Evaluating alignment between humans and neural network representations in image-based learning tasks

Can Demircan[1,2,*] **Tankred Saanum**[2] **Leonardo Pettini**[3,4] **Marcel Binz**[1,2]
Blazej M Baczkowski[4,5] **Christian F Doeller**[4,6,7,8] **Mona M Garvert**[4,9,10] **Eric Schulz**[1,2]

[1]Institute for Human-Centered AI, Helmholtz Computational Health Center, Munich, Germany
[2]Max Planck Institute for Biological Cybernetics - Tübingen, Germany
[3]Max Planck School of Cognition - Leipzig, Germany
[4]Max Planck Institute for Human Cognitive & Brain Sciences - Leipzig, Germany
[5]University of Tübingen - Tübingen, Germany
[6]Kavli Institute for Systems Neuroscience - Trondheim, Norway
[7]Leipzig University - Leipzig, Germany
[8]Technical University Dresden - Dresden, Germany
[9]Julius-Maximilians-Universität Würzburg - Würzburg, Germany
[10]Max Planck Institute for Human Development - Berlin, Germany
[*]{can.demircan@helmholtz-munich.de}

Abstract

Humans represent scenes and objects in rich feature spaces, carrying information that allows us to generalise about category memberships and abstract functions with few examples. What determines whether a neural network model generalises like a human? We tested how well the representations of 86 pretrained neural network models mapped to human learning trajectories across two tasks where humans had to learn continuous relationships and categories of natural images. In these tasks, both human participants and neural networks successfully identified the relevant stimulus features within a few trials, demonstrating effective generalisation. We found that while training dataset size was a core determinant of alignment with human choices, contrastive training with multi-modal data (text and imagery) was a common feature of currently publicly available models that predicted human generalisation. Intrinsic dimensionality of representations had different effects on alignment for different model types. Lastly, we tested three sets of human-aligned representations and found no consistent improvements in predictive accuracy compared to the baselines. In conclusion, pretrained neural networks can serve to extract representations for cognitive models, as they appear to capture some fundamental aspects of cognition that are transferable across tasks. Both our paradigms and modelling approach offer a novel way to quantify alignment between neural networks and humans and extend cognitive science into more naturalistic domains.

1 Introduction

Research on representational alignment between neural networks and humans has gained significant attention in recent years [1, 2]. Comparisons across the systems have provided important insights into neural network representations [3, 4], human cognition and the brain [5, 6, 7, 8, 9], and the development of more robust machine learning systems [10, 11, 12]. In the sensory domain, the comparisons have been predominantly made through two families of behavioural tasks. One common approach is to compare object recognition performance across humans and neural networks [13]. This

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-3890

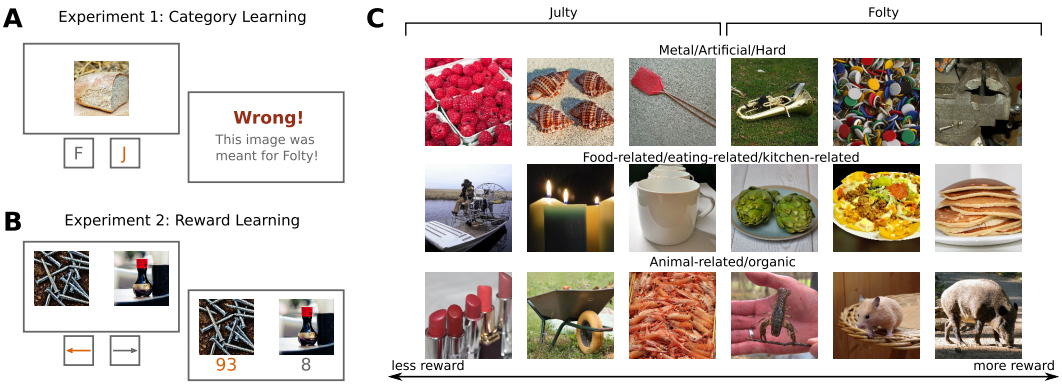

Figure 1: Task descriptions. (*A*) An example trial from the category learning task, where an incorrect decision is made. (*B*) An example trial from the reward learning task where the best option is chosen and highlighted in orange. (*C*) Example images from the THINGS database [30]. The database has a low dimensional semantically interpretable embedding [27], which is derived from human similarity judgements. The example images are placed in the most three prominent dimensions of this embedding. In both tasks, participants were randomly assigned to one of these three dimensions. The associated category membership and rewards for the two tasks are displayed.

is a fruitful approach for understanding if the two systems use the same features for object recognition [14, 15, 16], are susceptible to similar distortions [17, 18, 19, 20], and struggle with similar images [21]. Another common approach is to use similarity judgement tasks, which may entail reporting pairwise similarity scores [22, 23, 24], arranging stimuli in a 2D space based on their similarity [25, 26], or choosing the odd-one-out in triplets of stimuli [27, 28]. Using these tasks, previous work has identified the factors that contribute to neural networks representing stimuli similarly to humans, both in low-level perception [29] and semantic judgements [3].

However, similarity judgements do not begin to capture the complexity of tasks humans use their representations for. Humans rely on rich representations for making judgements and acting in the world. For example, an apple has a multitude of features, such as colour, taste, shape, and brand. Depending on the context, people can use these features and make predictions about the apple's taste, the environmental impacts of growing it, or the significance of it in different mythological and religious settings. What determines whether a neural network model represents an object like an apple with the same richness and flexibility?

In this work, we investigated people's ability to learn functional relationships on naturalistic images in a few-shot setting, and what neural network models best predict human choices. We adapted two commonly used learning paradigms from the cognitive psychology literature: category learning (Fig. 1A) and reward learning (Fig. 1B). However, instead of using repeating artificial stimuli, we presented human participants with unique naturalistic images sampled from the THINGS database [30] in each trial, requiring them to continuously generalise. To understand whether neural networks contain sufficiently rich representations that allow for such generalisation, we tested 86 different neural networks [27, 31, 32, 33, 34, 35, 11, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 10, 12, 52, 53, 54, 55]. These networks varied in their loss function, training diet, and the modality of training data. In summary, we found that:

- While almost all pretrained models generalised above chance level and predicted human behaviour in both naturalistic learning tasks, contrastive language image pretraining (CLIP) [45] consistently yielded the best predictions of human behaviour. We furthermore showed that this could not be fully attributed to the training diet alone.
- Multiple factors were important for human alignment, including task performance, model size, training diet, separation of different classes in representations, and the similarity of the representations to the generative embedding of the task.
- Of the tested human-aligned neural networks, no method consistently improved human alignment in our tasks compared to non-aligned baselines. However, two of the methods (Harmonization [11] and gLocal [10]) yielded improvements in task accuracy on average.

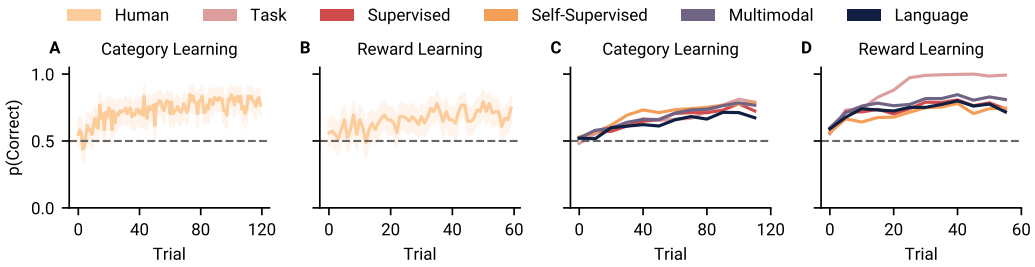

Figure 2: Learning trajectories of human participants and neural networks. Neural networks can perform as well as humans. (*A* & *B*) Accuracy of human participants across trials for the category and the reward learning tasks respectively. Shaded lines indicate 95% confidence intervals. (*C* & *D*) Example learning curves for the neural network representations in the category and the reward learning tasks respectively. The best-performing models from each model type are shown.

2 Experiments

We design our experiments around naturalistic images from the THINGS database [30, 27]. Each image in the database depicts a collection of entities (animals or objects) and comes with an embedding with 49 human interpretable features, which was built by Hebart et al. [27] to predict human similarity judgements of these objects. Each feature reflects a semantically meaningful property such as whether an image contains metallic objects, food, animals etc. In our experiments, humans learned functions defined over these individual embedding dimensions. We chose category learning and reward learning experiments, as they are well-established paradigms to test function learning and generalisation in human participants. However, unlike traditional paradigms, we used naturalistic images and no images were repeated, requiring generalisation.

Category learning: Human participants ($n = 91$) completed 120 trials of an online category learning task, where they were presented with a novel image in each trial. They were asked to deliver these images to one of two dinosaurs, *Julty* or *Folty*, using key presses. Participants were told that the two dinosaurs had completely non-overlapping preferences for what gifts they enjoyed. After each trial, we gave participants feedback on whether their choice of delivery was correct. An example trial from the task is shown in Fig. 1*A*. Participants were assigned to one of three conditions, where in each condition the category boundary was defined over a different THINGS embedding dimension. The three chosen dimensions map to how metallic, food-related, and animal-related the shown image is (Fig. 1*C*). For instance, in one condition non-metallic images should be classified to Folty, and metallic images to Jolty. For each participant, 120 unique stimuli from the THINGS database were sampled. A median split over the assigned feature of the sampled stimuli determined the category boundary.

Reward learning: Human participants ($n = 82$) completed 60 trials of a reward learning paradigm [56], in which they were asked to maximise their accumulated reward throughout the task. In each trial, participants were presented with two images and were asked to select one using key presses. After making a choice, the associated reward with each option was shown. An example trial from the task is shown in Fig. 1*B*. Participants were assigned to one of the three conditions, as was done in the category learning task. Stimuli were sampled in the same way as the category learning task. For each participant, the values of the task-relevant feature were re-scaled linearly between 0 and 100. Additional details about the experimental paradigms are described in Appendix A.

3 Behavioural analyses

Humans learn to generalise quickly. The learning curves of the participants are shown in Fig. 2A and 2B. To measure whether and how fast people learned in the two experiments, we analysed their choice data using mixed-effects logistic regression models. In the category learning task, we predicted whether a participant made the correct choice using an intercept and the trial number. We found that participants performed this task above chance level, as indicated by a significant

https://doi.org/10.52202/079017-3890

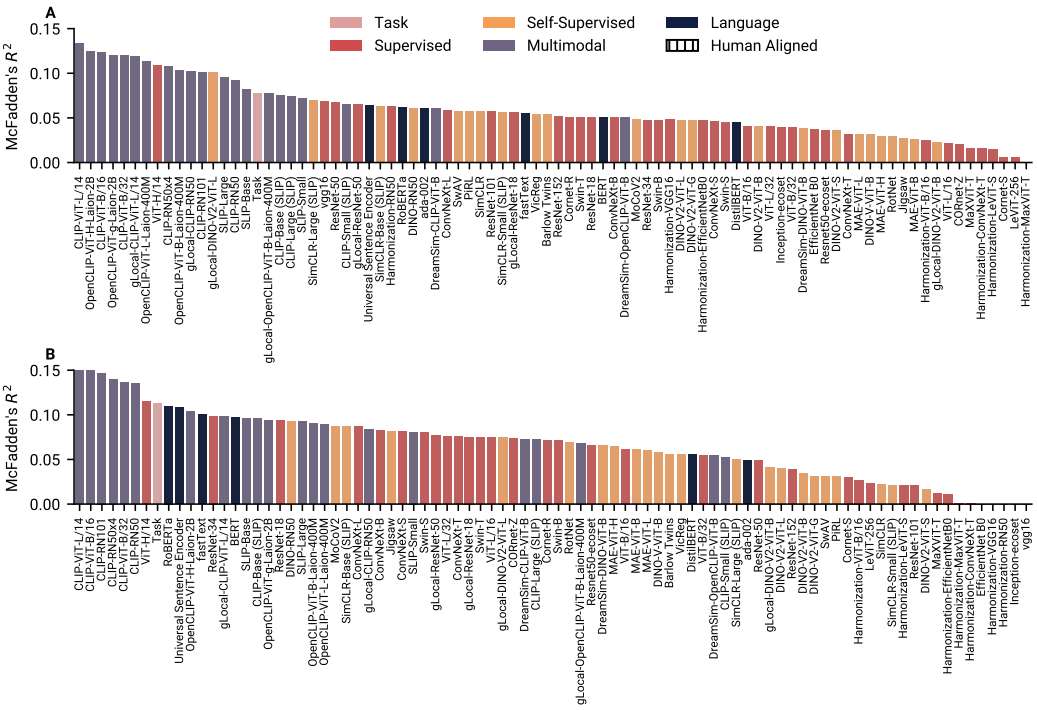

Figure 3: Model fits to human choice data. In both category learning (*A*) and reward learning tasks (*B*), several CLIP models predict human choices the best, even better than the generative features of the tasks. How well the models fitted human choice was more heterogeneously distributed for supervised, self-supervised, and language models. Plotted are the cross-validated McFadden's R^2 of each representation for the category learning and the reward learning tasks respectively. Higher values indicate better fits to human behaviour. 0 marks the alignment of a random model.

intercept ($\hat{\beta} = 1.14 \pm 0.09$, $z = 13.18$, $P < .001$), and that their performance improved over trials ($\hat{\beta} = 0.32 \pm 0.05$, $z = 6.89$, $P < .001$), indicating a learning effect. This suggests that people can very efficiently extract the relevant feature dimension in high-dimensional naturalistic environments despite seeing each stimulus only once. For the reward learning task, we predicted whether a participant chose the image on the right using the reward difference between the two images, the trial number, and the interaction of the two predictors. We found that the reward difference ($\hat{\beta} = 0.89 \pm 0.07$, $z = 12.56$, $P < .001$), and the interaction of this difference with the trial number ($\hat{\beta} = 0.34 \pm 0.04$, $z = 9.30$, $P < .001$) predicted choice, again indicating a learning effect. We further characterise how quickly humans learn the task in Fig. 10 in Appendix C and provide the full specification of the mixed-effects models in Appendix A.

4 Model-based analyses

To understand what kind of representations are needed to predict human choices, we tested representations extracted from several pretrained neural networks on our tasks.

Most representations predict human choice above chance level. CLIP makes the best predictions. The representations were extracted from the penultimate layer if the models had a classification layer, and from the final layer otherwise. For the transformer models, the `[CLS]` token representations were extracted. To extract representations from language models, we provided them with the prompt `A photo of` X where X was the category label of the task image. fastText was only provided with the category label instead.

We trained linear models to predict either reward or category membership from each neural network model's extracted representations. The models were provided with image-target pairs until trial

$t-1$ as training data and made predictions for the image on trial t. For the category learning task, we used an ℓ_2 regularised logistic regression model, and for the reward learning task, we used a Bayesian linear regression model with spherical Gaussian priors. We used the estimates from the linear models to predict participant choice using mixed-effects logistic regression in leave-one-trial-out cross-validation. For the category learning task, we regressed the probability estimates of the logistic regression models onto participant choice. For the reward learning task, we regressed the reward estimate differences between the left and the right options onto choice. Example learning curves for the two tasks are shown in Fig. 2C and Fig. 2D. Finally, we measure alignment to human choices using McFadden's R^2 [57], which is computed as follows:

$$\text{McFadden's } R^2 = 1 - \frac{\mathcal{L}_{\text{Model}}}{\mathcal{L}_{\text{Random}}} \tag{1}$$

where $\mathcal{L}_{\text{Model}}$ is the negative log likelihood of a given model and $\mathcal{L}_{\text{Random}}$ is the negative log likelihood of a random model.

We observed most of the representations we tested can do our task and predict human behaviour above chance level across the two tasks (as visualized in Fig. 3A and Fig. 3B). CLIP models were the top 7 (6) models for the category (reward) learning task in predicting human choices. In total, 16 (7) of the 86 candidate representations predicted participant behaviour better than the ground truth representations that were used to generate the task. Of these 16 (7) representations, 14 (6) were CLIP models. One was a large vision transformer, trained in a supervised manner on ImageNet [58]. A human-aligned variant of DINO-v2 provided a better fit than the generative task representations in the category learning task. The rest of the supervised and self-supervised vision models, as well as the language models, had a heterogeneous distribution in how well they predicted human behaviour. To provide better intuition for how human participants and CLIP were similar, we display example trials where both CLIP and humans make the same incorrect decisions in Fig. 11 in Appendix C.

Which factors contribute to alignment?

Why are CLIP models substantially better aligned with humans in our task? We conducted a series of analyses to better understand which model properties contribute to alignment. We pooled the data across the two tasks and excluded the language models from all analyses except those shown in Fig. 4A and Fig. 4E, as comparing other properties across vision and language models (e.g. model size) is not meaningful. We first tested if larger models predicted human choice better. While it is common for more expressive models to perform better at downstream computer vision tasks [59, 60, 61], previous work has shown that this is not a robust predictor of human alignment [3, 9]. In our tasks, we found that larger models predicted human choices better ($\rho = 0.48$, $p < .001$, Fig. 4B), which contradicts previous findings. Next, we considered the number of images seen during training, which is predictive of higher accuracy in image recognition [62] and human alignment. In our tasks, we found that models trained on more images were more predictive of human choices as well ($\rho = 0.52$, $p < .001$, Fig. 4C).

Then, we analysed which, if any, properties of the models' representations were predictive of their alignment with human choices. First, we considered how well the THINGS classes were separated in the representations of each model. Following Kornblith et al. [63], the class-separation was computed as follows:

$$R^2 = 1 - \bar{d}_{\text{within}}/\bar{d}_{\text{total}} \tag{2}$$

$$\bar{d}_{\text{within}} = \sum_{k=1}^{K}\sum_{m=1}^{N_k}\sum_{n=1}^{N_k}\frac{1-\cos(\mathbf{x}_{k,m},\mathbf{x}_{k,n})}{KN_k^2} \quad \bar{d}_{\text{total}} = \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{m=1}^{N_j}\sum_{n=1}^{N_k}\frac{1-\cos(\mathbf{x}_{j,m},\mathbf{x}_{k,n})}{K^2N_jN_k} \tag{3}$$

where $\mathbf{x}_{k,m}$ is the representation of image m in object class k. K is the total number of classes, and N_k is the total number of images in class k. $\cos(\cdot,\cdot)$ denotes cosine similarity between representations. The R^2 measure is between 0 and 1, where higher scores indicate a low within-class distance to across-class distance ratio, i.e. high class separation. Previous work has shown a positive link between class separation and image classification [63], as well as recall [64]. Similar to these findings, we found that models that had higher class separation were more predictive of human choices ($\rho = 0.29$, $p = .01$, Fig. 4D).

https://doi.org/10.52202/079017-3890

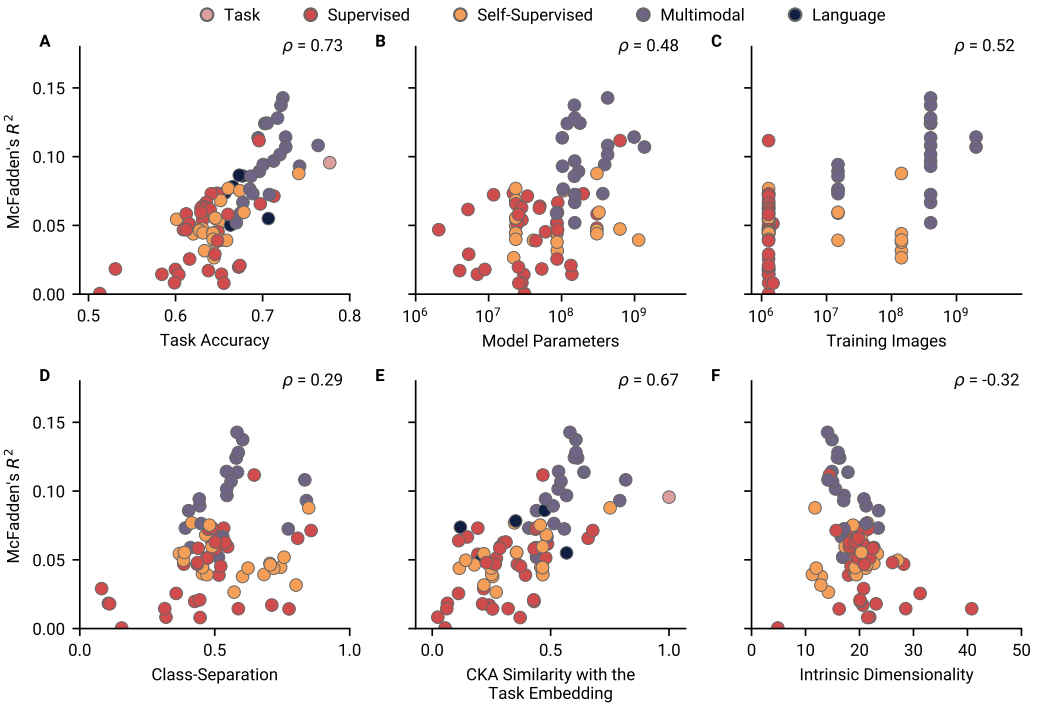

Figure 4: Several factors contribute to alignment. Models trained on more data and with more trainable parameters predict human choices with higher accuracy. Turning to representations, those that better separate image classes and are more similar to the generative task features exhibit stronger alignment with human choices.

We then considered whether the similarity of the representations with the generative task features was predictive of how well different representations predicted human choices. For this, we used linear Centered Kernel Alignment (CKA) [65], which computes the similarity between the generative task representations $\mathbf{T}$ and neural network representations $\mathbf{X}$ as follows:

$$\mathrm{CKA}(\mathbf{T}, \mathbf{X}) = \frac{||\mathbf{X}^T\mathbf{T}||_F^2}{||\mathbf{T}^T\mathbf{T}||_F||\mathbf{X}^T\mathbf{X}||_F} \tag{4}$$

where $||\cdot||_F$ denotes the Frobenius norm. We found that representations that were more similar to the generative task embedding predicted human choices better ($\rho = 0.67$, $p < .001$, Fig. 4E).

Lastly, we tested whether the intrinsic dimensionality of representations was related to alignment. Lower intrinsic dimensionality of neural networks in late layers is positively linked to better classification performance [66]. The degree to which a network *compresses* its inputs is also directly linked to its ability to generalize [67, 68, 69, 70, 71]. In the human alignment literature, a similar measure named expressed dimensionality has been studied in the context of neural representations.However, diverging from Ansuini et al. [66], one study found a negative correlation between alignment and this measure [72], and another study found no link [9]. We used the TwoNN method proposed by Facco et al. [73] to estimate intrinsic dimensionality, which makes use of the nearest neighbour distances. First, we linearly scaled all the features to be between 0 and 1. We then computed pairwise distances

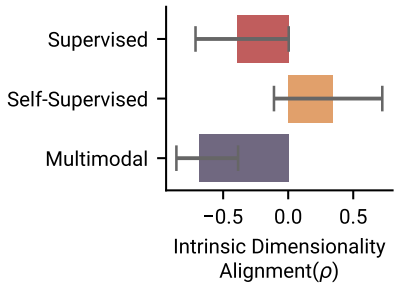

Figure 5: Lower intrinsic dimensionality is linked with higher alignment most strongly for the multimodal models, and to a lesser extent with supervised ones.

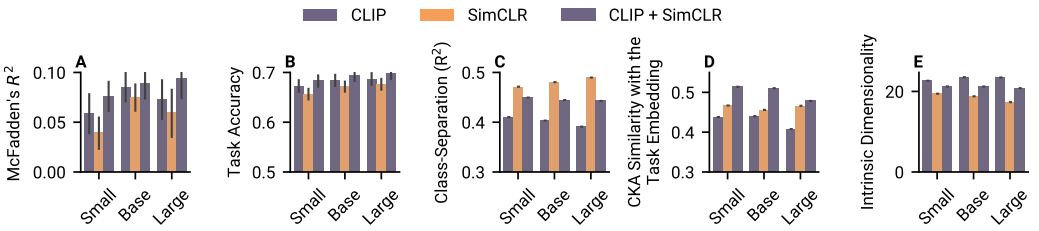

Figure 6: The effect of CLIP loss while controlling for model size and data. We observed that CLIP loss increases alignment when data size and architecture are controlled. Here plotted are (*A*) McFadden's R^2, (*B*) task accuracy, (*C*) class-separation, (*D*) similarity with the task embedding, and (*E*) intrinsic dimensionality across model sizes and loss functions.

for each pair of data points. Then, we calculated $\mu_i = {}^{r_2}/_{r_1}$ where r_1 and r_2 are the shortest distances from datapoint i. Later, the empirical cumulative distribution $F^{emp}(\mu)$ was computed by sorting μ_i and normalising by the total data points N. The slope of a linear model that maps $\log \mu_i$ to $-\log 1 - F^{emp}(\mu_i)$ with no intercept gives the intrinsic dimensionality measure.

Pooling over all model types, lower intrinsic dimensionality was significantly associated with alignment ($\rho = -0.32$, $p = .03$, Fig. 4F). However, we found that this relationship was most strongly driven by the multimodal models and to a lesser extent by supervised models (Fig. 5). That input compression and dataset size are positively related to alignment most strongly for CLIP models suggests that the contrastive multimodal training regime unlocks desirable scaling properties in these models. See Fig. 12 in Appendix C for pairwise correlations between the investigated factors.

Are CLIP models well aligned only due to their high data diet?

While we found that models trained with contrastive language image loss predicted human behaviour the best, there remains an important confound. These models are also the ones that are trained on the largest datasets (400M to 2B images). Therefore the direct benefits of multimodal training remain unclear. To address this point, we turned to models provided by Mu et al. [52]. Here, the same models are trained on a large dataset (YFCC15M [74, 45]) using three different losses: i) a CLIP loss that penalises for the distance between corresponding pairs of text and image representations ii) a SimCLR [39] loss that pushes the representations of the augmented and the original image close to each other and away from others, and iii) a CLIP + SimCLR loss.

First, we found that CLIP models always fit human data better than SimCLR models, and CLIP + SimCLR models made the best predictions when controlling for model size (Fig. 6A). This suggests that the advantage provided by the CLIP models cannot solely be attributed to the training data. We found the same ranking of models in terms of how well they did the tasks (Fig. 6B). Yet, contrary to our expectations, the SimCLR models had a higher class separation than CLIP models (Fig. 6C), as well as better alignment with the generative task features (Fig. 6D), and lower intrinsic dimensionality (Fig. 6E). This was surprising because, in our previous analyses, we found these properties to be associated with models that predicted human choice better. However, there still may be other confounds that impacted the findings. For example, controlling for training data is not straightforward, as text-image pairs may carry more information than augmented versions of the same image, providing an unfair advantage to the multimodal models.

Do alignment methods transfer to our learning tasks?

Lastly, we evaluated the performance of models that were explicitly aligned to be more human-like. This comparison included three sets of models. Fel et al. [11] have aligned models through a method called Harmonization. In addition to the standard supervised training, the models are trained to use the same visual features of images that humans use. The second part is achieved by aligning the networks' saliency maps with feature importance maps obtained from human judgment. This results in networks that perform better in ImageNet and that are aligned with humans. Next, Fu et al. [12] have curated human similarity judgements on a carefully created synthetic dataset. They later fine-tuned pretrained models such as CLIP using Low-Rank Adaptation [75] to derive a metric named DreamSim that outperforms other models in predicting human similarity judgements. Lastly, Muttenthaler et al.

https://doi.org/10.52202/079017-3890

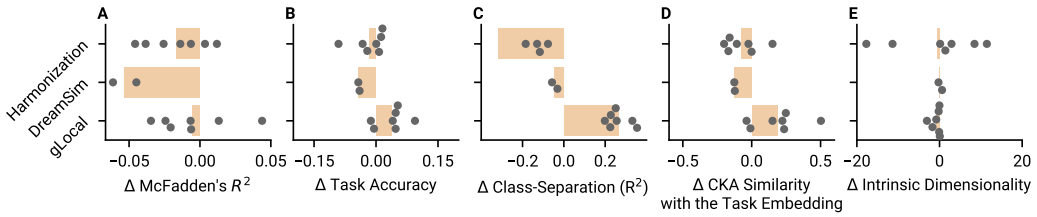

Figure 7: We compared models aligned to humans through three different methods against baselines that had the same architecture and that were pretrained on the same data. We found no consistent improvement in human alignment. Here plotted are (*A*) McFadden's R^2, (*B*) task accuracy, (*C*) class-separation, (*D*) similarity with the task embedding, and (*D*) intrinsic dimensionality across model sizes and loss functions.

[10] have fine-tuned representations of pretrained models through a novel transformation named gLocal, which aligns the global representational space to be more human-like by trying to predict human similarity judgements, while preserving the local structure through a contrastive loss that encourages the representations to stay close to their original positions. For these comparisons, we used the models openly provided by the authors.

First, we found that none of the alignment methods improved alignment in our task consistently, with some instances of Harmonised and gLocal models improving alignment (Fig. 7A). Alignment improved task accuracy on average for Harmonised and gLocal models (Fig. 7B). Class separation was lower for all Harmonization and DreamSim models, whereas it increased for all gLocal models tested (Fig. 7C). We also observed that the similarity between the representations and the task embedding decreased after alignment for Harmonization and DreamSim, but it increased in most of the models after gLocal alignment (Fig. 7D). Lastly, we observed heterogeneous patterns in the change of intrinsic dimensionality across the three alignment methods, with gLocal reducing the intrinsic dimensionality for all but one of the tested models (Fig. 7E).

How do our tasks compare to other alignment measures?

Lastly, to better characterise how our cognitive tasks fit in the alignment literature, we compared them to previously established measures (Fig. 8). We found the strongest correlation with the THINGS odd-one-out judgements [3, 27] ($\rho = 0.54$ for zero-shot, and $\rho = 0.61$ for probing). Given the two tasks use the same images, and the ground truth of our tasks was constructed from the odd-one-out judgements, this strong relationship is expected. However, the correlations are still moderate, indicating important differences across tasks. Comparisons with an independent similarity judgement [23] task showed a weaker correlation ($\rho = 0.35$), and we found no correlation with a fine-grained two-alternative forced choice task [12]. Lastly, we compared alignment in our task to alignment on the ClickMe dataset [76], which was used to build the Harmonization models [11]. We observed a negative correlation ($\rho = -0.48$) here, suggesting that pixel-level alignment and semantically bound global image alignment might be at odds.

5 Discussion

In this work, we investigated the alignment of neural network representations to humans. To study this, we measured how well different neural network representations predict human choices in two newly developed learning tasks. Of the 86 tested representations, all but one predicted human choice above chance level. We furthermore identified several important factors for human alignment, such as large model size, training regime, and low intrinsic dimensionality. These results expand on previous work in both human alignment and cognitive modelling. From an alignment perspective, we considered more challenging tasks compared to previous studies. Previous work has predominantly focused on simple image exposure and similarity judgments. We believe our findings complement this research by addressing unexplored aspects of alignment, which are generalisation and information integration across an extended horizon. From a cognitive modelling perspective, we demonstrated that off-the-shelf pretrained neural networks can serve as representations for cognitive models [77], which allows to push cognitive models into more naturalistic domains.

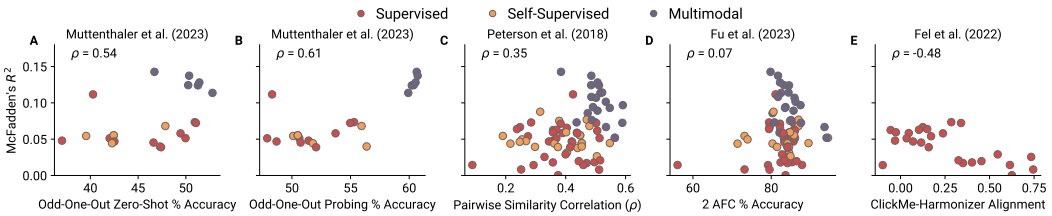

Figure 8: How do our tasks compare to other alignment methods? Our tasks offer similar (but not identical) results with two of the four similarity judgement tasks. There is a strong negative relationship with the ClickMe dataset, which focuses on localised pixel-level alignment.

5.1 Related work

How do our findings compare to previous work on alignment? First, previous research has shown that CLIP representations align well with human representations using brain imaging [78, 9, 79] and similarity judgments [3]. In line with this, we also found that the contrastive language-image loss improved alignment when controlling for data size and architecture. We furthermore found that training on large datasets generally improved alignment. Yet, it remains unclear whether supervised training on massive datasets alone can achieve high alignment similar to that of CLIP models. For example, Muttenthaler et al. [3] showed that models trained in a supervised manner on the JFT-3B dataset [61] can outperform CLIP in predicting human similarity judgments. However, since this is a proprietary dataset, we could not make this comparison.

We observed some findings that diverge from previous studies using different experimental paradigms. Muttenthaler et al. [3] and Conwell et al. [9] found no consistent correlation between a model's number of parameters and its alignment with human similarity judgments and visual cortex activity. In contrast, we found that models with more parameters were more predictive of human choices. Another significant divergence is how intrinsic dimensionality relates to alignment.Elmoznino and Bonner [72] found that vision models with higher latent dimensionality better predict visual cortex activity, Conwell et al. [9] found no correlation. In contrast to this, we observed that lower intrinsic dimensionality led to increased alignment for CLIP models. We hypothesize that both of the observed discrepancies are due to the higher cognitive demands required by our tasks, highlighting the importance of studying alignment in more complex settings. That being said, an alternative explanation for the latter discrepancy could be due to differences in measuring latent dimensionality. Both Elmoznino and Bonner [72] and Conwell et al. [9] use the squared sum of the eigenvalues of principal components divided by the sum of squares of eigenvalues, assuming representations lie on a linear manifold. However, previous work shows that later layers in vision models lie on a curved manifold [66]. Thus, using principal components might not be the best method for this estimation.

Lastly, we found that a method designed for increasing human alignment, DreamSim [12], actually hurt alignment in our task. On the other hand, gLocal [10] and Harmonization [11] improved both performance and human alignment for some models but not all of them. However, the gLocal transform heavily utilises the THINGS dataset, as it made use of the triplet odd-one-out similarity judgement data [27], making it difficult to interpret how well it generalises to other settings. Taken together, these results highlight the importance of studying how well different alignment methods transfer across tasks, as we have done in this work.

5.2 Limitations

There are several limitations and extensions of our work that deserve to be highlighted. The main limitation concerns the interactions between factors in the tested neural networks, making it difficult to isolate specific factors. For example, we would like to test the influence of loss function keeping all other factors equal. Ideally, we would train all combinations of architectures, model sizes, loss functions, and datasets, but this is computationally infeasible.

While we controlled for factors such as training data size and architecture in our comparison of CLIP to other models, there may still be confounding variables we haven't accounted for. For instance, it's not straightforward to compare the information content of image-text pairs used in CLIP training to

https://doi.org/10.52202/079017-3890

image-only data used in other models. Text-image pairs might inherently carry more information than single images, potentially giving multimodal models an advantage that's difficult to quantify. This and other subtle differences in training paradigms could influence our results in ways that are challenging to isolate and measure. Lastly, there can also be other families of models that may outperform CLIP models we haven't considered, such as video models, generative models, or image segmentation models.

We furthermore tested only two experimental paradigms. Future research should explore whether the considered representations predict human behaviour with nonlinear task rules and extend to other paradigms. In particular, one should also consider tasks beyond those generated through the embedding from Hebart et al. [27]. Previous work has shown that it is possible to automatically generate a large set of text-based category learning problems using large language models [80]. It might be interesting to test whether these methods can be extended to generate tasks involving visual stimuli and use these tasks to test whether our findings generalise to a wider setting.

Finally, we only measured human alignment by looking at behaviour. However, to fully confirm our results, it would also be important to investigate the alignment to brain data. Hence, future work should replicate our experiments in an MRI scanner and compare the representations of neural networks to people's brain activity.

5.3 Conclusion

The findings presented in this work have implications both for machine learning and cognitive science. For machine learning, our task and modelling approach offers a new way to measure the human alignment of neural network representations and use this as a metric while building human-aligned neural networks. Alignment at this level can pave the way for artificial systems that can generalise across semantically rich tasks, making them more robust and powerful. For cognitive science, our findings create the opportunity to study several other problems in naturalistic settings by showing that people can do learning tasks with naturalistic stimuli and that pretrained neural networks can be used to extract representations for cognitive models. This could open up the door for a whole new cognitive science that uses naturalistic tasks and environments and thereby increase the validity of the cognitive sciences more generally.

Acknowledgments and Disclosure of Funding

This work was funded by the Max Planck Society, the Volkswagen Foundation, as well as the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC2064/1–390727645.

References

[1] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment, 2023.

[2] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 2020. doi: 10.1101/407007. URL https://www.biorxiv.org/content/early/2020/01/02/407007.

[3] Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen, and Simon Kornblith. Human alignment of neural network representations. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ReDQ1OUQR0X.

[4] Ilia Sucholutsky and Tom Griffiths. Alignment with human representations supports robust few-shot learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 73464–73479. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/e8ddc03b001d4c4b44b29bc1167e7fdd-Paper-Conference.pdf.

[5] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology*, 10(11):e1003915, November 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003915. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003915. Publisher: Public Library of Science.

[6] Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C. Frank, James J. DiCarlo, and Daniel L. K. Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, January 2021. doi: 10.1073/pnas.2014196118. URL https://www.pnas.org/doi/10.1073/pnas.2014196118. Publisher: Proceedings of the National Academy of Sciences.

[7] Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H. McDermott. Many but not all deep neural network audio models capture brain responses and exhibit hierarchical region correspondence, November 2022. URL https://www.biorxiv.org/content/10.1101/2022.09.06.506680v3. Pages: 2022.09.06.506680 Section: New Results.

[8] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, November 2021. doi: 10.1073/pnas.2105646118. URL https://www.pnas.org/doi/10.1073/pnas.2105646118. Publisher: Proceedings of the National Academy of Sciences.

[9] Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *bioRxiv*, 2023. doi: 10.1101/2022.03.28.485868. URL https://www.biorxiv.org/content/early/2023/07/01/2022.03.28.485868.

[10] Lukas Muttenthaler, Lorenz Linhardt, Jonas Dippel, Robert A. Vandermeulen, Katherine Hermann, Andrew Kyle Lampinen, and Simon Kornblith. Improving neural network representations using human similarity judgments. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Nh5dp6Uuvx.

[11] Thomas Fel, Ivan Felipe, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[12] Stephanie Fu, Netanel Yakir Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using

synthetic data. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=DEiNSfh1k7.

[13] Rishi Rajalingham, Elias B. Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J. DiCarlo. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 38(33):7255–7269, 2018. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.0388-18.2018. URL https://www.jneurosci.org/content/38/33/7255.

[14] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision, October 2021. URL http://arxiv.org/abs/2106.07411. arXiv:2106.07411 [cs, q-bio].

[15] 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*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

[16] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J. Kellman. Deep convolutional networks do not classify based on global object shape. *PLOS Computational Biology*, 14(12): 1–43, 12 2018. doi: 10.1371/journal.pcbi.1006613.

[17] Robert Geirhos, Carlos R. M. Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. Generalisation in humans and deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/0937fb5864ed06ffb59ae5f9b5ed67a9-Paper.pdf.

[18] Brandon RichardWebster, Samuel E. Anthony, and Walter J. Scheirer. PsyPhy: A psychophysics driven evaluation framework for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2280–2286, 2019. doi: 10.1109/TPAMI.2018.2849989.

[19] 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)*, pages 1–7, 2017. doi: 10.1109/ICCCN.2017.8038465.

[20] Hossein Hosseini, Baicen Xiao, Mayoore Jaiswal, and Radha Poovendran. On the limitation of convolutional neural networks in recognizing negative images. In Xuewen Chen, Bo Luo, Feng Luo, Vasile Palade, and M. Arif Wani, editors, *16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017*, pages 352–358. IEEE, 2017. doi: 10.1109/ICMLA.2017.0-136.

[21] Kristof Meding, Luca M. Schulze Buschoff, Robert Geirhos, and Felix A. Wichmann. Trivial or impossible — dichotomous data difficulty masks model differences (on imagenet and beyond). In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=C_vsGwEIjAr.

[22] Kamila M. Jozwik, Nikolaus Kriegeskorte, Katherine R. Storrs, and Marieke Mur. Deep Convolutional Neural Networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, 8, 2017. ISSN 1664-1078. doi: 10.3389/fpsyg.2017.01726.

[23] Joshua C. Peterson, Joshua T. Abbott, and Thomas L. Griffiths. Evaluating (and Improving) the Correspondence Between Deep Neural Networks and Human Representations. *Cognitive Science*, 42(8):2648–2669, November 2018. ISSN 1551-6709. doi: 10.1111/cogs.12670.

[24] Raja Marjieh, Pol van Rijn, Ilia Sucholutsky, Theodore R. Sumers, Harin Lee, Thomas L. Griffiths, and Nori Jacoby. Words are all you need? Language as an approximation for human similarity judgments, February 2023. URL http://arxiv.org/abs/2206.04105. arXiv:2206.04105 [cs, stat].

[25] Radoslaw M. Cichy, Nikolaus Kriegeskorte, Kamila M. Jozwik, Jasper J.F. van den Bosch, and Ian Charest. The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *NeuroImage*, 194:12–24, 2019. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2019.03.031.

[26] Marcie L. King, Iris I.A. Groen, Adam Steel, Dwight J. Kravitz, and Chris I. Baker. Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage*, 197:368–382, 2019. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2019.04.079.

[27] Martin N. Hebart, Charles Y. Zheng, Francisco Pereira, and Chris I. Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4(11):1173–1185, November 2020. ISSN 2397-3374. doi: 10.1038/s41562-020-00951-3. URL https://www.nature.com/articles/s41562-020-00951-3. Number: 11 Publisher: Nature Publishing Group.

[28] Brett D. Roads and Bradley C. Love. Enriching ImageNet with human similarity judgments and psychological embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3547–3557, June 2021.

[29] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, April 2018. URL http://arxiv.org/abs/1801.03924. arXiv:1801.03924 [cs].

[30] Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10):e0223792, October 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0223792. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0223792. Publisher: Public Library of Science.

[31] 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, June 2021. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs].

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. URL http://arxiv.org/abs/1512.03385. arXiv:1512.03385 [cs].

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, August 2021. URL http://arxiv.org/abs/2103.14030. arXiv:2103.14030 [cs].

[34] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s, March 2022. URL http://arxiv.org/abs/2201.03545. arXiv:2201.03545 [cs].

[35] Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, and James J. DiCarlo. CORnet: Modeling the Neural Mechanisms of Core Object Recognition, September 2018. URL https://www.biorxiv.org/content/10.1101/408385v1. Pages: 408385 Section: New Results.

[36] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers, May 2021. URL http://arxiv.org/abs/2104.14294. arXiv:2104.14294 [cs].

[37] Ishan Misra and Laurens van der Maaten. Self-Supervised Learning of Pretext-Invariant Representations, December 2019. URL http://arxiv.org/abs/1912.01991. arXiv:1912.01991 [cs].

[38] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments, January 2021. URL http://arxiv.org/abs/2006.09882. arXiv:2006.09882 [cs].

[39] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations, June 2020. URL http://arxiv.org/abs/2002.05709. arXiv:2002.05709 [cs, stat].

[40] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning, January 2022. URL http://arxiv.org/abs/2105.04906. arXiv:2105.04906 [cs].

[41] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction, June 2021. URL http://arxiv.org/abs/2103.03230. arXiv:2103.03230 [cs, q-bio].

[42] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised Representation Learning by Predicting Image Rotations, March 2018. URL http://arxiv.org/abs/1803.07728. arXiv:1803.07728 [cs].

[43] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved Baselines with Momentum Contrastive Learning, March 2020. URL http://arxiv.org/abs/2003.04297. arXiv:2003.04297 [cs].

[44] Mehdi Noroozi and Paolo Favaro. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles, August 2017. URL http://arxiv.org/abs/1603.09246. arXiv:1603.09246 [cs].

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

[46] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal Sentence Encoder, April 2018. URL http://arxiv.org/abs/1803.11175. arXiv:1803.11175 [cs].

[47] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv:1810.04805 [cs].

[48] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, February 2020. URL http://arxiv.org/abs/1910.01108. arXiv:1910.01108 [cs].

[49] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019. URL http://arxiv.org/abs/1907.11692. arXiv:1907.11692 [cs].

[50] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020. URL http://arxiv.org/abs/2005.14165. arXiv:2005.14165 [cs].

[51] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in Pre-Training Distributed Word Representations, December 2017. URL http://arxiv.org/abs/1712.09405. arXiv:1712.09405 [cs].

[52] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training, 2021.

[53] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

[54] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773.

[55] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws

for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[56] Can Demircan, Leonardo Pettini, Tankred Saanum, Marcel Binz, Blazej Metody Baczkowski, Christian Doeller, Mona Garvert, and Eric Schulz. Decision-making with naturalistic options. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, 2022.

[57] Daniel McFadden. Conditional logit analysis of qualitative choice behavior. 1972.

[58] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[59] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 491–507, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58558-7.

[60] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019.

[61] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12104–12113, June 2022.

[62] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era, 2017.

[63] Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. Why do better loss functions lead to less transferable features?, 2021.

[64] Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Bjorn Ommer, and Joseph Paul Cohen. Revisiting training strategies and generalization performance in deep metric learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8242–8252. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/roth20a.html.

[65] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited, July 2019. URL http://arxiv.org/abs/1905.00414. arXiv:1905.00414 [cs, q-bio, stat].

[66] Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks, 2019.

[67] Raef Bassily, Shay Moran, Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. Learners that use little information. In *Algorithmic Learning Theory*, pages 25–55. PMLR, 2018.

[68] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[69] Tankred Saanum, Noémi Éltető, Peter Dayan, Marcel Binz, and Eric Schulz. Reinforcement learning with simple sequence priors. *Advances in Neural Information Processing Systems*, 36, 2024.

[70] Tankred Saanum, Peter Dayan, and Eric Schulz. Predicting the future with simple world models. *arXiv preprint arXiv:2401.17835*, 2024.

[71] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

[72] Eric Elmoznino and Michael F Bonner. High-performing neural network models of visual cortex benefit from high latent dimensionality. *PLOS Computational Biology*, 20(1):e1011792, 2024.

[73] Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140, 2017.

 https://doi.org/10.52202/079017-3890

[74] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: the new data in multimedia research. *Communications of the ACM*, 59(2):64–73, January 2016. ISSN 1557-7317. doi: 10.1145/2812802. URL http://dx.doi.org/10.1145/2812802.

[75] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[76] Drew Linsley, Dan Shiebler, Sven Eberhardt, and Thomas Serre. Learning what and where to attend. *International Conference on Learning Representations (ICLR)*, 2019.

[77] Ruairidh M. Battleday, Joshua C. Peterson, and Thomas L. Griffiths. Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications*, 11(1):5418, October 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-18946-z. URL https://www.nature.com/articles/s41467-020-18946-z. Number: 1 Publisher: Nature Publishing Group.

[78] Lukas Muttenthaler and Martin N. Hebart. THINGSvision: A Python Toolbox for Streamlining the Extraction of Activations From Deep Neural Networks. *Frontiers in Neuroinformatics*, 15, 2021. ISSN 1662-5196. URL https://www.frontiersin.org/articles/10.3389/fninf.2021.679838.

[79] Aria Y. Wang, Kendrick Kay, Thomas Naselaris, Michael J. Tarr, and Leila Wehbe. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5(12):1415–1426, November 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00753-y. URL http://dx.doi.org/10.1038/s42256-023-00753-y.

[80] Akshay K Jagadish, Julian Coda-Forno, Mirko Thalmann, Eric Schulz, and Marcel Binz. Ecologically rational meta-learned inference explains human category learning. *arXiv preprint arXiv:2402.01821*, 2024.

[81] Joshua R. de Leeuw. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, 47(1):1–12, March 2015. ISSN 1554-3528. doi: 10.3758/s13428-014-0458-y. URL https://doi.org/10.3758/s13428-014-0458-y.

[82] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

[83] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67:1–48, October 2015. ISSN 1548-7660. doi: 10.18637/jss.v067.i01. URL https://doi.org/10.18637/jss.v067.i01.

[84] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. ISSN 1533-7928. URL http://jmlr.org/papers/v12/pedregosa11a.html.

[85] Laura Mai Stoinski, Jonas Perkuhn, and Martin N. Hebart. THINGSplus: New Norms and Metadata for the THINGS Database of 1,854 Object Concepts and 26,107 Natural Object Images, July 2022. URL https://psyarxiv.com/exu9f/.

[86] Maarten Speekenbrink, Shelley Channon, and David R Shanks. Learning strategies in amnesia. *Neuroscience & Biobehavioral Reviews*, 32(2):292–310, 2008.

[87] Samuel J Gershman. A unifying probabilistic view of associative learning. *PLoS computational biology*, 11(11):e1004567, 2015.

[88] Marcel Binz, Samuel J. Gershman, Eric Schulz, and Dominik Endres. Heuristics from bounded meta-learned inference. *Psychological Review*, 129(5):1042–1077, October 2022. ISSN 1939-1471. doi: 10.1037/rev0000330.

[89] Raja Marjieh. The universal law of generalization holds for naturalistic stimuli, Jan 2024. URL osf.io/rbkgh.

A Methods

Participants For the category learning task, we recruited 98 participants (48 females, 50 males, mean age= 28.92y, SD= 7.32) on the Prolific platform. Participants with less than 50% accuracy were excluded from the analyses, leaving us with 91 participants. A base payment of £ 1.50 was made, and participants could earn an additional bonus of £ 6.00. The median completion time was 12 minutes and 38 seconds. The inclusion criteria included having a minimum approval rate of 97%, and a minimum number of 15 previous submissions on Prolific. Participation in the reward learning study was an exclusion criterion. For the reward learning task, 99 participants were recruited (49 females, 49 males, 1 other, mean age = 27.9 y, SD = 9.13). After applying the 50% accuracy criteria, we were left with 82 participants. A base payment of £ 2.00 was made, and an additional performance-dependent bonus of £4.00 was offered. The median completion time was 9 minutes and 26 seconds. The inclusion criteria included having a minimum approval rate of 95%, and a minimum number of 10 previous submissions on Prolific. All participants agreed to their anonymized data being used for research. The study was approved by the ethics committee of of the medical faculty of the University of Tübingen (number 701/2020BO). Participants gave consent for their data to be anonymously analyzed by agreeing to a data protection sheet approved by the data protection officer of the MPG (Datenschutzbeauftragte der MPG, Max-Planck-Gesellschaft zur Förderung der Wissenschaften).

Tasks and Stimuli Both tasks were run online in forced full-screen mode. Participants were shown written instructions and were asked to complete comprehension check questions before they could start the tasks. In both tasks, participants were given unlimited time to make decisions. In the category learning task, binary (correct versus wrong) feedback was given for $2s$. In the reward learning task, the associated reward with the stimuli was shown for $1.5s$, and there was an inter-trial interval of $1s$ where participants were shown a blank screen. Throughout both tasks, the estimated total payment of participants was shown on the upper part of the screen. At the end of the tasks, participants were asked whether they thought their data should be used for analysis. Across both tasks, all but one participant responded saying their data should be analyzed, whose data was anyway excluded due to poor performance. The category learning task was programmed using jsPsych [81], whereas the reward learning task was programmed in plain JavaScript.

For each participant, 120 stimuli were sampled independently from the THINGS database. Because the loadings of the features were not uniformly distributed, we made 5 equally sized bins of the loadings for the assigned feature and sampled object categories uniformly from these bins. From these object categories, the specific images were assigned randomly. For details on the used features and the embedding, see Hebart et al. [27].

Behavioural Analyses We used mixed-effects logistic models for both category and reward learning analyses. For category learning, we predicted correct responses per trial, using trial number as a fixed effect and including participant-specific random effects for intercept, trial number, and assigned task rule. In the reward learning model, we predicted whether the image on the right is selected, incorporating the trial number, reward difference between images, and their interaction as fixed effects. These factors, along with the assigned task rule, were also modelled as participant-specific random effects. Both models effectively captured task structure, learning progression, and individual variability in performance. In R formula notation[82, 83], the model for category learning is denoted as follows:

correct_choice $\sim$1 + trial + (1 + trial + dimension | participant)

For the reward learning task, the following model was used:

right_choice $\sim$−1 + trial $*$ right_left_reward_difference +
(−1 + trial + dimension + right_left_reward_difference | participant)

where -1 denotes no intercept.

Software, Data, & Compute Resources The code to reproduce the reported results is available at `https://github.com/candemircan/naturalcogsci` and we provide anonymised human choice data on `https://osf.io/h3t52/`. For the learning models we used `lme4` [83] and `scikit-learn` [84]. To extract representations from neural networks we used `thingsvision` [78].

https://doi.org/10.52202/079017-3890

Computations were performed on an academic SLURM cluster. Feature extraction was done on a single Nvidia A100 GPU (40GB) under 24 hours. The linear models were parallelised across several jobs that used single core and 8GB RAM and were completed in under 48 hours. The mixed-effects models were similarly parallelised and completed under 24 hours.

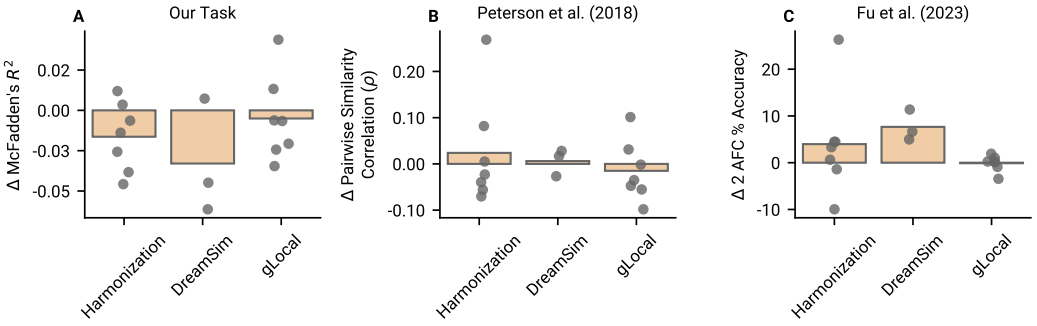

Figure 9: Change of human alignment for different methods on different datasets

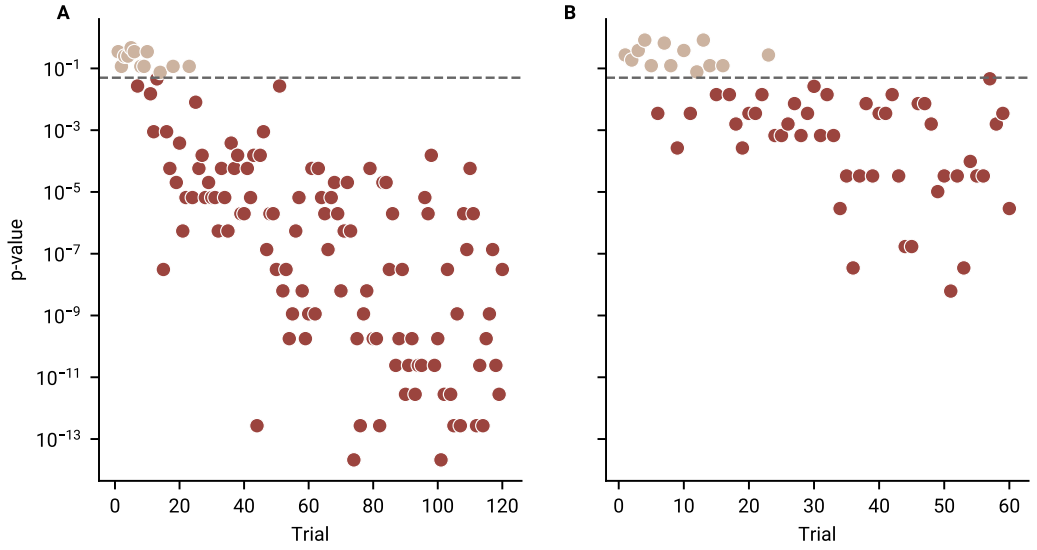

Figure 10: Participant Performance Against Chance Level at Each Trial. Trial-by-trial p-values from 1 sample t-tests testing accuracy against chance level for (*A*) category learning task and the (*B*) reward learning task.

Modelling

For the category learning task, we used an ℓ_2 regularised logistic regression model to optimize regression weights. We relied on `scikit-learn`'s `LogisticRegression` class which internally optimizes the following objective:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \sum_{i=1}^{N} -c_i \log(p(c_i|\mathbf{x}_i, \mathbf{w})) - (1 - c_i) \log(1 - \log(p(c_i|\mathbf{x}_i, \mathbf{w})) + \frac{1}{2}||\mathbf{w}||_2^2 \quad (5)$$

For the reward learning task, we used a Bayesian linear regression model to infer a posterior distribution over regression weights. We relied on `scikit-learn`'s `BayesianRidge` class which infers a posterior distribution assuming spherical Gaussian priors (i.e., $p(\mathbf{w}) = \mathcal{N}(0, \lambda^{-1}\mathbf{I})$) and Gaussian likelihood (i.e., $p(y_i|\mathbf{x}_i, \mathbf{w}) = \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i, \beta^{-1})$). Based on these assumptions, the posterior

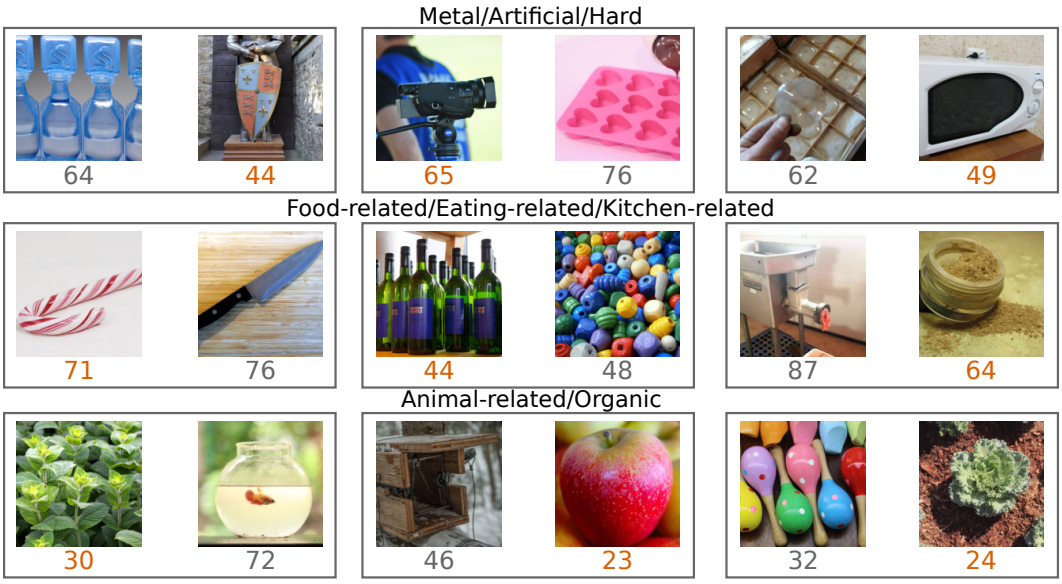

Figure 11: Example trials showing the similarity between CLIP and human decisions that show disagreement with the task embedding. Each row shows three trials from a different condition. Orange highlighted text shows the option chosen by all CLIP models and the human participant, whereas grey text shows the decision made by the task embedding. As the tasks were generated using the task embedding, all the choices shown here made by CLIP and humans are suboptimal. Shown examples are from the second half of the task, as to eliminate the learning process as a confound. The original images are replaced with copyright-free alternatives from the THINGSplus database [85].

distribution can be computed in closed form:

$$p(\mathbf{w}|\mathbf{X},\mathbf{y}) = \mathcal{N}(\mathbf{m}_N, \mathbf{S}_N) \tag{6}$$

$$\mathbf{m}_N = \beta \mathbf{S}_N \mathbf{X}^\top \mathbf{y} \tag{7}$$

$$\mathbf{S}_N^{-1} = \lambda \mathbf{I} + \beta \mathbf{X}^\top \mathbf{X} \tag{8}$$

where $\mathbf{X}$ and $\mathbf{y}$ denote the stacked inputs and targets respectively.

We run both models from scratch on each trial using all previously observed input-target pairs. The choice of these models was motivated by previous investigations in similar – but low-dimensional – settings [86, 87, 88].

The ℓ_2 penalty term for the logistic regression model described above was determined via grid search to maximise task performance, on a per participant basis. For the linear regression model, λ and β were fitted to maximise the log marginal likelihood on the task performances.

The estimates from these models were used in mixed-effects logistic regression models with leave-one-out predictions to assess participant choices. For category learning, we used logistic regression probability estimates as predictors. In the reward learning task, we used the difference in estimated rewards from linear regression models as predictors. In both cases, these predictors were included as both fixed and random effects, allowing us to account for individual differences while maintaining the group effects. These correspond to the following models in R formula notation for category and reward learning respectively:

human_choice ~−1 + probability_estimate + (−1 + probability_estimate | participant)

human_choice ~−1 + estimated_reward_difference + (−1 + estimated_reward_difference | participant)

For the mixed-effects models, the training data was centred and divided by its standard deviation. The same scaling parameters were applied to the test data.

https://doi.org/10.52202/079017-3890

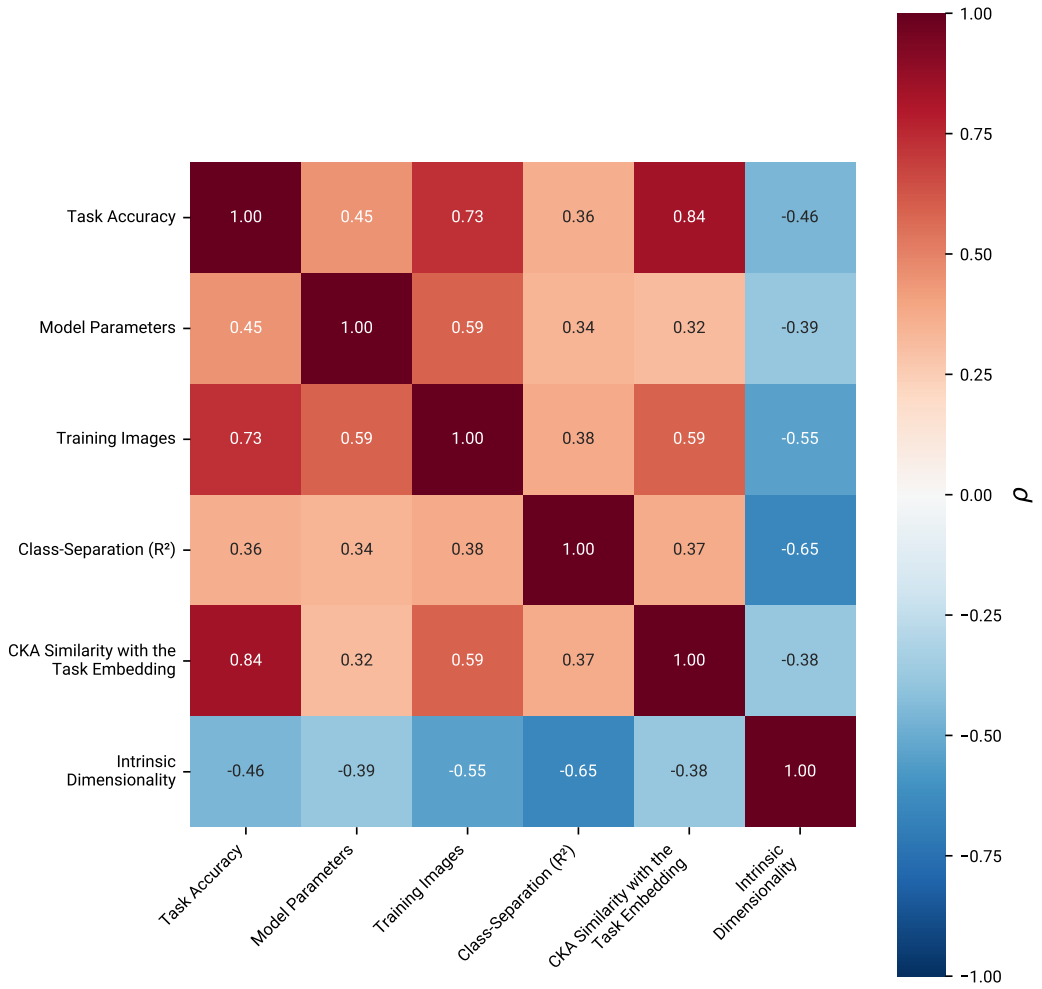

Figure 12: Pairwise Spearman correlations between the factors investigated that contribute to alignment.

Additional human-alignment tasks

Results reported in Fig. 8A & Fig. 8B include comparisons for the overlap of models between those reported by Muttenthaler et al. [3] and the ones we tested. For Fig. 8B & Fig. 8C, we tested all the vision models reported in our paper. However, for the Peterson et al. [23] dataset, we only found a subset of the original data reported in the paper[1]. The ClickMe-Harmonizer alignment was only computed for supervised models, as the method requires computing gradients for ImageNet classes, which we could only do for the supervised models that had ImageNet classification heads.

B Testing aligned models on other datasets

Above, we also report how different alignment methods perform on different datasets (Fig. 9). Harmonization is on average slightly more human-like on the two external tested datasets compared to baselines. DreamSim shows mixed results for the Peterson et al. [23] dataset, but it shows improvement on the NIGHTS dataset [12]. This is not surprising, as this dataset was used to build DreamSim. Lastly, gLocal shows mixed results.

[1]Specifically, we tested similarity judgements obtained on Animal, Fruit, and Vegetable categories. The data was obtained from [89]

C Additional results

Above we provide some additional results supporting our claims in the main text. Fig. 10 shows participants can do both tasks above chance level very early on in the task. Fig. 11 shows some incorrect choices made by humans and also by CLIP models, and Fig. 12 shows pairwise correlations between the factors we investigated that contribute to alignment.

https://doi.org/10.52202/079017-3890

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions are clearly marked in the introduction, and they directly match the provided results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalise to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, there is a dedicated section for this.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: There are no theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all the relevant details for the experiments and the analyses that are needed for reproducing the results. We will also share the data and the code before the conference on GitHub and OSF.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/079017-3890

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All the data and the code are available in online repositories. We will make them available before the conference.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: These details are provided in the **Results** section of the main text, and in the **Methods** section in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The error bars and lines are shown in the figures. The captions explain what the error bars represent (95% confidence intervals).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the approximate compute resources used for our work in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our work conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There are no potential positive or negative societal impacts that may arise directly from our work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

https://doi.org/10.52202/079017-3890

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The data we release do not have a risk of misuse. No new models are provided.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all the pretrained models we used. The used images either had a public domain or CC0 copyright license.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: All the code and the data, which will be released before the conference, are documented.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We provide a summary of the task and the instructions given to the participants both in the main text and in the Appendix. The experiment code, including the full instructions given to the participants, will be made available before the conference.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The behavioural studies carried no risks for participants. Both studies were approved by the local ethics committee.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

https://doi.org/10.52202/079017-3890

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.