Individual differences in artificial neural networks capture individual differences in human behavior

Herrick Fung, N. Apurva Ratan Murty, & Dobromir Rahnev

School of Psychology, Georgia Institute of Technology

Keywords: Individual differences; artificial neural networks; human-model alignment; perceptual decision making; confidence

Author Contributions: All authors jointly designed the research, interpreted the results, and contributed to revising and editing the manuscript. H.F. conducted the research, collected and analyzed the data, and drafted the manuscript. N.A.R.M and D.R. reviewed the manuscript.

Competing Interests: The authors declared no competing interests.

Acknowledgment: This work was supported by the National Institute of Health (award: R01MH119189). We thank Alish Dipani for helpful suggestions and discussions.

Correspondence:
Herrick Fung
School of Psychology
Georgia Institute of Technology
654 Cherry Str NW
Atlanta, GA 30332
Email: herrickfung@gmail.com

Abstract

Human behavior differs substantially across individuals. While artificial neural networks (ANNs) are regarded as promising models of human perception, they are often assumed to lack such individual differences. Here, we demonstrate that multiple instances of the same ANN architecture exhibit substantial individual differences in behavior that mimic those observed in humans. We trained and tested 60 ANN instances from three architectures on a digit recognition task and found notable individual differences in overall accuracy, confidence, and response time (RT). Critically, these individual differences in ANN instances mapped consistently onto the individual differences produced by 60 humans performing the same task, with the mapping strength often approaching the human-to-human benchmark across all three behavioral metrics (accuracy, confidence, RT). The mapping generalized even across behavioral metrics: an ANN instance that aligned with an individual human on accuracy also aligned with the same individual on confidence and RT. These findings generalized to a more complex, 10-choice blurry object recognition task, though the human-ANN mapping was generally less robust than the human-human benchmark. Overall, these findings open the possibility of using ANN ensembles as computational proxies for probing the mechanisms underlying human variability.

1 Introduction

No two individuals perceive, think, or act in the same way. Understanding and predicting individual differences in human behavior is a critical challenge in the quest to uncover the complexity of human intelligence. In recent years, artificial neural networks (ANNs) have emerged as the leading computational models of the primate visual system and perceptual decision making for complex images (Kriegeskorte, 2015; McGrath et al., 2024; Wichmann & Geirhos, 2023). Yet, since Turing's original proposal of the imitation game (Turing, 1950), AI and, more recently, NeuroAI (Feather et al., 2025) have largely aimed to model the behavior of an *average* human. Whether ANNs can capture human behavior at the individual level, that is, the full spectrum of human behavioral diversity, is yet to be explored.

Indeed, ANNs are often considered uniform within a given architecture (e.g., AlexNet), with little meaningful variability between different model instances. It is only recently that studies challenged this assumption, demonstrating that even within a fixed ANN architecture, different random weight initialization (Mehrer et al., 2020) and different order of the training images (Chow & Palmeri, 2024) can lead to substantial differences in the internal representations. Critically, these representational differences are likely to translate into behavioral variability and several recent papers have found clear behavioral differences between ANN instances that only differ in their random initializations (Baek et al., 2021; Green et al., 2026; Rafiei et al., 2024; Shekhar & Rahnev, 2024).

While differences between ANN instances are beginning to gain attention, it remains unclear whether such individual differences in ANNs can be meaningfully related to individual differences in human behavior. If individual differences in ANNs mimic those observed in humans, they could potentially serve as proxy models for human variability, supporting advances in both theoretical insight and empirical investigation. For instance, by manipulating model instances and testing their alignment with individuals, we could use ANNs to reverse engineer the origins of individual differences in humans, offering a new approach to probe the mechanisms underlying behavioral variability (Chow & Palmeri, 2024; Strock et al., 2025; Wood et al., 2024). A proxy model for human variability would also allow new *in*

silico experiments to precisely manipulate human behavior the individual level (Bashivan et al., 2019; Mueller et al., 2023; Ponce et al., 2019; Shahbazi et al., 2024; Walker et al., 2019).

Here we examine whether individual differences in ANNs capture individual differences in human behavior. We trained and evaluated 60 instances from three ANN architectures on a digit recognition task and observed substantial variability in accuracy, confidence, and response time (RT). We then tested the alignments between human individuals and individual ANN instance and found that individual differences in ANN instances mapped consistently onto 60 human subjects performing the same task, with the strength of the relationship often approaching the strength of human-to-human mapping consistency. Remarkably, these human-ANN mappings were not confined to a single behavioral metric: ANN instances that mimic an individual in one behavioral metric also tend to mimic the same individual on other metrics. These results even generalized to a 10-choice blurry object recognition task, although the human-ANN mappings were less robust than the human-human benchmark. Taken together, these findings open the possibility of using ANNs in uncovering the mechanisms underlying individual variability in human behavior.

2 Results

To explore whether the individual differences in ANNs are related to individual differences in humans, we first reanalyzed data from a previous study where N=60 human subjects performed an 8-choice noisy digit recognition task (Figure 1a; Rafiei et al., 2024). We recorded accuracy, confidence, and RT on each trial. We trained models based on three different ANN architectures – RTNet (Rafiei et al., 2024), AlexNet (Krizhevsky et al., 2012), and ResNet18 (He et al., 2015). For each architecture, we trained 60 independent instances by changing the random initialization. All models achieved at least 97% accuracy on a validation dataset of noiseless images. We then adjusted the noise level in the images separately for each network architecture, such that the average accuracy of the 60 instances from that architecture (RTNet: 0.72; AlexNet: 0.72; ResNet18: 0.71) matched the average accuracy of human subjects (0.71; Figure 1b).

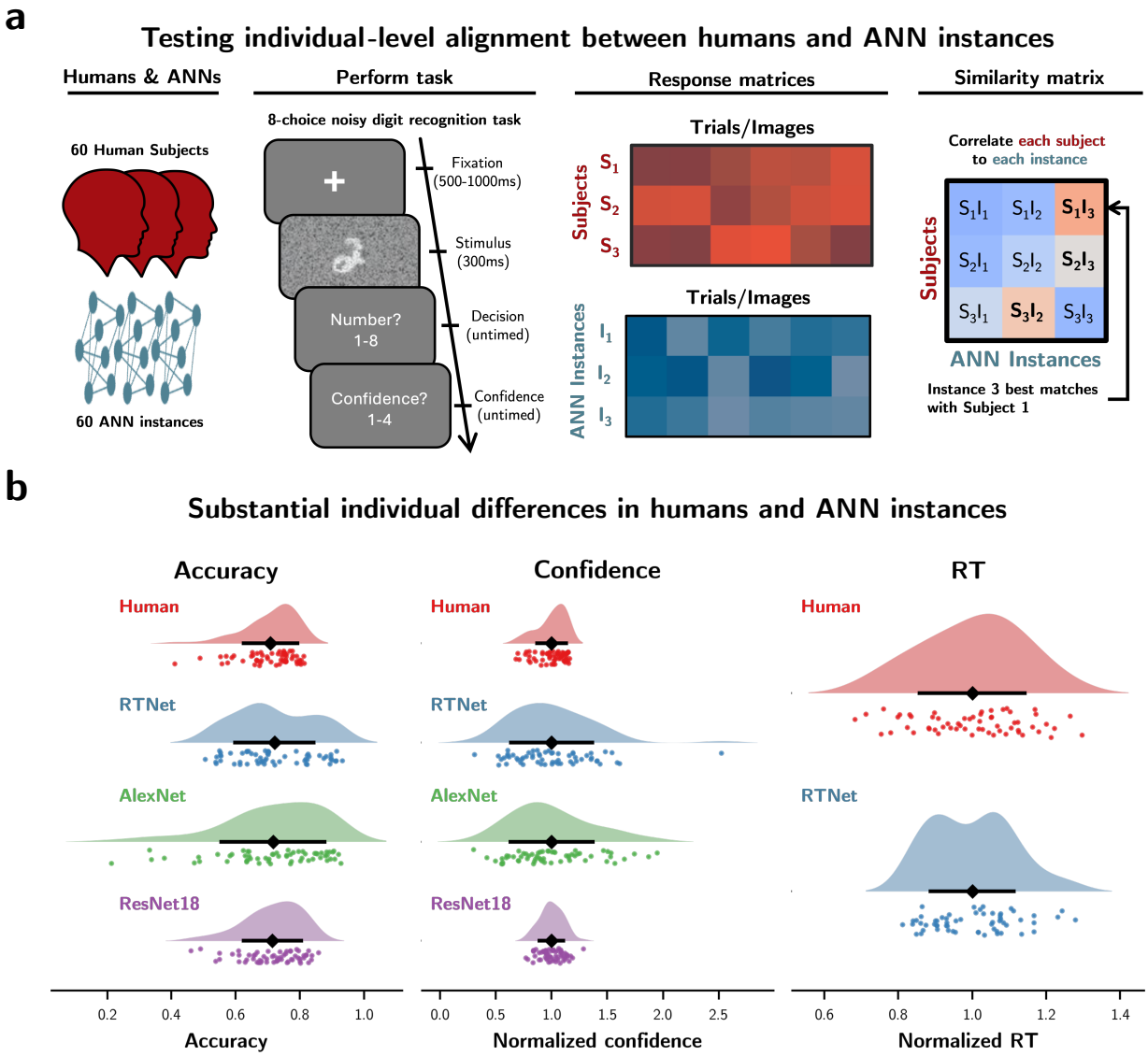


Figure 1. Task schematics and behavioral differences in humans and ANN instances. (a) Schematics of testing individual-level alignment. Human subjects (N=60) and ANN instances (N = 60 from each of three architectures: RTNet, AlexNet, ResNet18) performed an 8-choice noisy digit recognition task. On each trial, human subjects first saw a fixation cross for 500-1000 ms, followed by a noisy image of a handwritten digit from the MNIST dataset presented for 300 ms. Subjects then reported their decision, followed by their confidence rating. ANN instances were presented the same MNIST images and we extracted responses and confidence ratings equivalent to those in humans. Responses from human subjects and ANN instances were structured into response matrices. We correlated the image-by-image responses of each subject to each ANN instance to construct the similarity matrix, where each cell indicates how well each subject aligns to each ANN instance. (b) Substantial individual differences in humans and ANN instances. Confidence and RT are normalized by their mean across subjects or instances for visualization. Diamond markers denote the mean. Error bars show standard deviation. Dots represents individual subjects or individual ANN instances.

2.1 Substantial individual differences in humans and ANN instances

We first confirmed that there were substantial individual differences in both ANN instances and humans (Figure 1b). Indeed, even though ANN instances from each architecture differed only in their random initializations, they showed a very large range in average accuracy when tested on noisy images (RTNet: range = [0.50, 0.93], SD = 0.12; AlexNet: range = [0.21, 0.93], SD = 0.16; ResNet18: range = [0.46, 0.86], SD = 0.09). Surprisingly, the variability among ANN instances was even larger than the variability among human subjects (range = [0.41, 0.81], SD = 0.08). We obtained similar results for confidence and RT. Since confidence and RT are on different scales across humans and each ANN architecture, we quantified the individual variability as the coefficient of variation (CV, standard deviation divided by the mean). Using this metric, we observed greater individual variability in confidence for RTNet and AlexNet than in humans (humans: CV = 0.13; RTNet: CV = 0.37; AlexNet: CV = 0.39), whereas ResNet18 showed a similar level of variability than humans (ResNet18: CV = 0.11). RT was assessed only for RTNet (as the other models do not produce RT) and showed comparable level of variability between humans and RTNet (humans: CV = 0.14; RTNet: CV = 0.11). Similar results were found when examining variability in the full trial-level response distributions using Wasserstein distances (Supplementary Figure 1 & Supplementary Results). These results demonstrate that ANNs that only differ in their random initialization exhibit large individual differences that either match or exceed the variability among human subjects.

2.2 Individual ANN instances vary in their alignment with individual human subjects

Having confirmed the existence of substantial individual differences in ANNs and humans, we explored how well each ANN instance aligns with each human subject. For each behavioral metric (accuracy, confidence, and RT), we computed the correlations between all pairs of human subjects and network instances. We found a wide range of correlations between ANN instances and individual humans (see Figure 2a for accuracy and Supplementary Figure 2 for confidence and RT). For example, when considering the correlations for accuracy only, we found a wide range of correlation values for human-

human (-0.02 to 0.57), human-RTNet (-0.22 to 0.50), human-AlexNet (-0.15 to 0.48), and human-ResNet18 (-0.01 to 0.47) correlations. Similar results were obtained for confidence (human-human: [-0.09, 0.55], human-RTNet: [-0.20, 0.51], human-AlexNet: [-0.22, 0.57]; human-ResNet18: [-0.08, 0.47]) and RT (human-human: [-0.04, 0.47], human-RTNet: [-0.22, 0.47]).

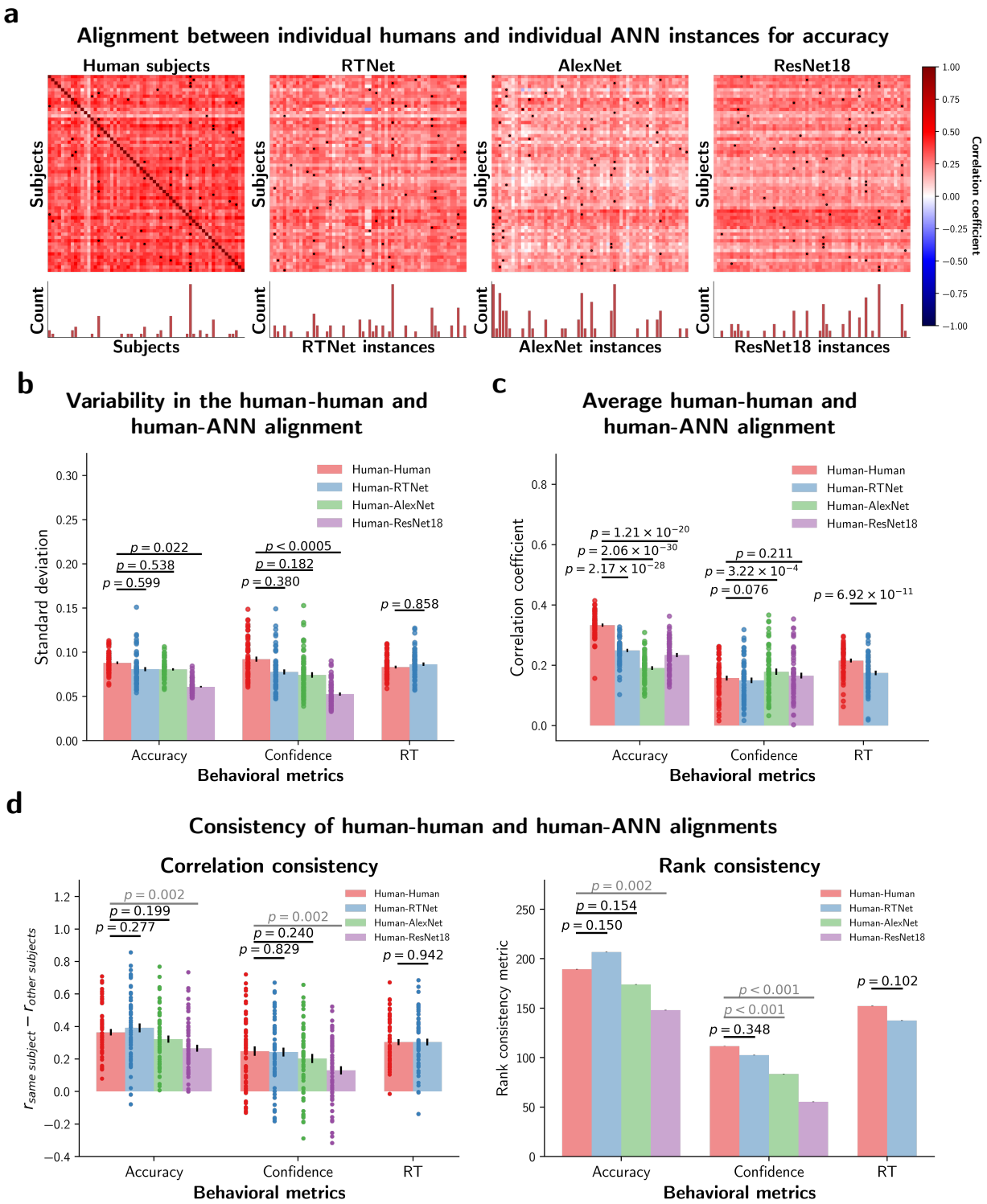


Figure 2. Individual ANN instances vary in their alignment with individual human subjects. (a) Similarity between each subject and each ANN instance in accuracy. We constructed the similarity matrix of 60 human subjects and the 60 instances of each ANN, with higher values indicating a better alignment between an individual human and a specific ANN instance. The black dots indicate the model instance with the best fit for each subject (i.e., there is one black dot for each row). The histograms underneath show how often each subject or ANN instance was the best match across all human subjects. (b)

Variability in the human-human and human-ANN alignment. The variability in alignment between individual humans and ANN instances was similar to the variability in alignment between each human subjects and the remaining subjects. P-values indicate the significance of pairwise differences in the standard deviation of the similarity matrix. (c) Average human-human and human-ANN alignment. Individual ANN instances showed slightly lower overall alignment with humans compared to the alignment between different humans. Error bars show SEM. Dots show individual human subjects or ANN instances. P-values indicate the significance of pairwise differences in the average correlation. (d) Consistency of human-human and human-ANN mapping. Using both the correlation and rank methods of quantifying consistency converge on the same findings that individual ANN instances map to human subjects as consistently as humans do. This effect is true for RTNet and AlexNet, whereas ResNet18 shows lower consistency than the human benchmark. Error bars show SEM. Dots show individual human subjects. P-values for correlation consistency indicates the significance of pairwise differences in the adjusted correlation between human-human and human-ANN mapping, computed using paired t-tests. P-values for rank consistency indicates the significance of pairwise differences in the rank consistency metric.

The similarity matrices in Figure 2a show that the same human subject or same ANN instance sometimes provides the best match to the data for multiple human subjects. We counted how often each subject or ANN instance was the best match across all human subjects and demonstrated largely similar patterns for the human-human and human-ANN alignment (Supplementary Figure 3 & Supplementary Results).

To further assess the amount of variation in the human-human and human-ANN alignment, we computed the standard deviation of the human-human and human-ANN correlation values for each of the 60 human subjects. This allowed us to examine, for each subject, the variability in how well other humans or ANN instances correlate with that subject. We found that variability in human-human correlations was broadly similar to the variability in human-ANN correlations (Figure 2b). Specifically, human-RTNet variability was indistinguishable from human-human variability for accuracy ($p = 0.31$) and confidence ($p = 0.371$), though human-RTNet variability was higher than human-human variability for RT ($p = 0.005$). Human-AlexNet variability was also similar to human-human variability for both accuracy ($p = 0.258$) and confidence ($p = 0.549$). In contrast, human-ResNet18 correlations demonstrated significantly lower variability than human-human correlations for both accuracy ($p = 0.009$) and confidence ($p = 0.01$), though the standard deviations decreased by a relatively small amount (up to 31%). Overall, the variability in how well individual humans are aligned with other individual humans was either comparable or only slightly higher than the variability in how well they are aligned with ANN instances.

For completeness, we also examined the strength of human-human and human-ANN alignment. We found that the average human-human correlations were higher than the average human-ANN correlations for all architectures and metrics (all p's < 0.01; Figure 2c) except for RTNet and ResNet18 in the case of confidence where no significant difference was observed (RTNet: $p = 0.076$; ResNet18: $p = 0.211$). These results demonstrate that individual ANN instances are not as good as human subjects at predicting individual human performance, although the difference is relatively small for RTNet where the decrease in correlation was 25%, 5%, and 19% for accuracy, confidence, and RT, respectively.

2.3 Individual ANN instances show consistent mapping with human subjects for different stimulus subsets

The results in the previous section demonstrated the existence of substantial individual differences between ANN instances in their alignment with individual human subjects. However, some amount of individual difference is expected from noise alone. Therefore, it is critical to demonstrate that the differences between ANN instances in how they align with individual subjects are consistent and persist even when assessed using separate sets of images. To this end, we performed 1000 iterations of randomly splitting the trials into two equal parts and computed the correlation matrix for each split. Finally, we evaluated the consistency of the correlation matrix and compared the obtained consistency values to human-human consistency as a benchmark.

We first quantified the consistency of the human-ANN similarity matrices by computing the Pearson correlation between the two halves of each of the 1000 trial splits within each individual subject. We then corrected these values for each trial split by subtracting the average correlation between the same subject and all other subjects, which removes any correlations stemming from ANN instances that may perform uniformly well or uniformly badly in matching all human subjects. We found that human-RTNet correlations were as consistent as the human-human benchmark for accuracy ($t(59) = 1.09$, $p = 0.277$, BF_{01}

= 3.00; Figure 2d), confidence ($t(59) = 0.217$, $p = 0.829$, $BF_{01} = 5.03$), and RT ($t(59) = 0.07$, $p = 0.942$, $BF_{01} = 5.13$). The same was also true for AlexNet (accuracy: $t(59) = 1.29$, $p = 0.199$, $BF_{01} = 2.42$; confidence: $t(59) = 1.18$, $p = 0.240$, $BF_{01} = 2.74$) but not for ResNet18, which exhibited lower consistency than the human-human benchmark for both accuracy ($t(59) = 3.10$, $p = 0.002$, $BF_{10} = 13.3$) and confidence ($t(59) = 3.17$, $p = 0.002$, $BF_{10} = 16.4$).

One concern with using Pearson correlation is that high consistency values could be driven by a few ANN instances that always show bad alignment with most subjects and it could be that our subtraction method doesn't fully correct for this effect. To mitigate this concern, we conducted a complementary analysis. For each subject, we computed the rank order of how well each ANN instance (or each subject except the selected one) is aligned with that subject. Then, for each pair of subjects, we calculated the difference in the ranks of ANN instances (or other subjects) for each of two subsets of the data. We then took the average of the product of these differences across the two subsets, with positive values indicating consistent rank ordering. Using this alternative metric of consistency, we found that human-RTNet similarity matrix had comparable consistency to the human-human benchmark for accuracy ($p = 0.15$, 95% bootstrap CI = [-40.7, 6.66]), confidence ($p = 0.348$, 95% bootstrap CI = [-9.90, 28.5]), and RT ($p = 0.102$, 95% bootstrap CI = [-3.95, 32.9]). Similarly, human-AlexNet similarity matrix was as consistent as the human benchmark for accuracy ($p = 0.154$, 95% bootstrap CI = [-5.10, 35.6]), but not for confidence ($p < 0.001$, 95% bootstrap CI = [11.6, 44.5]). Lastly, the human-ResNet18 similarity matrix exhibited significantly lower consistency than the human benchmark for both accuracy and confidence (both p's < 0.003). Overall, both the Pearson correlation and the rank order methods of quantifying consistency demonstrate that individual ANN instances for RTNet and AlexNet map to human subjects as consistently as human subjects map among themselves (while ResNet18 instances show lower consistency). These results demonstrate that the variability in human-ANN mapping is not simply due to noise but reflects meaningful variability among ANN instances.

2.4 ANN instances that mimic an individual on one metric also mimic the same individual on other metrics

Our results so far show that a given ANN instance is more similar to some human subjects than others. However, we showed this fact separately for each of our three behavioral metrics (accuracy, confidence, and RT). Here we test whether an ANN instance that mimics a subject's behavior on one metric (e.g., accuracy) also capture that subject's behavior for another metrics (e.g., confidence). To this end, we computed the correlations between the similarity matrices for each subject across pairs of behavioral metrics (Figure 3a). We found that the human-ANN across-metric correlations were significantly positive for all pairs of behavioral metrics, including accuracy-confidence (human-RTNet: average $r = 0.347$; $t(59) = 11.9$, $p = 2.09 \times 10^{-17}$; human-AlexNet: average $r = 0.367$; $t(59) = 15.7$, $p = 1.14 \times 10^{-22}$; human-ResNet18: average $r = 0.109$; $t(59) = 4.14$, $p = 1.14 \times 10^{-4}$; Figure 3b), accuracy-RT (human-RTNet: average $r = 0.384$; $t(59) = 12.4$, $p = 4.14 \times 10^{-18}$), and confidence-RT (human-RTNet: average $r = 0.519$; $t(59) = 13.3$, $p = 1.92 \times 10^{-19}$). Moreover, the human-RTNet across-metric correlations were statistically indistinguishable from the human-human control for accuracy-confidence ($t(59) = -1.34$, $p = 0.186$, $BF_{01} = 3.05$) and accuracy-RT ($t(59) = 1.28$, $p = 0.205$, $BF_{01} = 3.26$), and significantly outperform the human-human control for RT-confidence ($t(59) = 7.19$, $p = 1.30 \times 10^{-9}$, $BF_{10} = 8.55 \times 10^{6}$). Similarly, the human-AlexNet correlation was comparable to the human-human control for accuracy-confidence ($t(59) = -1.02$, $p = 0.313$, $BF_{01} = 4.33$). In contrast, the human-ResNet18 correlation was significantly worse than the human-human control for accuracy-confidence ($t(59) = -9.36$, $p = 2.87 \times 10^{-13}$, $BF_{10} = 2.76 \times 10^{10}$).

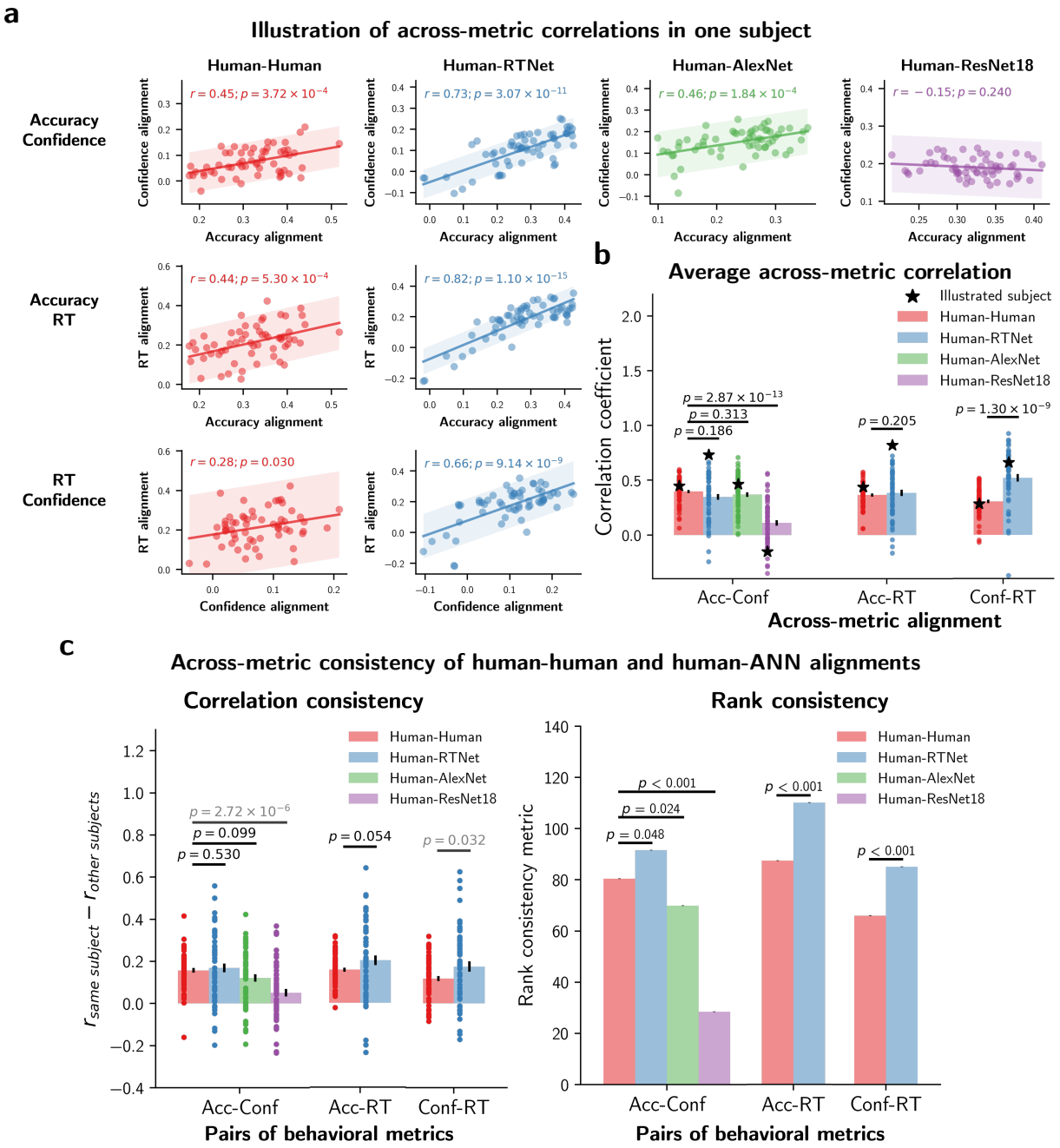


Figure 3. ANN instances that mimic an individual on one metric also mimic the same individual on other metrics. (a) Across-metric scatterplots for one example human subject. Each scatterplot shows alignment with the example subject (r-values) for each ANN instance (or other subject) on two of the three behavioral metrics. Dots show individual ANN instances (or all other subjects). (b) Average across-metric correlation. Human-ANNs across-metric correlation shows comparable or even higher correlation than the human-human benchmark, except for human-ResNet18, which shows lower correlation. Error bars show SEM. Dots show individual human subjects. P-values indicate the significance of pairwise differences in average across-metric correlation of all subjects. (c) Across-metric consistency of human-human and human-ANN alignments. We assessed across-metric consistency using the same methods previously applied to quantify within-metric consistency. Both methods of quantifying consistency yielded similar results: individual ANN instances map to human subjects as consistently as humans do, which is true for both RTNet and AlexNet, whereas ResNet18 shows lower consistency than the human-

human benchmark. Error bars show SEM. Dots show individual human subjects. P-values for correlation consistency indicates the significance of pairwise differences in the adjusted correlation between human-human and human-ANN alignment, computed using paired t-tests. P-values for rank consistency indicates the significance of pairwise differences in the rank consistency metric.

As before, we confirmed that these across-metric mappings reflect robust relationships by evaluating their consistency over 1000 iterations of random splits of the data (Figure 3c). Critically, across-metric consistency was assessed by comparing split 1 of one metric to split 2 of another, which ensures that the observed consistency cannot be attributed to shared data variance. Using Pearson correlation, we found that the human-RTNet across-metric consistency was either similar to or higher than the human-human benchmark in all pairs of behavioral metrics (accuracy-confidence: $t(59) = -0.631$, $p = 0.530$, $BF_{01} = 4.29$; accuracy-RT: $t(59) = -1.95$, $p = 0.054$, $BF_{10} = 1.07$; RT-confidence: $t(59) = -2.18$, $p = 0.032$, $BF_{10} = 1.62$). The same was true for the human-AlexNet accuracy-confidence consistency ($t(59) = 1.66$, $p = 0.099$, $BF_{01} = 1.48$), whereas human-ResNet18 showed significantly lower consistency than to the human benchmark ($t(59) = 4.97$, $p = 2.72 \times 10^{-6}$, $BF_{10} = 6.68 \times 10^{4}$). Similar results were also obtained using rank order. However, using the rank metric, human-RTNet exhibited significantly greater consistency compared to the human-human benchmark (accuracy-confidence: $p = 0.048$, 95% bootstrap CI = [-21.6, -0.130]; accuracy-RT: $p < 0.001$, 95% bootstrap CI = [-34.1, -12.4]; RT-confidence: $p < 0.001$, 95% bootstrap CI = [-30.0, -8.35]), whereas human-AlexNet and human-ResNet18 demonstrated significantly lower consistency than to the human benchmark in the accuracy-confidence pair (AlexNet: $p < 0.001$, 95% bootstrap CI = [6.19, 28.1]; ResNet18: $p < 0.001$, 95% bootstrap CI = [42.1, 62.4]). These results show that the alignment between individual differences in humans and ANN instances extends even across metrics, suggesting that an ANN instance may mimic an individual on a deeper level than what is reflected by a single metric.

2.5 Leveraging individual differences in ANN instances to improve prediction of single-subject responses

Thus far, we demonstrated that ANN instances differing only in their random initialization exhibit

substantial individual differences that could be mapped consistently with individual human subjects across data subsets and behavioral metrics. We next asked whether this human-ANN map can be leveraged for practical applications. Specifically, if the individual differences in ANN instances truly reflect the individual differences in humans, the human-ANN mapping should also enable prediction of human behavior in held-out data. To test this hypothesis, we computed the human-ANN similarity matrix from one data subset to learn the alignment strength between each subject and ANN instance. We then used the correlations in the similarity matrix as weights to predict behavior in a held-out data subset. This procedure weighs the ANN instances in proportion to how similar they are to a given human subject. Finally, we compared this alignment-weighted prediction with an unweighted prediction that simply took averages of all ANN instances.

We evaluated the prediction performance within each behavioral metric (Figure 4). All ANNs showed significantly better alignment-weighted prediction compared to unweighted prediction. This was true for accuracy (RTNet: $r_{difference} = 0.015$, $t(59) = 5.61$, $p = 5.71 \times 10^{-7}$; AlexNet: $r_{difference} = 0.015$, $t(59) = 4.98$, $p = 5.86 \times 10^{-6}$; ResNet18: $r_{difference} = 0.006$; $t(59) = 5.89$, $p = 2.00 \times 10^{-7}$), confidence (RTNet: $r_{difference} = 0.016$; $t(59) = 5.59$, $p = 6.02 \times 10^{-7}$; AlexNet: $r_{difference} = 0.004$; $t(59) = 2.06$, $p = 0.044$; ResNet18: $r_{difference} = 0.006$; $t(59) = 5.39$, $p = 1.27 \times 10^{-6}$), and RT (RTNet: $r_{difference} = 0.007$; $t(59) = 2.73$, $p = 0.008$). While these improvements were modest (maximum improvement of 0.016 in correlation), they were remarkably consistent across different subjects. As a benchmark, we used the same procedure to assess the difference between human alignment-weighted prediction and unweighted prediction in held-out data for each subject (note that this comparison doesn't consider the raw prediction accuracy for each type of prediction, which was higher for humans than ANNs; Supplementary Figure 4). We found that both RTNet and AlexNet outperformed the human benchmark for accuracy (RTNet: $t(59) = 3.41$, $p = 0.001$; AlexNet: $t(59) = 2.99$, $p = 0.004$), confidence (RTNet: $t(59) = 6.25$, $p = 5.02 \times 10^{-8}$; AlexNet: $t(59) = 3.18$, $p = 0.002$), and RT (RTNet: $t(59) = 2.08$, $p = 0.042$). However, ResNet18 showed more improvement than the human benchmark only in confidence ($t(59) = 4.30$, $p = 6.60 \times 10^{-5}$) but not for accuracy ($t(59) = 1.97$, $p = 0.054$). These results show that the individual differences in ANNs can be leveraged for the practical purpose of improving the

prediction for how individual human subjects would respond to yet unseen stimuli.

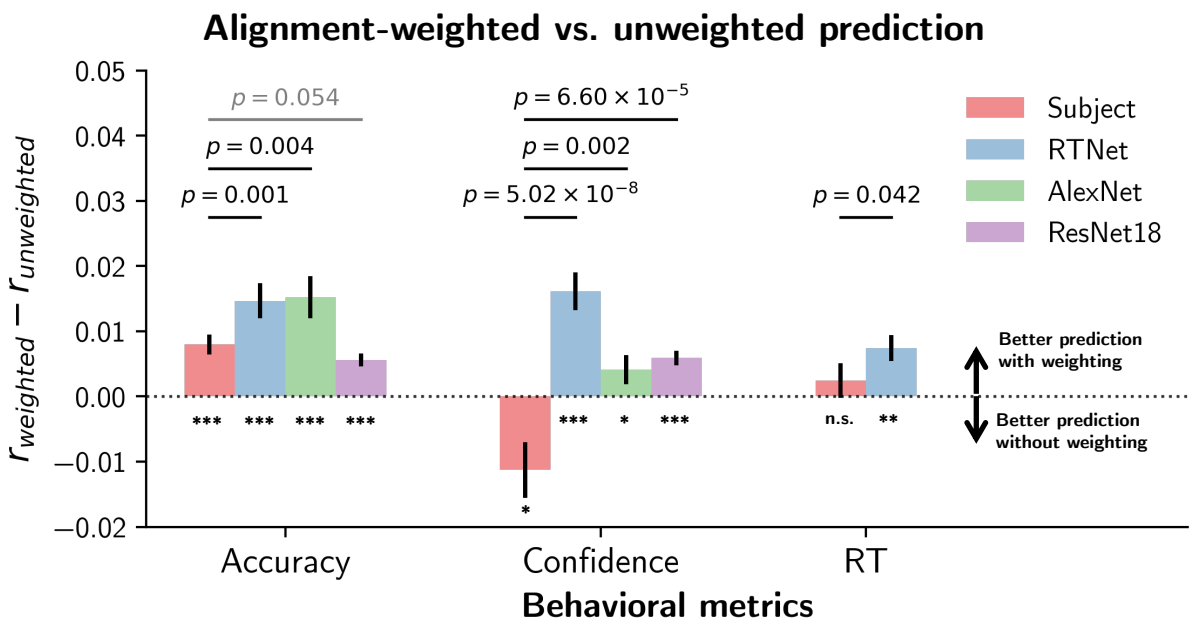


Figure 4. Leveraging individual differences in ANN instances improves prediction of individual human responses in a held-out dataset. We used the correlation in the similarity matrix as weights to predict human behavior in a held-out data subset. This alignment-weighted prediction was then compared to an unweighted average of all ANN instances. All ANNs showed significant improvement in prediction with the alignment weighting and even outperformed the human benchmark. Error bars show SEM. P-values above the bars indicate the significance of pairwise differences between humans and ANNs in the improvement from weighted versus unweighted predictions. Asterisks below the bars indicate significance for one-sample t-tests against zero. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; n.s., not significant.

2.6 Generalization to 10-choice object recognition

The results above were obtained using a relatively simple stimulus space – hand-written digits. Therefore, we further tested how well these results generalize to a more complex stimulus space comprised of naturalistic objects. Generalizing to a complex object space comes with several caveats. First, unlike for hand-written digits, ANN models are harder to train to near-perfect accuracy for common objects datasets. Thus, matching average human and ANN accuracy requires substantially lower image degradation for ANNs compared to humans. Second, the added complexity of object space (compared to digits space) means that we are likely to observe lower average alignments between

humans and ANNs. Given these factors, ANN instances are unlikely to fully capture the variability observed in human behavior when testing on objects. Nonetheless, we examined whether individual differences across ANN instances could still account for some components of human individual variability.

To this end, we conducted a preregistered experiment where N=60 human subjects performed a 10-choice blurry object recognition task (Figure 5a) and recorded accuracy, confidence, and RT on each trial. As with our MNIST analyses, we trained 60 unique instances of RTNet (Rafiei et al., 2024), AlexNet (Krizhevsky et al., 2012), and ResNet18 (He et al., 2015) by changing the random initialization. All model instances were trained in identical fashion to achieve at least 75% accuracy on a validation dataset of images with no blurring. We then adjusted the blur level in the images separately for each network architecture, so that at the population-level, the average accuracy of the 60 ANN instances from each architecture matched the average accuracy of the 60 human subjects.

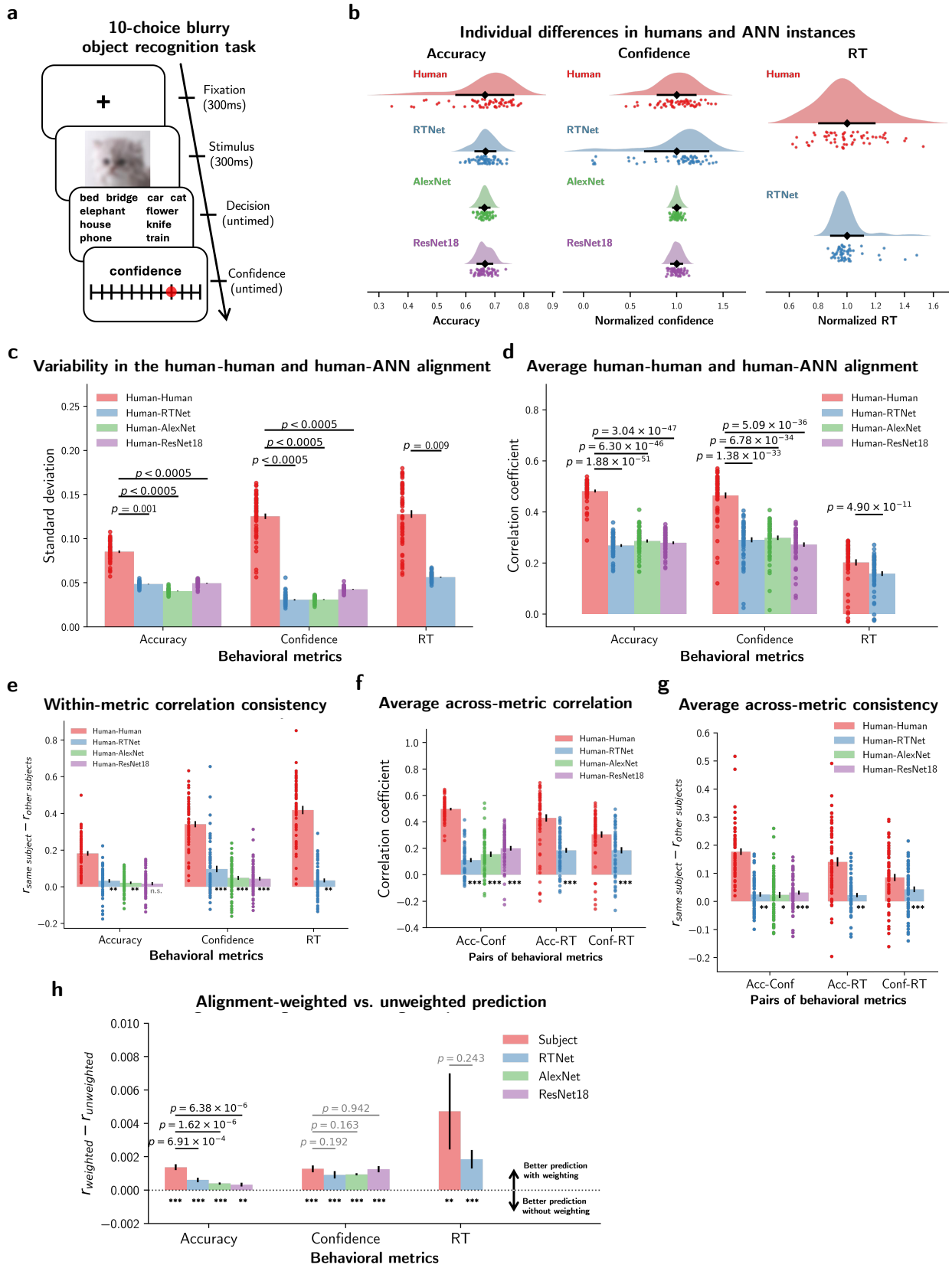


Figure 5. Generalization to the 10-choice object recognition task. (a) Task schematics for the 10-choice

blurry object recognition task. On each trial, subjects saw a blurred image of an object from the EcoSet dataset and reported their decision, followed by their confidence rating. (b) Individual differences in humans and ANN instances. Confidence is normalized by the mean across subjects or instances for visualization. Dots represent individual subjects or ANN instances. Unlike the 8-choice task, we found that humans demonstrate significantly more individual variability than ANN instances. (c) Variability in the human-human and human-ANN alignment. The variability in how well individual humans are aligned with other individual humans is significantly higher than the variability in how well they are aligned with ANN instances. P-values indicates the significance of pairwise differences in the standard deviation of the similarity matrix, computed using bootstrap resampling with 2000 data subsets. (d) Average human-human and human-ANN alignment. Individual ANN instances showed significantly lower overall alignment with humans compared to the alignment between different humans. P-values indicate the significance of pairwise differences in the average correlation. (e) Within-metric correlation consistency of human-human and human-ANN alignments. Most human-ANN alignments demonstrate significantly higher consistency than what is expected from chance, although lower than the human-human benchmark. Asterisks below the bars indicate significance for one-sample t-tests against zero. (f) Average across-metric correlation. All human-ANNs across-metric correlations are significantly higher than zero. (g) Across-metric correlation consistency of human-human and human-ANN mappings. All human-ANN mappings demonstrate significantly better across-metric consistency than chance, although lower than the human-human benchmark. (h) Alignment-weighted improvement among humans and three ANN architectures. All ANNs showed significant improvement in prediction with the individual alignment weighting compared to the unweighted control. All ANNs' improvement magnitudes were statistically indistinguishable from the human benchmark for confidence and RT but were lower for accuracy. P-values indicate the significance of pairwise differences between humans and ANNs in the improvement from weighted versus unweighted predictions. For all panels, unless otherwise noted, error bars show SEM, dots show individual human subjects, and asterisks below the bars indicate significance for one-sample t-tests against zero (*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; n.s., not significant).

We first examined whether there were substantial individual differences in both ANN instances and humans (Figure 5b). Unlike the 8-choice digit task, where ANN instances and humans showed comparable behavioral variability, ANN instances exhibited markedly reduced variability across instances compared to human subjects for both accuracy (human: range = [0.34, 0.79], SD = 0.098; RTNet: range = [0.58, 0.74], SD = 0.034; AlexNet: range = [0.63, 0.70], SD = 0.017; ResNet18: range = [0.62, 0.74], SD = 0.025) and RT (human: CV = 0.191; RTNet: CV = 0.111). An exception was confidence, for which RTNet (CV = 0.340) exhibited higher variability than humans (CV = 0.204), while AlexNet (CV = 0.038) and ResNet18 (CV = 0.058) showed substantially reduced variability. Similar results were obtained using Wasserstein distances (Supplementary Figure 5 & Supplementary Results).

We next computed the variations in the human-ANN and human-human similarity matrices. Again, diverging from the 8-choice digit task results, variability in the human-ANN alignments was significantly lower than in the human-human counterparts for both accuracy and confidence (all $p's < 0.0005$; Figure

5c). On average, we also found that the average strength of human-human correlations was significantly higher than human-ANN correlations across all three network architectures and behavioral metrics (all $p's < 0.001$; Figure 5d). These results demonstrate that in the 10-choice object recognition task, individual ANN instances that differ only in random initialization produce less variability and weaker alignment than human subjects themselves in predicting individual human performance.

To assess the robustness of the human-ANN mappings, we examined the consistency between human-human and human-ANN alignments across two subsets of the data. We found that human-RTNet alignment showed above-chance consistency for all three behavioral metrics (accuracy: $t(59) = 3.29$, $p = 0.002$; confidence: $t(59) = 5.09$, $p = 3.96 \times 10^{-6}$; RT: $t(59) = 3.07$, $p = 0.003$; Figure 5e). The same was true for human-AlexNet alignment (Accuracy: $t(59) = 3.17$, $p = 0.002$; Confidence: $t(59) = 4.28$, $p = 7.07 \times 10^{-5}$), but not for human-ResNet18 alignment, which demonstrated above-chance consistency only for confidence ($t(59) = 4.19$, $p = 9.58 \times 10^{-5}$) but not for accuracy ($t(59) = 1.92$, $p = 0.060$). Even though human-ANN alignments were generally consistent, that consistency was significantly lower than the human-human benchmark for all three behavioral metrics (all $p's < 0.001$). Similar results were obtained using rank order consistency (Supplementary Figure 6a).

We then assessed the human-human and human-ANN alignments across pairs of behavioral metrics (accuracy, confidence, and RT). We found the human-ANN across-metric correlations were significantly positive for all pairs of behavioral metrics, including accuracy-confidence (human-RTNet: average $r = 0.110$; $t(59) = 6.80$, $p = 5.83 \times 10^{-9}$; human-AlexNet: average $r = 0.155$; $t(59) = 7.79$, $p = 1.23 \times 10^{-10}$; human-ResNet18: average $r = 0.200$; $t(59) = 11.3$, $p = 1.87 \times 10^{-16}$; Figure 5f), accuracy-RT (human-RTNet: average $r = 0.184$; $t(59) = 9.92$, $p = 3.47 \times 10^{-14}$), and confidence-RT (human-RTNet: average $r = 0.185$; $t(59) = 7.68$, $p = 1.94 \times 10^{-10}$). However, all these human-ANN across-metric correlations were significantly lower than the human-human benchmark (all $p's < 0.001$). We further examined the consistency of these across-metric mappings. We found that the human-RTNet across-metric correlation showed significantly higher consistency than chance in all pairs of behavioral metrics (accuracy-confidence: $t(59) = 3.15$, $p = 0.003$;

accuracy-RT: $t(59) = 2.90$, $p = 0.005$; RT-confidence: $t(59) = 4.16$, $p = 1.05 \times 10^{-4}$; Figure 5g). The same was true for both human-AlexNet and human-ResNet18 accuracy-confidence correlation consistency (AlexNet: $t(59) = 2.14$, $p = 0.036$; ResNet18: $t(59) = 4.16$, $p = 1.06 \times 10^{-4}$). Yet, the strengths of these mappings were lower than the human-human benchmark for all three pairs of behavioral metrics (all p's < 0.001). Similar results were also obtained using rank order (Supplementary Figure 6b). Together, these results show that individual ANN instances failed to fully mimic the variability observed in human subjects, but they do reliably capture some patterns across all three behavioral metrics, revealing a meaningful, albeit partial, mimicking of human behavioral variability.

Lastly, as in the earlier analysis, we tested whether leveraging the individual differences in ANN instances improves prediction of individual human responses in a held-out dataset. We evaluated the prediction performance within each behavioral metric (Figure 5h). All ANNs showed significantly better alignment-weighted predictions as compared to the unweighted predictions for accuracy (all p's < 0.004), confidence (all p's < 0.001), and RT ($p = 8.53 \times 10^{-4}$). As a benchmark, we applied the same procedure to evaluate, for each subject, the difference between alignment-weighted and unweighted predictions on held-out data. While the raw prediction accuracy for each type of prediction was higher for humans than ANNs (Supplementary Figure 7), we found that RTNet showed comparable level of improvement to the human benchmark for confidence ($t(59) = 1.32$, $p = 0.192$) and RT ($t(59) = 1.18$, $p = 0.243$). Similarly, AlexNet and ResNet18 also exhibited improvement that is statistically indistinguishable from the human benchmark for confidence (AlexNet: $t(59) = 1.41$, $p = 0.163$; ResNet18: $t(59) = 0.07$, $p = 0.942$). However, all ANN architectures demonstrated smaller improvement than the human benchmark in accuracy (all p's $<$ 0.001). These results suggest that incorporating individual differences in ANNs could yield modest but reliable gains in predicting individual human responses to unseen data. Overall, these findings corroborate the results above, showing that individual differences among ANN instances capture aspects of human behavioral variability, demonstrating partial mapping between variability in humans and ANN instances when judging blurred objects.

3 Discussion

We investigated whether the variability across ANN instances trained with different weight initialization captures the individual differences observed in human behavior. Across two experiments, we demonstrated substantial behavioral variability between ANN instances. Critically, individual differences in ANN instances reliably captured the patterns of human behavioral variability. The mapping between ANN instances and individual humans generalized not only across different image subsets but also across multiple behavioral metrics (accuracy, confidence, and RT). Notably, the mapping was particularly strong for the digit recognition task, approaching the strength of human-to-human mapping benchmarks. Overall, these results reveal that the variability in ANN instances can serve as a proxy for human behavioral variability.

One of the central goals of NeuroAI is to build neural networks that behave like humans. Efforts in this space have shown, for instance, that aligning the visual diet of neural networks with the developmental trajectory of the human visual system produces more human-like behavior, including reduced texture bias and improved robustness to visual noise (Jang & Tong, 2024; Jinsi et al., 2023; Lu et al., 2025; Vogelsang et al., 2024). Most studies in this area, however, have focused on alignment at the population level, effectively tuning ANNs to resemble the behavior of an *average* human observer. In parallel, several recent papers have demonstrated that seemingly minor factor such as weight initialization (Mehrer et al., 2020), the image order of the training image (Chow & Palmeri, 2024), and data augmentation (Jordan, 2024) can drive substantial variability in neural network within the same architecture. This work, however, primarily explored variability in internal representations, leaving open how such instance-to-instance differences relate to individual differences in human behavior. Our work extends and bridges these two lines of work by demonstrating substantial individual differences in neural networks across three behavioral metrics (accuracy, confidence, RT) and by directly evaluating human-ANN alignment at the individual level, establishing that individual differences in neural networks are not purely stochastic but can capture meaningful aspects of human behavioral variability.

Our work introduces a largely overlooked dimension of building human-like ANN systems – the ability to model individual differences among humans. Specifically, we propose that a truly human-like neural network should not only exhibit strong population-level alignment with the *average* human behavior, but also capture the structured idiosyncrasies that characterize individual humans. One way to conceptualize this distinction between average alignment and structured idiosyncrasies is in terms of the *solution space*, where the center corresponds to the average behavior, and the shape reflects structured idiosyncrasies (Figure 6). From this perspective, existing work focuses primarily on bringing the center of the ANN solution space closer to the center of the human solution space. This approach however overlooks the fact that the human solution space has a constrained structure that may or may not match the structure of the ANN solution space. As progress in building human-like neural networks has begun to plateau (Conwell et al., 2024; Feather et al., 2025; Linsley et al., 2025; McNeal et al., 2024; Ratan Murty et al., 2021), we need new criteria for human-like neural networks that go beyond average alignment. Importantly, our data suggest that the structural alignment of individual differences is dissociable from average alignment. Specifically, although the three architectures we tested (RTNet, AlexNet, ResNet18) exhibited similar average alignment to humans (Figure 2c), there was a clear hierarchy of how well the ANNs captured the structured idiosyncrasies, with RTNet performing the best and ResNet18 performing the worst. These results highlight the need to explicitly evaluate and target the structure of individual differences when building human-like models.

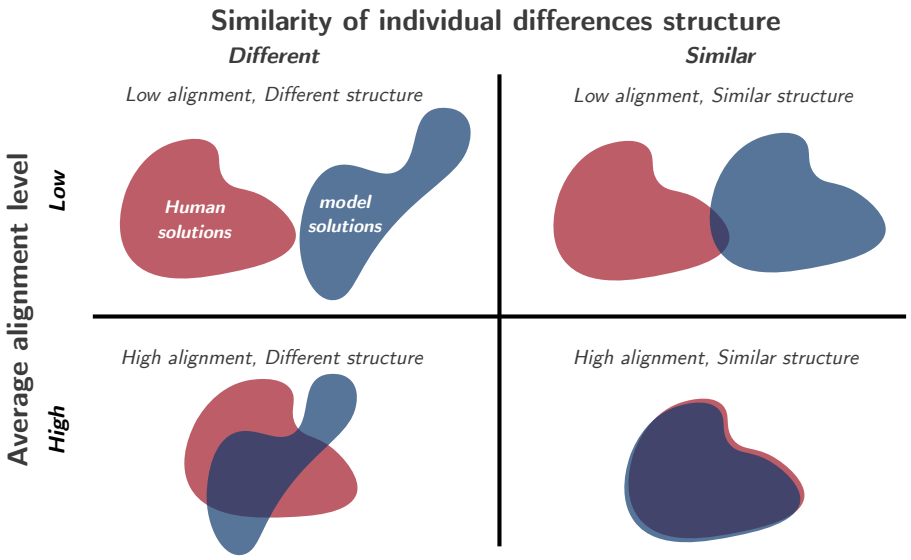


Figure 6. Illustration of the distinction between average alignment and structured individual differences. The 2 x 2 matrix illustrates the four possible scenarios in evaluating whether ANNs are human-like. Most existing work in NeuroAI focuses on average alignment by seeking to bring human solutions closer to the neural network solutions. We propose that a truly human-like neural network should not only exhibit strong average alignment with humans but also capture the structure of individual differences that characterize individual humans.

Although achieving high average alignment and a similar structure of individual differences are dissociable among different ANN architectures trained on the same task, the two objectives may be linked at the task level. Specifically, ANNs may be less human-like on some tasks than others, which would manifest in average alignment and quality of individual differences mapping changing in tandem across tasks. In our data, we observed both excellent average alignment and excellent individual differences mapping between humans and ANNs in the digit recognition task, whereas both these measures of how human-like a network is were reduced in the object recognition task. Future studies should explore a larger space of tasks to further probe whether average alignment and individual differences mapping vary together across tasks.

An open question that arises from our current study is why the ANN models were less human-like in the object recognition task. We speculate that the reason has to do with our ability to train the models to near-perfect level for undistorted images. Unlike the digit recognition task where ANNs can be easily trained

to achieve near ceiling performance for noiseless images, the object recognition task is too complex (and we lack sufficient number of training images) to achieve such near-perfect accuracy for images with no blur (typical range = 75–85% accuracy). One consequence is that because humans are near perfect for undistorted images in the object recognition task, matching accuracy between humans and ANNs necessitates the use of lower image distortion for the ANNs compared to humans. Perhaps due to this difference in average image distortion, ANNs also exhibited substantially lower variability than humans in the object recognition task for all metrics (accuracy, confidence, RT). If this explanation is correct, it would suggest that human-like individual differences for a given task would emerge more strongly when ANNs can perform at or above the human level for undistorted images in that task.

Another critical question raised by our results is why random initialization produces individual differences in ANNs that map well onto individual differences in humans. One possibility is that random weight initialization in neural networks loosely resembles the early sources of variability in human brain development. In humans, anatomical variability in neuronal organization and connectome structure emerge prenatally (Gao et al., 2014; Xu et al., 2019). Additionally, functional stochasticity may be generated by retinal waves and spontaneous cortical activity prior to any visual experience (Ackman et al., 2012; Hanganu et al., 2006; Wong, 1999). These factors introduce early variability in the developing human visual system, which could contribute to shaping the human solution space. We speculate that many aspects of variability in human visual behavior stem from these early system differences, which are broadly analogous to the stochasticity inherent in the random weight initialization of neural networks.

Our work breaks new ground by exploring human individual differences in visual behavior at the stimulus level. While individual differences in vision have been widely studied, this literature tends to focus on the overall task performance. For example, previous studies have documented individual differences in sensory encoding (Rahnev, 2021), color perception (Bosten, 2022), object recognition ability (Richler et al., 2019), etc. However, such task-level summaries provides only a coarse view of behavior: distinct individuals can have similar overall task performance but diverge substantially in their responses

to specific stimuli. For example, two subjects may both achieve 70% accuracy on the same task while making errors on entirely different stimuli. In contrast to most existing work, we examined the individual differences in accuracy, confidence, and RT for each stimulus, revealing a high-dimensional structure of individual differences. These results demonstrate that individual differences in perceptual decision making extend far beyond task-level summaries, highlighting the importance of stimulus-level analysis and broadening the landscape of research on human perceptual variability.

In conclusion, we found that neural networks capture the structured individual variability in humans. Our results demonstrate that neural networks can serve as a proxy for studying individual differences in human behavior and highlight the need to develop models that are not only human-like on average but also human-like in their idiosyncrasy.

4 Methods

4.1 Human experiments

We reanalyzed data from an existing study where subjects performed an 8-choice noisy digit recognition task (Rafiei et al., 2024) and collected new data for a 10-choice blurry object recognition task.

4.1.1 8-choice noisy digit recognition task (MNIST)

The dataset included 60 human subjects (Rafiei et al., 2024). Each trial began with the subject fixating on a white fixation cross for 500-1000 ms, followed by an image of a handwritten digit (between 1 – 8) embedded in low (0.25) or high (0.4) level of random noise, shown for 300 ms (Figure 1a). Subjects then sequentially reported their decision and confidence (4-point rating scale) with no time constraints. The stimulus set contains a total of 480 images randomly selected from the MNIST validation dataset (Deng, 2012), each tested twice for a total of 960 trials. On different trials, subjects were instructed to prioritize either speed (speed-focus) or accuracy (accuracy-focus). More details for the experiment and data preprocessing procedures can be found in the original publication.

4.1.2 10-choice blurry object recognition task (EcoSet10)

This study's sample size, experimental design, included variables, hypotheses, and planned analyses were preregistered on the Open Science Framework (https://osf.io/n6m7b) prior to any data being collected.

As we preregistered, we targeted a total of 60 subjects after applying the exclusion criteria. A total of 64 subjects (38 females; age range = 18-26, mean = 19.4, SD = 1.46) were recruited from the student body at Georgia Tech. Per our preregistration, we excluded four subjects whose accuracy fell below 25% or deviated more than 2.5 standard deviations from the group mean. All subjects completed the experiment online and received 1 SONA credit as compensation. Subjects were naïve to the purpose of the study, reported normal or corrected-to-normal vision, and provided informed consent. Ethical approval for the

study was obtained from the Georgia Tech Institutional Review Board (Protocol number: H21041).

The task stimuli consisted of 400 images from the EcoSet dataset (Mehrer et al., 2021). EcoSet is an ecologically motivated image dataset that incorporates linguistic corpus statistics and human ratings to define categories that are most relevant to humans. From the 565 categories in EcoSet, we selected those that met the following criteria: contains more than 4500 images in EcoSet, a concreteness rating above 4.75, and a frequency concreteness index greater than 0.5. From the resulting set, we then selected 10 categories to maximize perceptual and conceptual distinctiveness, avoiding closely related categories (e.g. dog and cat) to minimize confusion between categories. The final set had 10 categories: bed, bridge, car, cat, elephant, flower, house, knife, phone, and train. For each category, 40 unique images were randomly selected from the testing set of EcoSet to ensure the selected stimuli were novel for both human subjects and the trained models. All images were resized to 400×400 pixels for optimal viewing. Two levels of blurring were created by applying a Gaussian blur filter with a standard deviation of σ=6 and σ=10, using kernel sizes of $6\sigma + 1$. The two blur levels were randomly interleaved throughout the experiment.

The experiment was programmed in JavaScript with the jsPsych library 7.3.3 (de Leeuw, 2015). The virtual chinrest method was used to control differences in viewing distance, display size, and stimulus size (Fung et al., 2025; Li et al., 2020). Specifically, subjects were first asked to resize an image of a credit card on the screen to match the size of a physical credit card, thus adjusting for discrepancies in display size. Then, subjects fixated on a central point and tracked the lateral movement of a target in their periphery, pressing the spacebar when the target disappeared (i.e., when it entered the blind spot). The procedure was repeated three times to estimate the viewing distance, assuming the blind spot is located at 13.5° of the visual angle.

Each trial began with a black fixation cross displayed for 300 ms, followed by a stimulus shown for 300 ms (Figure 5a). Subjects categorized the image by clicking one of the 10 category labels. They then rated their confidence on an 11-point scale from 0% to 100% in 10% increments. Both the perceptual and

confidence responses had no time constraints, and the next trial began after both responses had been made.

Each of the 400 images was shown twice for a total of 800 trials, completed in the form of 4 runs, each containing 4 blocks of 50 trials. Before the main experiment, subjects completed two practice blocks. In the first block, subjects performed 10 trials with undistorted images to become familiar with the 10 categories and the task. This was followed by a second practice blocks of 20 trials where the images had the two levels of blur used in the experiment. Feedback was provided during the practice block but not during the main experiment. The full experimental procedure took about 60 minutes to complete.

We followed the same procedure as in the digit recognition task to remove individual trials with extreme RT values based on Tukey's interquartile criterion (Rafiei et al., 2024). Specifically, for each subject and blur condition, we computed the 25th (Q1) and 75th (Q3) percentiles of the response time (RT) distribution. Trials with RTs outside the range [Q1 – 1.5 × (Q3 – Q1), [Q3 + 1.5 × (Q3 – Q1)] were excluded from further analyses. This procedure resulted in the removal of 8.76% of trials per subject (range = 2.75-13.8%). Note that this trial exclusion was not preregistered but was performed for consistency in the analyses of the two datasets.

4.2 General ANN training and testing procedures

We considered ANNs of three architectures: RTNet (Rafiei et al., 2024), AlexNet (Krizhevsky et al., 2012), and ResNet18 (He et al., 2015). For each architecture, we trained 60 instances to match the number of human subjects by varying only the random seed used for weight initialization. All model instances were trained on undistorted images from the dataset corresponding to the task (MNIST and EcoSet10). We then tested on the same images that were presented to the human subjects in the experiment and adjusted the stimulus difficulty (i.e., noise level in MNIST and blur level in EcoSet10) to align population-level accuracy between humans and models. For all instances, we analyzed the categorical decisions. We

further computed the confidence of the ANNs as the logit difference between the first and second highest predicted class, a method shown to better predict human confidence than alternative approaches (Shekhar et al., 2025).

4.2.1 8-choice noisy digit recognition task (MNIST)

We tested RTNet, AlexNet and ResNet18 on the 8-choice noisy digit recognition task completed by the human subjects. All instances were trained on 60,000 noiseless MNIST training images. Prior to training, these images were resized to 227 × 227 pixels and normalized to have a mean of 0.1307 and a standard deviation of 0.3081. All models achieved at least 97% on the validation dataset with noiseless images. The last epoch of all instances was then tested on the 480 images presented to the human subjects.

4.2.2 10-choice blurry object recognition task (EcoSet10)

We tested RTNet, AlexNet and ResNet18 on the 10-choice blurry object recognition task completed by the human subjects. All instances were trained exclusively on the 10 categories tested in humans from EcoSet (EcoSet10; Mehrer et al., 2021). Prior to training, these images were resized to 227×227 pixels and normalized to the ImageNet mean (RGB means: 0.485, 0.456, 0.406) and standard deviation (RGB SDs: 0.229, 0.224, 0.225). All models were evaluated using images from the EcoSet validation dataset from the 10 trained categories. The last epoch of all instances was then tested on the 400 images presented to the human subjects.

4.3 ANN architectures

4.3.1 RTNet

RTNet is a recently developed neural network specifically designed for capturing the human characteristics of perceptual decision-making (Rafiei et al., 2024). Unlike the other two standard feedforward convolutional networks, RTNet is a Bayesian neural network with probabilistic weights and incorporates an evidence accumulation mechanism, repeatedly processing an image until the

accumulated evidence for a choice reaches a threshold. This mechanism allows RTNet to produce a model-equivalent of RT, defined as the number of repetition needed to reach the threshold. Additionally, RTNet generates stochastic decisions. Therefore, we tested the same images twice and averaged across the two repetitions to match it with the human subject testing.

For the digit recognition task, we used the published dataset of 60 RTNet instances that differ only in their random initialization during training (Rafiei et al., 2024). These instances were trained for 15 epochs with a batch size of 500, using the ELBO loss function and Adam optimizer with the default parameters (learning rate = 0.001; weight decay = 0, $\epsilon = 10^{-8}$). During testing, we matched the population-level accuracy between humans and models across the four experimental conditions (accuracy- vs. speed-focus, low- vs. high-noise) by searching stimulus noise levels from 2 to 5 in increments of 0.1, and the threshold level from 2 to 8 in increments of 0.5. The closest match to human accuracy was achieved for noise levels of 2.1 for low noise images and 4.1 for high noise images, and threshold values of 3 for the speed-focus condition and 6 for the accuracy-focus condition.

For the object recognition task, we trained 60 instances of RTNet instances that differ in their initializations. Unlike standard convolutional neural network, RTNet weight initializations were performed by different combinations of means and standard deviations, allowing each weight distribution to capture distinct prior uncertainties. Specifically, the means of the weights and biases were set to a value between 0.1 and 1.2 with 0.1 increments, and all standard deviations of the weights and biases were set to a value ranging from 1 to 5 with increments of 1 (total: 12×5=60 instances). Models that failed to converge were retrained with the standard deviation sampled from 6 to 10 with increments of 1, ensuring that none of the instances has the same initialization distribution. We applied random horizontal flipping (50%) to the image input during training to reduce overfitting and improve model generalization. All instances were trained for 150 epochs with a batch size of 512, using the ELBO loss function and Adam optimizer with learning rate = 10^{-5}; weight decay = 0, $\epsilon = 10^{-8}$. The KL divergence term in the ELBO loss function was gradually scaled using KL annealing, with the KL weight β following

a sigmoid schedule:

$$\beta = \frac{1}{1 + exp(-0.2 \cdot (epoch - 15)}$$

The gradually increasing KL divergence term enables the model to first focus on fitting the data in early training and then progressively regularize the weights according to the priors. All instances achieved above 75% accuracy in the validation dataset. During testing, we matched the population-level accuracy between humans and models across two blur conditions by adjusting the blur level (search from 0.5 to 2 in increments of 0.025). The closest match to human accuracy was achieved by $\sigma = 0.9$ and $\sigma = 1.525$ for low and high blur, respectively.

4.3.2 AlexNet

AlexNet is a standard feedforward convolutional neural network with five convolutional layers and three fully connected layers (Krizhevsky et al., 2012). For both tasks, we initialized 60 different random weight initialization using Kaiming uniform initialization and trained them independently as detailed below.

For the digit recognition task, all instances were trained for 15 epochs with a batch size of 128, using the cross-entropy loss function and the Adam optimizer with default parameters (learning rate = 0.001; weight decay = 0, $\epsilon = 10^{-8}$). During testing, since AlexNet has no inherent adjustment for threshold, we matched the population-level accuracy between humans and models across the four experimental conditions only by adjusting the noise level (search from 1 to 3.5 in increments of 0.01). The closest match to human accuracy was achieved by 1.425 and 2.7 for low and high noise in the accuracy-focus condition, and 1.75 and 2.825 for low and high noise in the speed-focus condition.

For the object recognition task, we applied random horizontal flipping (50%) to the image input during training. All instances were trained for 30 epochs with a batch size of 512, using the cross-entropy loss function and the Adam optimizer with learning rate = 10^{-5}; weight decay = 10^{-4} and $\epsilon = 10^{-8}$. The

learning rate was adjusted using a reduce-on-plateau scheduler that monitored validation loss and reduced the learning rate by a factor of 0.5 when no improvement was observed for three consecutive epochs. All instances achieved above 80% accuracy in the validation dataset. During testing, we matched the population-level accuracy between humans and models across two blur conditions by adjusting the blur level (search from 0.5 to 2 in increments of 0.025). The closest match to human accuracy was achieved by $\sigma = 1.25$ and $\sigma = 1.90$ for low and high blur, respectively.

4.3.3 ResNet18

ResNet18 is a convolutional architecture with residual connections that incorporate skip connections, allowing information to bypass intermediate layers for training deeper networks (He et al., 2015). For both tasks, we initialized 60 different random weight initialization using Kaiming uniform initialization and trained them independently as detailed below.

For the digit recognition task, all instances were trained for 15 epochs with a batch size of 128, using the cross-entropy loss function and the Adam optimizer with default parameters (learning rate = 0.001; weight decay = 0, $\epsilon = 10^{-8}$). During testing, we matched the population-level accuracy between humans and models across the four experimental conditions only by adjusting the noise level (search from 0 to 5 in increments of 0.1). The closest match to human accuracy was achieved by 0.52 and 0.86 for low and high noise in the accuracy-focus condition, and 0.6 and 0.9 for low and high noise in the speed-focus condition.

For the object recognition task, we applied random horizontal flipping (50%) and random rotation (15°) to the image input during training. All instances were trained for 30 epochs with a batch size of 512, using the cross-entropy loss function the Adam optimizer with learning rate = 10^{-5}; weight decay = 10^{-4} and $\epsilon = 10^{-8}$. The learning rate was adjusted using a reduce-on-plateau scheduler that monitored validation loss and reduced the learning rate by a factor of 0.5 when no improvement was observed for two

consecutive epochs. All instances achieved above 85% accuracy in the validation dataset. During testing, we matched the population-level accuracy between humans and models across two blur conditions by adjusting the blur level (search from 0.5 to 2 in increments of 0.025). The closest match to human accuracy was achieved by $\sigma = 1.225$ and $\sigma = 1.725$ for low and high blur, respectively.

4.4 Analyses

4.4.1 Testing the alignment of individual ANN instances to individual human subjects

To test the alignment of individual ANN instances to individual human subjects, we first organized the data into the form of a response matrix R, where each row represents an individual (human or ANN instance) and each column represents an image:

$$R \in \mathbb{R}^{N_{individuals} \times N_{images}};\ R_{ij} = \text{response of individual } i \text{ to image } j$$

For human subjects and RTNet instances (which generate stochastic decisions), we averaged their responses across the two stimulus repetitions in the response matrix. In contrast, AlexNet and ResNet18 produce deterministic outputs for a given image, so no averaging was performed for these models. Thus, for both humans and different ANN architectures, we created a response matrix for each behavioral metric (accuracy, confidence, and RT).

We then computed the pairwise correlations between the responses of each human subject and each ANN instance to obtain a similarity matrix (S) of size $N_{human} \times N_{ANN}$, where each entry is computed as the dot product of z-scored response matrix, which is mathematically equivalent to the Pearson correlation. A higher correlation values indicate better alignment between a human subject and an ANN instance:

$$S_{human-ANN} = \frac{z(R_{human}) \cdot z(R_{ANN})^{\top}}{N_{images}};\ S \in \mathbb{R}^{N_{human} \times N_{ANN}}$$

Note that to compute the similarity matrix, S, we used all images regardless of their difficulty level.

Further, the 8-choice digit recognition task included speed and accuracy-focused conditions. Since these can be seen as slightly different tasks, we computed the similarity matrix separately for each of these two conditions and then averaged them to obtain a single similarity matrix.

For benchmarking, we also computed human-human similarity matrices using the same procedure. Since the diagonal of the human-human similarity matrix is always 1 (as it corresponds to an individual being correlated with themselves), we removed the diagonal in subsequent analyses.

4.4.2 Statistical properties of the human-human and human-ANN similarity matrices

We first examined the basic statistical properties of the human-ANN and human-human similarity matrices to gain an overall understanding of how individual human subjects align with ANN instances and with one another. We computed the standard deviation of the similarity matrix to quantify variability. To enable statistical comparison between the variability of the human-ANN and human-human similarity matrices, we performed 1000 iterations of randomly splitting the images into two equal halves, yielding a total of 2000 subsets. For each subset, we computed the corresponding similarity matrix and its standard deviation. We then assessed statistical significance using a non-parametric bootstrap test. Specifically, for each bootstrap sample, we computed the difference in standard deviation between human-ANNs and human-human similarity matrix. The two-sided p-value was the proportion of bootstrap samples where the difference crossed zero, and the 2.5^{th} and 97.5^{th} percentiles provided a 95% confidence interval. For completeness, we also examined mean similarity. For each subject, we computed the average correlation with all instances or all other subjects (excluding the target subject). Differences in mean similarity between human-ANN and human-human were tested using paired-sample t-tests.

4.4.3 Consistency of mapping between ANN instances and human subjects

To ensure that the mapping between individual ANN instances and individual human subjects is not spurious, we tested the consistency of this mapping across different subsets of the data. Specifically, we

randomly split the images into two equal halves and computed the corresponding similarity matrix for each split. This allowed us to quantify the consistency of the resulting mappings across the two halves of the data. This procedure was repeated for 1000 bootstrap iterations with different random image splits to ensure our results were not specific to a particular data subset.

Consistency between the two similarity matrices was quantified in two ways: (1) by Pearson correlation, and (2) by rank. First, for Pearson correlation, we computed the correlation between the two splits within each individual subject. We then corrected this value by subtracting the average correlation between the same subject and all other subjects, which removes any correlations stemming from ANN instances that may perform uniformly well or uniformly badly in matching all human subjects. Mathematically, this is defined as:

$$r_i^{corr} = corr\left(S_{i,:}^{(1)}, S_{i,:}^{(2)}\right) - \frac{1}{N-1}\sum_{i' \neq i} corr\left(S_{i,:}^{(1)}, S_{i',:}^{(2)}\right)$$

where r_i^{corr} refers to the corrected Pearson correlation for subject i; $S_{i,:}^{(1)}$ and $S_{i,:}^{(2)}$ are the rows corresponding to subject i in the similarity matrices from splits 1 and 2, respectively; N is the total number of subjects; and the summation runs over all other subjects without self-correlations, $i' \neq i$.

The corrected correlation consistency values for each subject were then averaged across 1000 bootstrap iterations. We statistically compared the human-human and human-ANN correlation consistencies using paired t-test across subjects. In addition, we computed the Bayes factor for this comparison using the default Cauchy scale factor in the Pingouin Python package (https://pingouin-stats.org/build/html/index.html).

Since using Pearson correlation in measuring consistency could partially be driven by a few ANN instances that always show good or bad alignment with most subjects, we conducted a complementary analysis by quantifying consistency using the rank order relationship. Specifically, for each subject, we

ranked all ANN instances (or each subject except the target subject) based on their degree of alignment with the target subject. For each subject *i* and each unique pair of ANN instances (or subjects) (j_1, j_2) with $j_1 < j_2$, we computed the rank differences separately in each split:

$$\Delta_{i,j_1,j_2}^{(1)} = S_{i,j_2}^{(1)} - S_{i,j_1}^{(1)}; \quad \Delta_{i,j_1,j_2}^{(2)} = S_{i,j_2}^{(2)} - S_{i,j_1}^{(2)}$$

where $S_{i,j}^{(k)}$ denotes the ranked similarity between subject *i* and ANN instance (or subject) *j* in split *k*. These differences Δ_{i,j_1,j_2} were then multiplied across the two splits:

$$P_{i,j_1,j_2} = \Delta_{i,j_1,j_2}^{(1)} \cdot \Delta_{i,j_1,j_2}^{(2)}$$

The idea behind the multiplication is that positive products would indicate that the relative ordering of the pair of ANN instances (or subjects) is consistent across splits, which would increase the score multiplicatively, whereas negative products indicate the ordering flips between splits and are penalized. Finally, we averaged over all pairs and all subjects to obtain a single rank consistency score:

$$\text{Rank consistency} = \frac{1}{N \cdot \binom{M}{2}} \sum_{i=1}^{N} \sum_{j_1<j_2} P_{i,j_1,j_2}$$

where *N* is the total number of subjects and *M* is the total number of items being ranked: $M = N$ for the human-ANN mapping, and $M = N - 1$ for the human-human mapping.

To assess statistical significance of the differences between human-human and human-ANN mappings in the rank consistency metric, we used a non-parametric bootstrap test. Specifically, for each of 1000 bootstrap iteration, we compared the difference in the rank consistency score between human-human and human-ANN similarity matrices. The proportion of bootstrap samples with differences greater than or less than zero was used to calculate a two-sided *p*-value, and the 2.5th and 97.5th percentiles of the bootstrap differences were used to obtain a 95% confidence interval.

4.4.4 Across-metric analyses

To investigate whether the mapping consistency between individual ANN instances and individual human subjects extends across the three behavioral metrics (accuracy, confidence, and RT), we applied a procedure analogous to that described above, but computed correlations between similarity matrices obtained from pairs of behavioral metrics (e.g. accuracy vs. confidence). Statistical significance was assessed by comparing human-human and human-ANN correlations against zero using one-sample t-tests. We then performed the same mapping consistency analyses (corrected Pearson correlation and rank consistency) to confirm that these across-metric mappings reflect systematic individual differences. Critically, we evaluated the across-metric consistency by comparing the first split of data of one metric with the second split of the other metric, ensuring that the measured consistency was not driven by carryover effects within the same subset of images.

4.4.5 Predicting human behavior in held-out data

To test whether human-ANN mappings could improve behavioral predictions for unseen data, we performed a cross-data subset prediction analysis. Specifically, we computed the human-ANN similarity matrix from one data subset to learn the alignment between each subject and ANN instance. These correlations in the similarity matrix were then used as weights to combine ANN responses, predicting the human response matrix in the held-out data subset. Mathematically:

$$\hat{R}_{human} = \sum_{j} w_j \cdot R_{ANN,j}; w_j \propto S_{human-ANN,j}$$

where j refers to the index of instance, $\hat{R}_{human}$ refers to the response of ANN instance on the held-out data subset, and w_j is proportional to the human-ANN similarity matrix estimated from the training subset.

The above procedure was repeated for 1000 bootstrap iterations with different random image splits to ensure our results were not specific to a particular data subset. To assess the benefit of the alignment weighting, we compared the weighted prediction with an unweighted average prediction that simply

averaged responses across all ANN instances. We then evaluated the statistical significance of the improvement by computing the difference between the alignment-weighted and unweighted predictions and testing the difference against zero using one-sample t-tests. Additionally, these differences were compared to the human-human mapping benchmark using paired-sample t-tests.

4.5 Data and code availability

All data and codes are publicly available at https://github.com/herrickfung/midb_data_code.

The full analysis pipeline is packaged separately in a custom Python library, *IndiMap*, for analyzing and mapping individual differences between humans and artificial neural networks. The full repository is available at https://github.com/herrickfung/indimap.

5 References

Ackman, J. B., Burbridge, T. J., & Crair, M. C. (2012). Retinal waves coordinate patterned activity throughout the developing visual system. *Nature*, *490*(7419), 219–225. https://doi.org/10.1038/nature11529

Baek, S., Song, M., Jang, J., Kim, G., & Paik, S.-B. (2021). Face detection in untrained deep neural networks. *Nature Communications*, *12*(1), 7328. https://doi.org/10.1038/s41467-021-27606-9

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, *364*(6439), eaav9436. https://doi.org/10.1126/science.aav9436

Bosten, J. M. (2022). Do You See What I See? Diversity in Human Color Perception. *Annual Review of Vision Science*, *8*(Volume 8, 2022), 101–133. https://doi.org/10.1146/annurev-vision-093020-112820

Chow, J. K., & Palmeri, T. J. (2024). Manipulating and measuring variation in deep neural network (DNN) representations of objects. *Cognition*, *252*, 105920. https://doi.org/10.1016/j.cognition.2024.105920

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2024). A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, *15*(1), 9383. https://doi.org/10.1038/s41467-024-53147-y

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Deng, L. (2012). The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. *IEEE Signal Processing Magazine*, *29*(6), 141–142. https://doi.org/10.1109/MSP.2012.2211477

Feather, J., Khosla, M., Murty, N. A. R., & Nayebi, A. (2025). *Brain-Model Evaluations Need the NeuroAI Turing Test* (No. arXiv:2502.16238). arXiv. https://doi.org/10.48550/arXiv.2502.16238

Fung, H., Shekhar, M., Xue, K., Rausch, M., & Rahnev, D. (2025). Similarities and differences in the effects of different stimulus manipulations on accuracy and confidence. *Consciousness and Cognition*, *136*, 103942. https://doi.org/10.1016/j.concog.2025.103942

Gao, W., Elton, A., Zhu, H., Alcauter, S., Smith, J. K., Gilmore, J. H., & Lin, W. (2014). Intersubject variability of and genetic effects on the brain's functional connectivity during infancy. *Journal of Neuroscience*, *34*(34), 11288–11296. https://www.jneurosci.org/content/34/34/11288.short

Green, M. L., Hu, M., Denison, R. N., & Rahnev, D. (2026). Using Artificial Neural Networks to Relate External Sensory Features to Internal Decisional Evidence. *Open Mind*, *10*, 29–46. https://doi.org/10.1162/OPMI.a.317

Hanganu, I. L., Ben-Ari, Y., & Khazipov, R. (2006). Retinal Waves Trigger Spindle Bursts in the Neonatal Rat Visual Cortex. *The Journal of Neuroscience*, *26*(25), 6728–6736. https://doi.org/10.1523/JNEUROSCI.0752-06.2006

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition* (No. arXiv:1512.03385). arXiv. https://doi.org/10.48550/arXiv.1512.03385

Jang, H., & Tong, F. (2024). Improved modeling of human vision by incorporating robustness to blur in convolutional neural networks. *Nature Communications*, *15*(1), 1989. https://doi.org/10.1038/s41467-024-45679-0

Jinsi, O., Henderson, M. M., & Tarr, M. J. (2023). Early experience with low-pass filtered images facilitates visual category learning in a neural network model. *PLOS ONE*, *18*(1), e0280145. https://doi.org/10.1371/journal.pone.0280145

Jordan, K. (2024). *On the Variance of Neural Network Training with respect to Test Sets and Distributions* (No. arXiv:2304.01910). arXiv. https://doi.org/10.48550/arXiv.2304.01910

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1), 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*. https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

Li, Q., Joo, S. J., Yeatman, J. D., & Reinecke, K. (2020). Controlling for Participants' Viewing Distance in

Large-Scale, Psychophysical Online Experiments Using a Virtual Chinrest. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-019-57204-1

Linsley, D., Feng, P., & Serre, T. (2025). Better artificial intelligence does not mean better models of biology. *Trends in Cognitive Sciences*, *0*(0). https://doi.org/10.1016/j.tics.2025.11.016

Lu, Z., Thorat, S., Cichy, R. M., & Kietzmann, T. C. (2025). *Adopting a human developmental visual diet yields robust, shape-based AI vision* (No. arXiv:2507.03168). arXiv. https://doi.org/10.48550/arXiv.2507.03168

McGrath, S. W., Russin, J., Pavlick, E., & Feiman, R. (2024). How Can Deep Neural Networks Inform Theory in Psychological Science? *Current Directions in Psychological Science*.

McNeal, N., Deb, M., & Murty, A. R. (2024, October 10). *Small-scale adversarial perturbations expose differences between predictive encoding models of human fMRI responses*. UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models. https://openreview.net/forum?id=1g7HWFmvY0#discussion

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, *118*(8), e2011417118. https://doi.org/10.1073/pnas.2011417118

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., & Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature Communications*, *11*(1), 5725. https://www.nature.com/articles/s41467-020-19632-w

Mueller, K. N., Carter, M. C., Kansupada, J. A., & Ponce, C. R. (2023). Macaques recognize features in synthetic images derived from ventral stream neurons. *Proceedings of the National Academy of Sciences*, *120*(10), e2213034120. https://doi.org/10.1073/pnas.2213034120

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., & Livingstone, M. S. (2019). Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences. *Cell*, *177*(4), 999-1009.e10. https://doi.org/10.1016/j.cell.2019.04.005

Rafiei, F., Shekhar, M., & Rahnev, D. (2024). The neural network RTNet exhibits the signatures of human perceptual decision-making. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-024-

01914-8

Rahnev, D. (2021). Response Bias Reflects Individual Differences in Sensory Encoding. *Psychological Science*, *32*(7), 1157–1168. https://doi.org/10.1177/0956797621994214

Ratan Murty, N. A., Bashivan, P., Abate, A., DiCarlo, J. J., & Kanwisher, N. (2021). Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature Communications*, *12*(1), 5540. https://doi.org/10.1038/s41467-021-25409-6

Richler, J. J., Tomarken, A. J., Sunday, M. A., Vickery, T. J., Ryan, K. F., Floyd, R. J., Sheinberg, D., Wong, A. C.-N., & Gauthier, I. (2019). Individual differences in object recognition. *Psychological Review*, *126*(2), 226. https://psycnet.apa.org/record/2019-09750-002

Shahbazi, E., Ma, T., Pernuš, M., Scheirer, W., & Afraz, A. (2024). Perceptography unveils the causal contribution of inferior temporal cortex to visual perception. *Nature Communications*, *15*(1), 3347. https://doi.org/10.1038/s41467-024-47356-8

Shekhar, M., Fung, H., Saxena, K., Rafiei, F., & Rahnev, D. (2025). Using artificial neural networks to reveal the human confidence computation. *PLOS Computational Biology*, *21*(12), e1013827. https://doi.org/10.1371/journal.pcbi.1013827

Shekhar, M., & Rahnev, D. (2024). Human-like dissociations between confidence and accuracy in convolutional neural networks. *PLOS Computational Biology*, *20*(11), e1012578. https://doi.org/10.1371/journal.pcbi.1012578

Strock, A., Mistry, P. K., & Menon, V. (2025). Personalized deep neural networks reveal mechanisms of math learning disabilities in children. *Science Advances*, *11*(23), eadq9990. https://doi.org/10.1126/sciadv.adq9990

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*, 433–460. https://www.edwardfrenkel.com/turing-intelligence.pdf

Vogelsang, M., Vogelsang, L., Gupta, P., Gandhi, T. K., Shah, P., Swami, P., Gilad-Gutnick, S., Ben-Ami, S., Diamond, S., Ganesh, S., & Sinha, P. (2024). Impact of early visual experience on later usage of color cues. *Science*, *384*(6698), 907–912. https://doi.org/10.1126/science.adk9587

Walker, E. Y., Sinz, F. H., Cobos, E., Muhammad, T., Froudarakis, E., Fahey, P. G., Ecker, A. S., Reimer, J.,

Pitkow, X., & Tolias, A. S. (2019). Inception loops discover what excites neurons most using deep predictive models. *Nature Neuroscience*, *22*(12), 2060–2065. https://doi.org/10.1038/s41593-019-0517-x

Wichmann, F. A., & Geirhos, R. (2023). Are Deep Neural Networks Adequate Behavioral Models of Human Visual Perception? *Annual Review of Vision Science*, *9*(1), 501–524. https://doi.org/10.1146/annurev-vision-120522-031739

Wong, R. O. L. (1999). Retinal waves and visual system development. *Annual Review of Neuroscience*, *22*(Volume 22, 1999), 29–47. https://doi.org/10.1146/annurev.neuro.22.1.29

Wood, J. N., Pandey, L., & Wood, S. M. W. (2024). Digital Twin Studies for Reverse Engineering the Origins of Visual Intelligence. *Annual Review of Vision Science*, *10*(Volume 10, 2024), 145–170. https://doi.org/10.1146/annurev-vision-101322-103628

Xu, Y., Cao, M., Liao, X., Xia, M., Wang, X., Jeon, T., Ouyang, M., Chalak, L., Rollins, N., Huang, H., & He, Y. (2019). Development and Emergence of Individual Variability in the Functional Connectivity Architecture of the Preterm Human Brain. *Cerebral Cortex*, *29*(10), 4208–4222. https://doi.org/10.1093/cercor/bhy302

6 Supplementary Results

6.1 Wasserstein distance in humans and ANN instances

In the main text, we demonstrated substantial individual differences in humans and ANN instances based on summary statistics computed for each individual (Figure 1b). However, these summary statistics do not capture differences at the level of trial-by-trial response distributions. To directly assess distributional differences in humans and ANNs, we computed the pairwise Wasserstein distance between the trial-level response distributions for every possible pair of individual human subjects or individual ANN instances ($C_2^{60} = 1770$). Formally, for individuals i and j, the Wasserstein distance was defined as:

$$W_1(R_i, R_j) = \int_0^1 \left| F_i^{-1}(q) - F_j^{-1}(q) \right| \, dq$$

where R_i and R_j denote the trial-by-trial response from individual i and j, respectively. $F_i^{-1}(q)$ and $F_j^{-1}(q)$ are the quantile functions (inverse cumulative distribution functions), giving the values at the q-th percentile of each individual's trial responses. In other words, the Wasserstein distance measures the average difference between two individual's response across all percentiles, with larger values indicating larger variability.

For the 8-choice digit recognition task, we found that RTNet and AlexNet generally showed slightly greater individual variability than humans for accuracy (Human: 0.091 ± 0.076; RTNet: 0.143 ± 0.098; AlexNet: 0.177 ± 0.143; Supplementary Figure 1) and confidence (Human: 0.150 ± 0.112; RTNet: 0.403 ± 0.326; AlexNet: 0.424 ± 0.305), whereas ResNet18 exhibited similar level of variability (Accuracy: 0.101 ± 0.077; Confidence: 0.138 ± 0.078). For RT, however, RTNet showed slightly smaller variability than humans (Human: 0.170 ± 0.112; RTNet: 0.133 ± 0.089). Overall, these results replicate the results from Figure 1b by demonstrating that ANNs that only differ in their random initialization exhibit substantial individual differences that either match or exceed the variability among human subjects.

For the 10-choice object recognition task, ANN instances generally showed markedly reduced variability as compared to humans for both accuracy (Human: 0.104 ± 0.093; RTNet: 0.040 ± 0.0.028; AlexNet: 0.020 ± 0.014; ResNet18: 0.029 ± 0.021; Supplementary Figure 5) and RT (Human: 0.225 ± 0.157; RTNet: 0.105 ± 0.119). The only exception was confidence, for which RTNet demonstrated higher variability than humans (Human: 0.264 ± 0.156; RTNet: 0.375 ± 0.316), while AlexNet (0.062 ± 0.025) and ResNet18 (0.080 ± 0.040) showed substantially reduced variability. These results showed that ANNs had substantially less variability than the human subjects in the 10-choice object recognition task.

6.2 Comparing distributions of best-matching subjects and instances

The similarity matrices in Figure 2a and Supplementary Figure 2 demonstrate that the same human subject or same ANN instance sometimes provides the best match to the data for multiple human subjects. If individual differences in ANN instances are a good model of individual differences in human subjects, we may expect that this effect would be comparable in human-human and human-ANN similarity matrices.

To investigate whether a small subset of subjects or ANN instances dominated the alignment with most human subjects, we analyzed the distribution of best-match instances (or subjects) across repeated splits of the data. Specifically, we performed 1000 iterations of randomly splitting the images into two equal halves, yielding a total of 2000 subsets. For each subset, we computed a similarity matrix and identified, for each subject, the ANN instance (or subject) which has the highest similarity. We then counted how often each ANN instance (or subject) was selected as the best match across all human subjects. These counts were ranked in descending order to obtain the best-match frequency distribution (Supplementary Figure 3a), which we modeled with an exponential decay function:

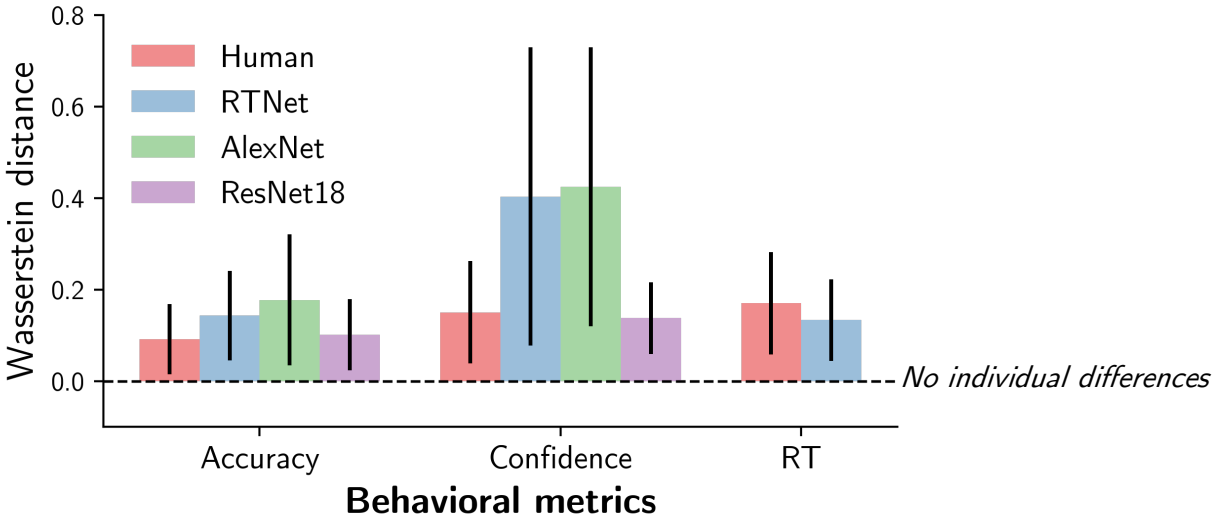
$$f(x) = Ae^{-\lambda x}$$

where A is a scaling factor, and λ is the decay rate, indicating how quickly best-match frequency

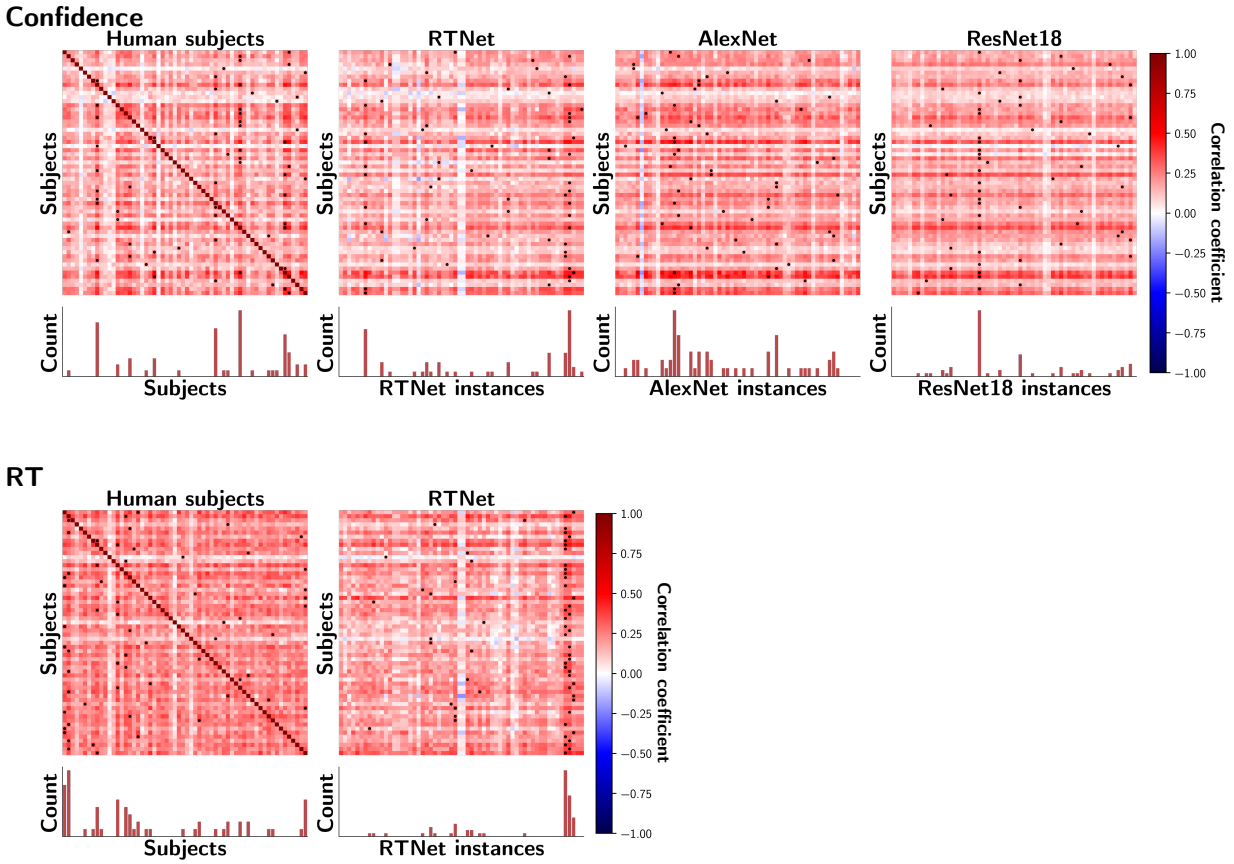
decreases down the ranked list of instances/subjects. Comparing the decay rates between humans and different ANN architectures allowed us to quantify differences in their alignment structure. Statistical significance in the decay rates was assessed using a non-parametric bootstrap test. For each bootstrap sample, we computed the difference in the decay rate between humans and ANNs. The two-sided p-value was the proportion of bootstrap samples where the difference crossed zero, and the 2.5th and 97.5th percentiles provided a 95% confidence interval.

We found that the best-match frequency distributions for humans and ANN instances have very similar decay rate (Supplementary Figure 3b). Specifically, for accuracy, the decay rate for humans was not statistically different from that of RTNet, AlexNet, or ResNet18 (all p's > 0.429). A similar pattern was observed for confidence (all p's > 0.187). For RT, however, human-RTNet best-match frequency distribution exhibited a significantly steeper decay than human-human frequency distribution ($p = 0.033$). These results demonstrate that the best-match frequency distributions, which describe the variability in the similarity of individual subjects or ANN instances to the 60 human subjects, are comparable for humans and ANN instances.

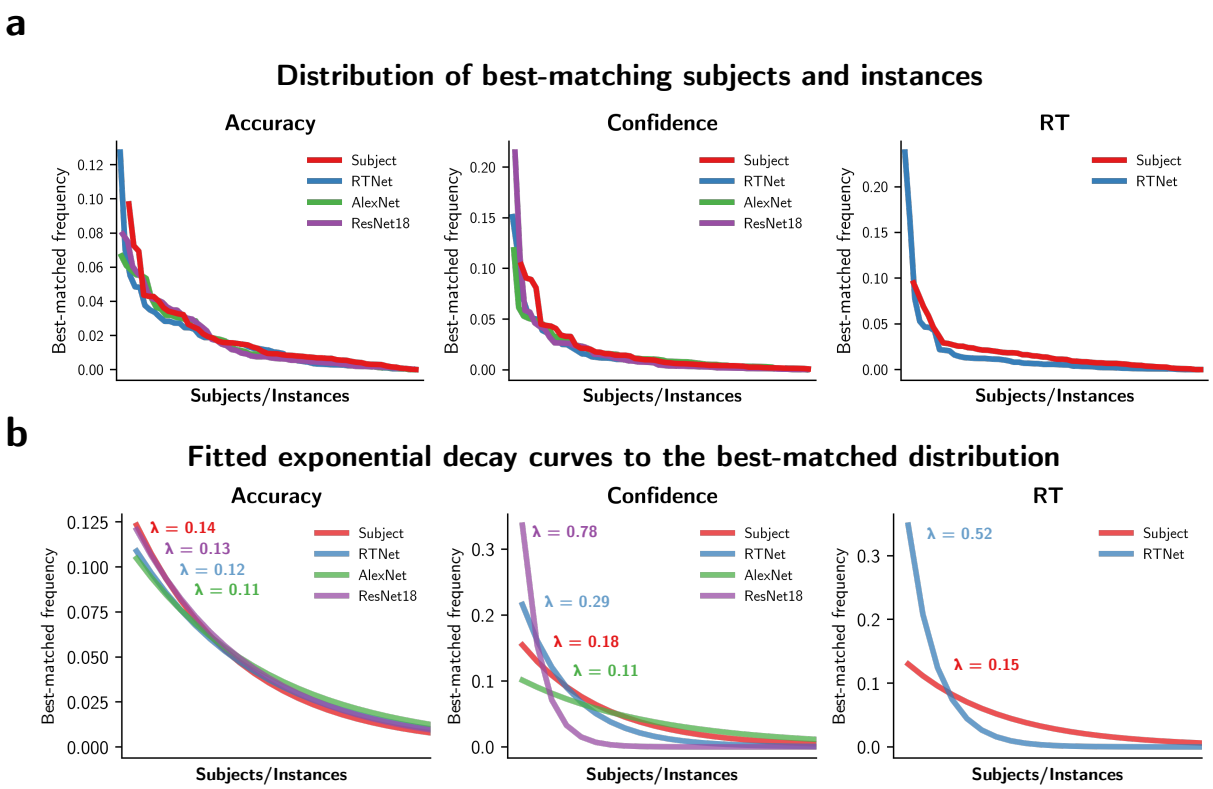
7 Supplementary Figures



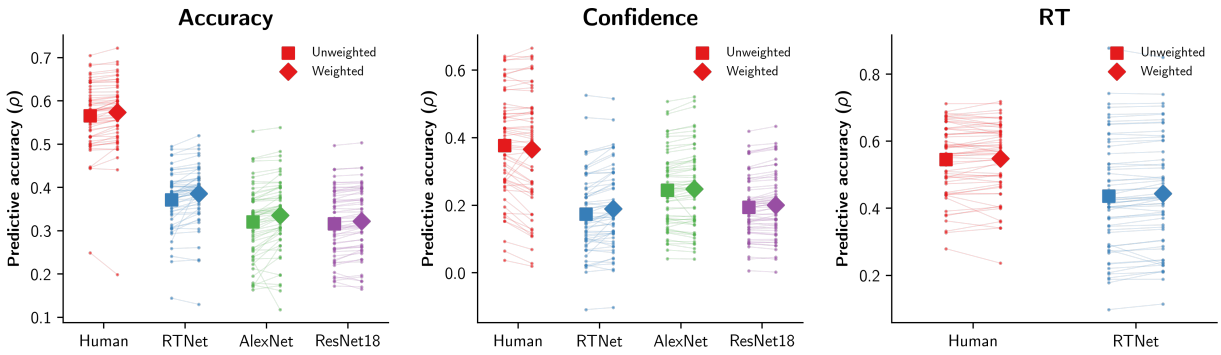
Supplementary Figure 1. Wasserstein distance for humans and ANN instances for the 8-choice digit recognition task. Figure 1b demonstrated substantial individual differences in humans and ANN instances based on summary statistics computed for each individual. However, these summary statistics do not capture differences at the level of trial-by-trial response distributions. To directly assess distributional differences, we computed the Wasserstein distances between the trial-level response distributions for each pair of human subjects and for each pair of individual ANN instances, separately for the three ANN architectures. Larger distance indicates larger individual variability. Consistent with the main text, both humans and ANNs exhibited substantial individual differences. RTNet and AlexNet generally showed slightly greater individual variability than human for accuracy and confidence, whereas ResNet18 exhibited a comparable level of variability. For RT, RTNet showed slightly lower variability than that observed in humans. Error bars show standard deviation across all subject/instance pairs. Dashed line at zero indicate the baseline with no individual differences.



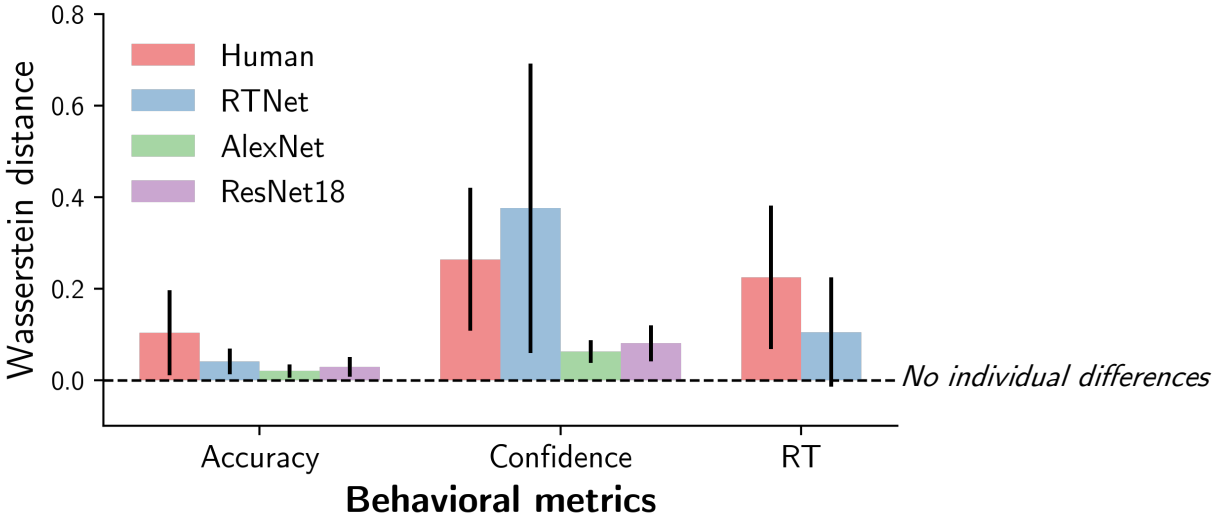
Supplementary Figure 2. Similarity between each subject and each ANN instance in confidence and RT. We constructed the similarity matrix of 60 human subjects and the 60 instances of each ANN, with higher values indicating a better alignment between an individual human and a specific ANN instance. The black dots indicate the model instance with the best fit for each subject (i.e., there is one black dot for each row). The histograms underneath show how often each subject or ANN instance was the best match across all human subjects. Consistent with the similarity matrix for accuracy as reported in Figure 2a, we observed a wide range of correlation and variability from aligning individual ANN instances to individual human subjects across three ANN architectures for confidence and RT. We also observed variability in the model instances that provided the best fit for each subject. We further explored these differences in variability in the main text and in Supplementary Figure 3.



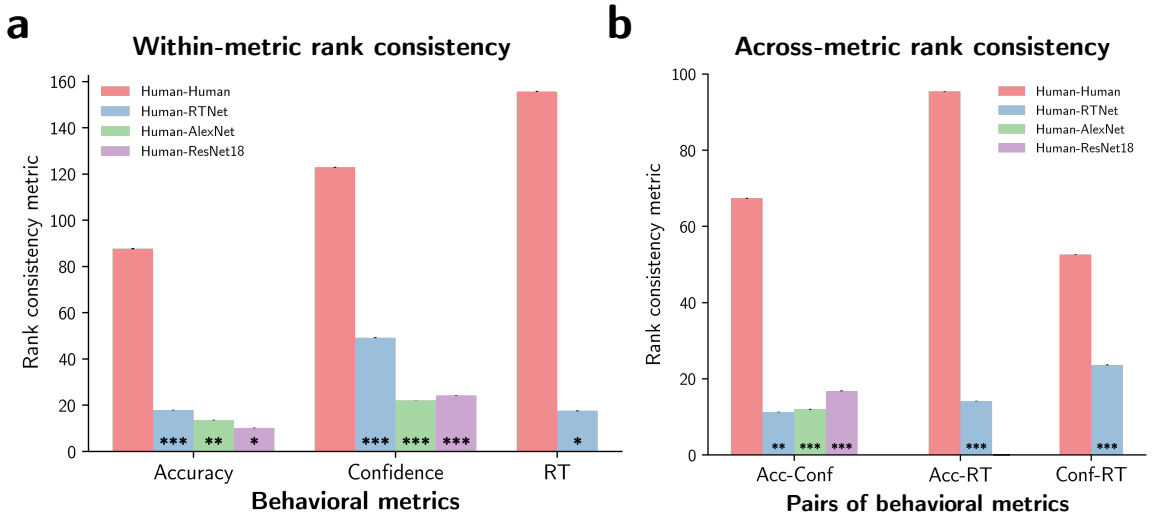
Supplementary Figure 3. Distribution of best-matching subjects and instances and its exponential decay curve fitting. (a) Distribution of best-matching subjects and instances. The figure shows the frequency of each instance (or subject) being identified as the best match in descending order, where we observed that human and ANN instances have relatively similar rate of decay. (b) The fitted exponential decay curve to the best-matched distribution. λ denotes the decay rate. The similarity matrices in Figure 2a and Supplementary Figure 2 show that the same human subject or same ANN instances sometimes provides the best-match to the data for multiple human subjects. To quantify and compare this effect between the human-human and human-ANN similarity matrices, we counted how often each subject or ANN instance was the best matched across all human subjects. We then sorted these counts in descending order to obtain "best-match" frequency distribution. We modeled these distributions using an exponential decay function, allowing us to compare the decay rate between humans and ANNs. We performed this procedure in 2000 data subsets for statistical comparison. In the case of accuracy and confidence, we found that the decay rate for humans was not statistically different from RTNet, AlexNet, and ResNet18 (all p's > 0.187). However, for RT, human-RTNet best-matching distribution shows a steeper decay than the human-human distribution ($p = 0.033$). These results demonstrate that the best-match frequency distributions, which describe the variability in the similarity of individual subjects or ANN instances to the 60 human subjects, are quite similar for humans and ANN instances.



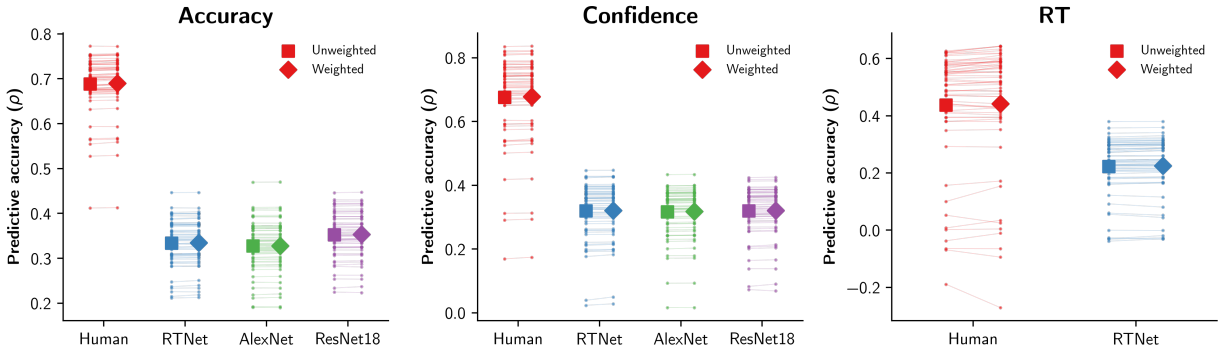
Supplementary Figure 4. Raw prediction accuracy for alignment-weighted vs. unweighted prediction for the 8-choice digit recognition task. We used the correlations in the similarity matrix as weights to predict behavior in a held-out data subset. For comparison, we computed an unweighted prediction that simply averaged across all subjects or ANN instances. Across all three behavioral metrics, human subjects consistently exhibited higher prediction accuracy than ANNs, regardless of the weighting strategy. For all ANN architectures, the alignment-weighted predictions showed statistically significant improvements over the unweighted control. Although these improvements were modest in magnitude, they were highly consistent across different subjects. We further explored these differences in improvement magnitude across humans and three ANNs in Figure 4 of the main text. Dots show individual human subjects, lines connect the alignment-weighted and unweighted predictions for the same subjects.



Supplementary Figure 5. Wasserstein distance in humans and ANN instances for the 10-choice object recognition task. Figure 5b demonstrated the individual differences in humans and ANN instances using summary statistics computed for each individual. However, these summary statistics do not capture differences at the level of trial-by-trial response distributions. To directly assess distributional differences, we computed the pairwise Wasserstein distance between the trial-level response distributions for each pair of human subjects and for each pair of individual ANN instances, separately for the three ANN architectures. Larger distance indicates larger individual variability. Consistent with the main text, we found that ANN instances showed markedly reduced variability as compared to humans for both accuracy and RT. The only exception was confidence, for which RTNet demonstrated higher variability than humans, while AlexNet and ResNet18 showed substantially reduced variability. These results showed that ANNs had substantially less variability than the human subjects in the 10-choice object recognition task. Error bars show standard deviation across all subject/instance pairs. Dashed line at zero indicate the baseline with no individual differences.



Supplementary Figure 6. Within-metric and across-metric rank consistency results in 10-choice object recognition task. (a) Within-metric rank consistency. (b) Across-metric rank consistency. In the main text, within-metric and across-metric consistency were quantified using correlation consistency metrics (Figure 5e,g). However, Pearson correlation may inflate consistency estimates if driven by a small number of ANN instances that consistently show poor alignment with most subjects, and this effect is not fully corrected by the subtraction. To address this concern, we computed consistency using rank ordering, based on how well each ANN instance (or each non-target subject) aligns with a target subject. These rank consistency results were largely consistent with the correlation consistency. All human-ANN mappings showed consistency significantly above chance, but lower than the human-human mapping benchmark, indicating that individual ANN instances capture some, but not all, patterns of human behavioral variability. Asterisks below the bars indicate significance against zero computed with 1000 bootstrapped samples (*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$).



Supplementary Figure 7. Raw prediction accuracy for alignment-weighted vs. unweighted prediction for the 10-choice object recognition task. As in the 8-choice digit recognition task, we used the correlations in the similarity matrix as weights to predict behavior in a held-out data subset. For comparison, we computed an unweighted prediction that simply averaged across all subjects or ANN instances. Across all three behavioral metrics, human subjects consistently exhibited higher prediction accuracy than ANNs, regardless of the weighting strategy. For all ANN architectures, the alignment-weighted predictions showed statistically significant improvements over the unweighted control. Although these improvements were modest in magnitude, they were highly consistent across different subjects. We further explored these differences in improvement magnitude across humans and three ANNs in Figure 5h of the main text. Dots show individual human subjects, lines connect the alignment-weighted and unweighted predictions for the same subjects.