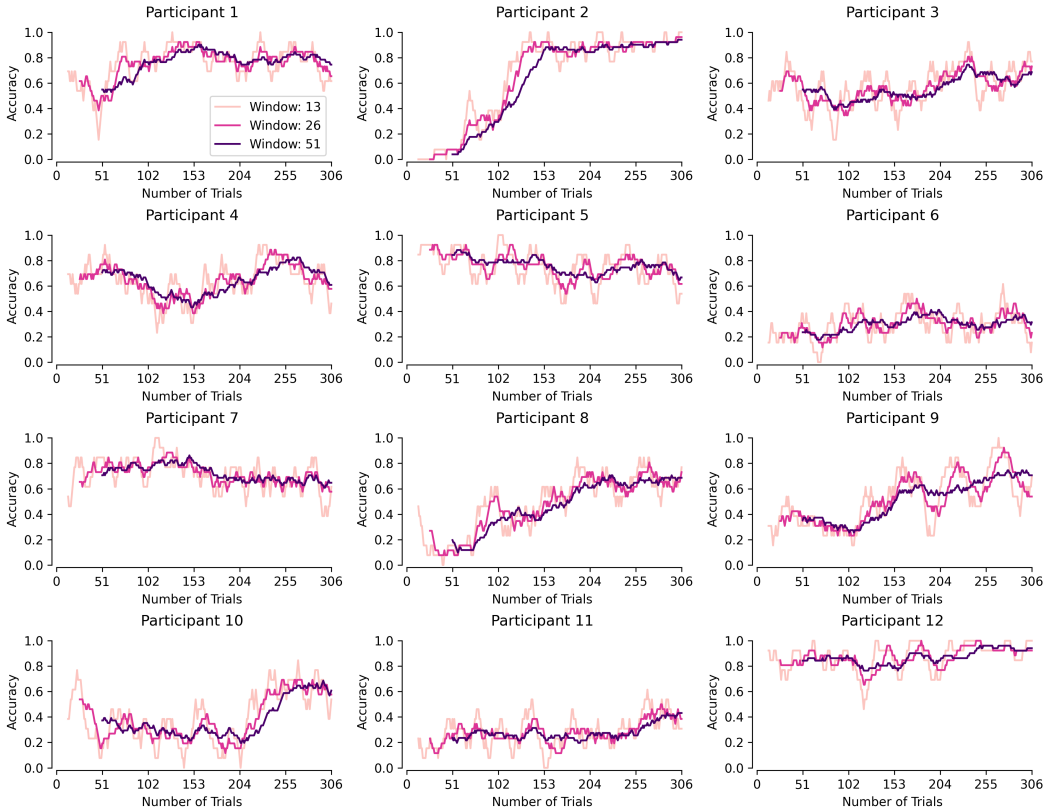

Figure 4: **Inter-individual differences in human test performance**. For each participant we show the evolution of test performance across the entire training period—depicted by moving averages of varying window sizes (different shades of pink). The largest window size of 51 equals the length of a test set.

C APPENDIX: A CLOSER LOOK AT TEST PERFORMANCE

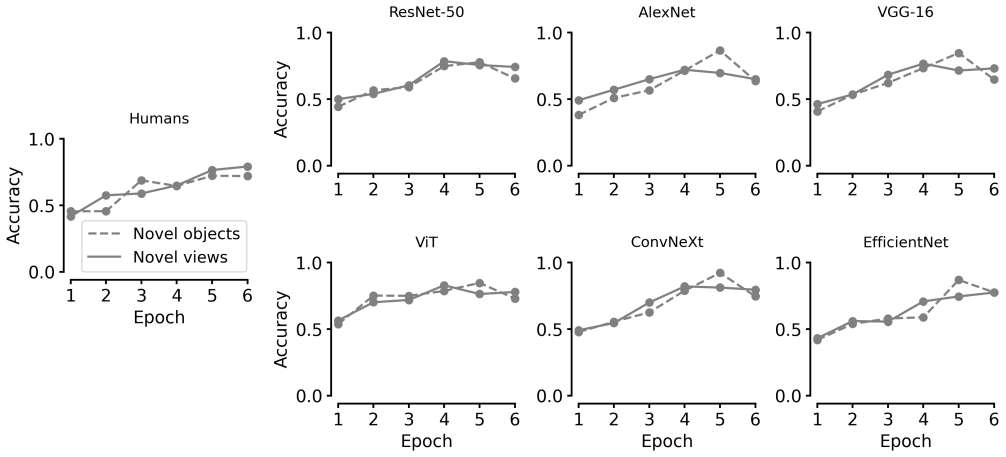

Figure 5: **Neither humans nor any model demonstrated a consistent advantage for either novel perspectives or novel objects.** All images were previously unseen; novel perspectives are viewpoints of training objects not encountered during training, while novel objects are entirely new and also not encountered during training. In each plot, test performance is divided into accuracy for novel perspectives (solid lines) and accuracy for novel objects (dashed lines).

D APPENDIX: DETAILS ON MODELS AND FINE-TUNING RUNS

All ImageNet pre-trained weights were obtained through the `PyTorch` model zoo (except the ViT which was obtained through the `Hugging Face` model library), where they are referred to as: "ResNet50_Weights.IMAGENET1K_V1" (ResNet-50), "AlexNet_Weights.IMAGENET1K_V1" (AlexNet), "VGG16_Weights.IMAGENET1K_V1" (VGG-16), "google/vit-base-patch16-224" (ViT), "ConvNeXt_Base_Weights.IMAGENET1K_V1" (ConvNeXt), and "EfficientNet_V2_M_Weights.IMAGENET1K_V1" (EfficientNet). Figure 6 shows the results of the fine-tuning of those pre-trained model on our stimuli set.

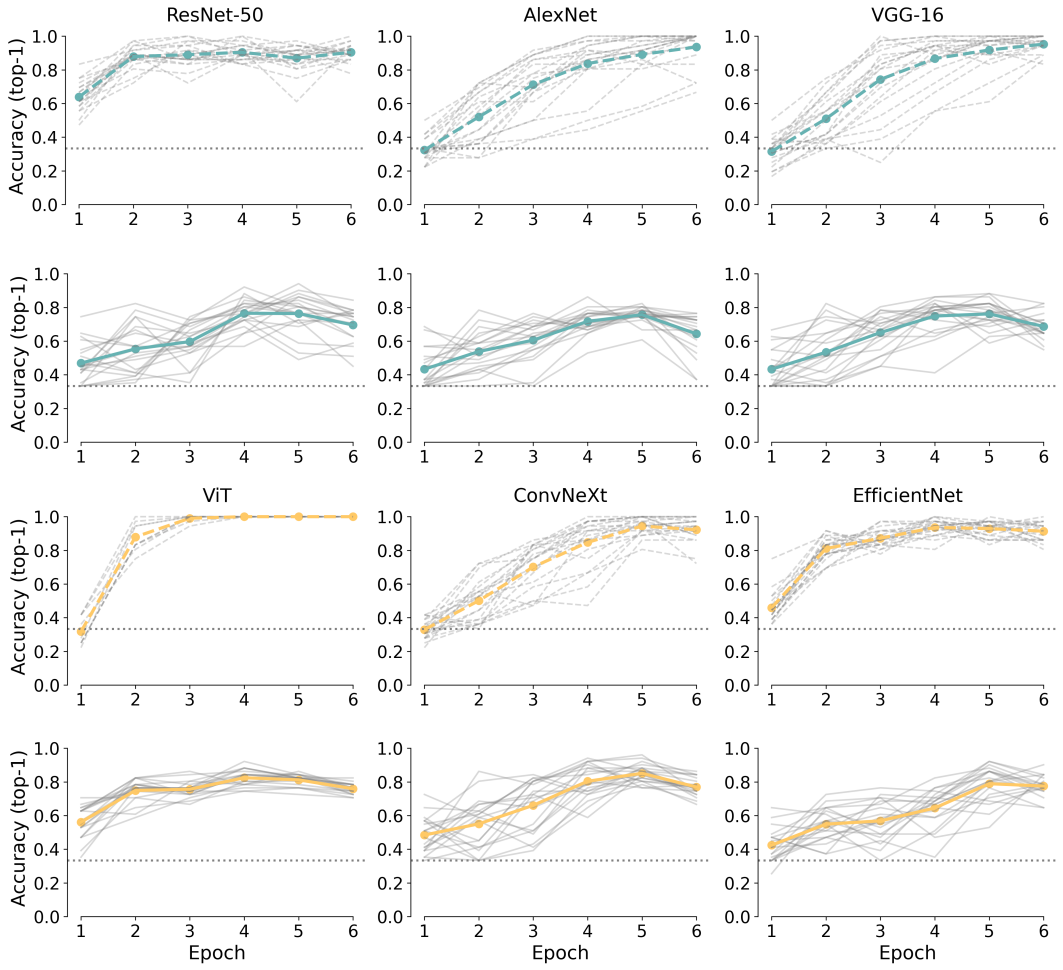

Figure 6: **Overview of all ten fine-tuning runs for all classic CNNs (teal) and SOTA models (yellow).** Classification accuracy is plotted as a function of learning epochs for training (dashed lines) and test (solid lines) trajectories. Subtle grey lines indicate individual fine-tuning runs and colored trajectories inidcates average performance as reported is Figure E, Panel (**a**). The large variation across fine-tuning runs seems not surprising given the small size of the training dataset (36 images) and might be attributed to the stochastic nature of the Adam optimizer, leading to different convergence paths and solutions.

E APPENDIX: CALCULATION OF GENERALISATION LAG

The generalisation lag is given by

$$\Delta G = \frac{1}{|E|} \sum_{i \in E} (acc_{train,i} - acc_{test,i})$$

whereby $|E|$ denotes the number of epochs included, $acc_{train,i}$ represents the training accuracy at the ith epoch, and $acc_{test,i}$ is the test accuracy at the ith epoch. The set E includes epochs where $acc_{train,i}$ is significantly above chance and $acc_{test,i}$ has not started to decline, ensuring that only relevant epochs contribute to the ΔG calculation, thereby providing a standardized measure of generalisation lag.