Human-like individual differences emerge from random weight initializations in neural networks

Herrick Fung, N. Apurva Ratan Murty, Dobromir Rahnev
School of Psychology
Georgia Institute of Technology
Atlanta, GA 30332
herrickfung@gmail.com, ratan@gatech.edu, rahnev@psych.gatech.edu

Abstract

Much of AI research targets the behavior of an *average* human, a focus that traces to Turing's imitation game. Yet, no two human individuals behave exactly alike. In this study, we show that artificial neural networks (ANNs) trained with different random initializations exhibit substantial individual differences that resemble those in humans. Using a large dataset (N = 60) of human responses (accuracy, confidence, & response time) in a digit recognition task, we trained multiple instances of three ANN architectures on the same task, creating as many ANN instances as human subjects. We found that these ANN instances vary significantly from one another. Critically, ANN instances showed consistent variation in their alignment with specific human subjects. This consistency in alignment between ANN instances and humans extended across behavioral metrics, indicating that an ANN instance mimicking an individual on one metric also does so on others. Finally, we showed that leveraging these alignments improves predictions of individual human responses. Our findings highlight the potential of ANNs to capture human variability, opening new directions to develop models that go beyond aligning the *average* human and instead aligning the idiosyncratic behavior of specific *individuals*.

1 Introduction

Artificial neural networks (ANNs) have become the leading computational model of the primate visual system [1–7]. Deep convolutional architectures trained on large datasets achieve human-level performance [8, 9], demonstrate similar error patterns [10], and exhibit signatures of human perceptual decision making [11] in complex image recognition tasks. Since Turing's original proposal of the imitation game [12], much of the AI (and now NeuroAI [13]) research has been evaluated against the behavior of an *average* human, overlooking a fundamental fact: individual humans perceive, think, and behave differently. Given ANNs' success at perceptual tasks, we ask: do ANNs with the same architecture also exhibit individual differences in behavior? If so, do these individual differences in ANNs capture the range of variability observed across human individuals? Can these individual differences in ANNs serve as computational proxy models for individual differences in human behavior? Here, we show that neural networks trained with different initializations exhibit substantial individual differences that resemble those observed in humans.

Prior work has demonstrated that seemingly minor factors such as weight initialization [14] or the order of training images [15] can drive substantial variability in the features a model learns. Even though the variability in models' internal representations has been investigated, far less attention has been paid to whether different model instances produce distinct behavioral patterns. Specifically, no work to date has investigated how such individual differences in ANNs relate to the well-documented

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Data on the Brain and Mind.

individual differences in humans. Thus, it remains an open question whether the idiosyncracies of ANNs capture the spectrum of human behavioral variability.

We make three main contributions. We demonstrate that (1) ANN instances differing only in their random initialization exhibit significant yet consistent variation in alignment with specific human subjects, paralleling the variability and stability in human-human alignment (Figure 1), (2) human-ANN alignment generalizes across behavioral metrics: an ANN instance best matches an individual on one metric also tends to mimic the same individual on other metrics (Figure 2), (3) leveraging these human-ANN alignments improves predictions of individual behavioral responses in held-out data (Figure 3).

2 Methods

Details of the human and model data acquisition are provided in the Supplementary Information. To assess alignment between ANN instances and human subjects, we computed pairwise correlations between all human subjects' and ANN instances' responses to obtain a 60×60 similarity matrix for each model architecture, where higher values indicate better alignment between a human subject and an ANN instance. This procedure was repeated over 1000 bootstrap iterations with random image splits to estimate consistency (Figure 1c & 2b), calculated as the within-subject correlation across splits. We then corrected this correlation by subtracting the average across-subject correlation to account for similarity expected by shared correlations. To test whether these alignments could improve behavioral predictions for unseen data, we computed the similarity matrix from one subset to estimate the human-ANN alignments, which then served as weights to predict behavior in the held-out subset (Figure 3). We then compared this alignment-weighted prediction to an equal-weighted average prediction to assess its improvement. For benchmarking, a human-human counterpart of these analyses was performed by replacing 60 ANN instances' responses with all human subjects except the target (60 - 1), yielding a 60×59 similarity matrix that measures the alignment between the target subject and all other individual subjects except themselves.

3 Results

3.1 ANN instances show differences in alignment with specific human subjects

We first established that individual differences arise in both ANN instances and human subjects (Figure 1a). Although these ANN models were trained on noiseless images and evaluated on images with added noise and differed only in their random initialization seeds, these ANN models demonstrated a large range of accuracy ranging from 21% to 93%, with all ANN architectures demonstrating greater variability than the variability observed in human subjects (41% - 81%). Similar results were found for confidence but not RT, where we found larger variability in humans than RTNet.

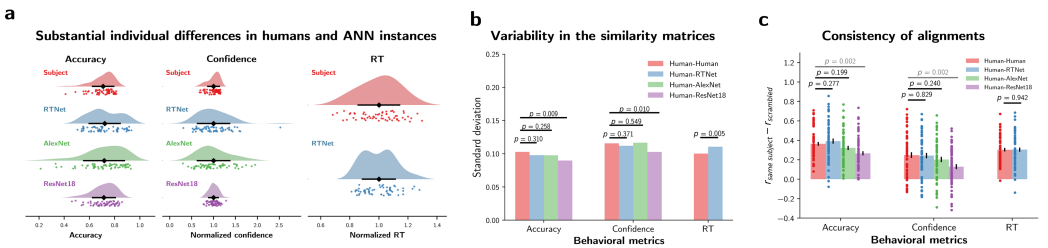

Figure 1: **ANN instances vary in alignment with specific human subjects in a consistent manner. a.** Substantial individual differences in humans and ANN instances. Confidence and RT are normalized by their mean across subjects or instances for visualization. **b.** Variability in aligning human to human and human to ANN instances. P-values: Pairwise differences tested with 2000 bootstraps. **c.** Consistency of human-human and human-ANN alignments between data subsets. Error bars: SEM. Dots: Individual human subjects. P-values: Significance of paired t-tests.

Having established individual differences in both ANNs and humans, we next explored whether individual ANN instances vary in how well they align with specific human subjects. Across all three ANN architectures, we observed substantial variability in the human-ANN similarity matrices, most of which was statistically indistinguishable from the human-human variability (all $p's > 0.257$; Figure 1b). The only exception was ResNet18 which demonstrated reliably lower variability (both $p's < 0.011$). To verify that the observed variability in human-ANN alignment was not spurious but instead reflected meaningful variability among ANN instances, we assessed human-ANN alignment consistency using separate sets of images (Figure 1c). Across 1000 bootstrap samples, human-RTNet and human-AlexNet alignments were as consistent as the human-human benchmark for all behavioral metrics (all $p's > 0.19$). In contrast, ResNet18 showed lower consistency than the human-human benchmark for both accuracy and confidence (both $p's = 0.002$). These results demonstrate that individual ANN instances of RTNet and AlexNet align reliably with specific human subjects, exhibiting variability and consistency comparable to that observed among humans themselves.

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

Having shown that ANN instances align with human subjects on each behavioral metric (accuracy, confidence, RT) individually, we next examined whether the alignment between ANN instances and humans generalizes across metrics. Across all pairs of behavioral metrics, human-ANN alignments were significantly positive (all $p's < 0.001$, Figure 2a), indicating systematic correspondence between humans and ANNs even across behavioral metrics. Moreover, except for the human-ResNet18 correlation, which was significantly worse than the human-human benchmark ($p = 2.87 \times 10^{-13}$), all other human-ANN alignments across behavioral metric pairs were indistinguishable from the human-human benchmark (each $p > 0.18$). Notably, for the case of human-RTNet and RT-confidence, it even outperformed the human-human benchmark ($p = 1.30 \times 10^{-9}$).

To ensure robustness, we validated the across-metric consistency using 1000 bootstrap samples (Figure 2b). We assessed the across-metric consistency by comparing split 1 of one metric to split 2 of another, ruling out the possibility that the observed consistency was driven by shared data variance. Human-RTNet and Human-AlexNet alignments again showed across-metric consistency comparable to the human benchmark in all metric pairs (all $p's > 0.05$), whereas human-ResNet18 alignment was significantly weaker than the human benchmark ($p = 2.72 \times 10^{-6}$). These findings indicate that the correspondence between individual differences in humans and ANN instances extends across multiple behavioral metrics, supporting that ANN instances may capture individuals' diverse behavioral characteristics at a more fundamental level.

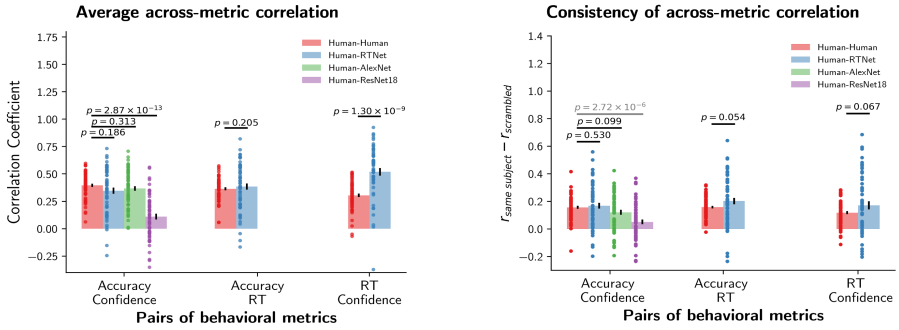

Figure 2: **ANN instances that mimic an individual on one metric also mimic the same individual on other metrics. a.** Average across-metric correlation. **b.** Across-metric consistency of human-human and human-ANN alignments. Error bars: SEM. Dots: Individual human subjects. P-values: Significance of paired t-tests.

3.3 Leveraging individual differences in ANN improves single-subject response prediction

So far, we demonstrated that neural networks trained with different initialization exhibit substantial individual differences that align consistently with individual human subjects across both data subsets

and behavioral metrics. We next asked if these human-ANN alignments can be leveraged for practical applications. Specifically, we tested whether this alignment could improve predictions of individual human behavior in held-out data. We first evaluated the within-metric prediction and found that incorporating these human-ANN alignments improved predictions for unseen data across all behavioral metrics (all $p's < 0.001$; Figure 3a). We further compared these improvements with the subject counterparts and found that all three ANNs either showed comparable improvements or improve significantly more than the subject benchmark (Figure 3a). Similar benefits were also observed in most across-metric predictions, though the effect was generally weaker as compared to the within-metric predictions (Figure 3b) but are expected because of cross-validation. These results demonstrate that leveraging individual differences in ANN instances improves single-subject response prediction for both within-metric and across-metric contexts.

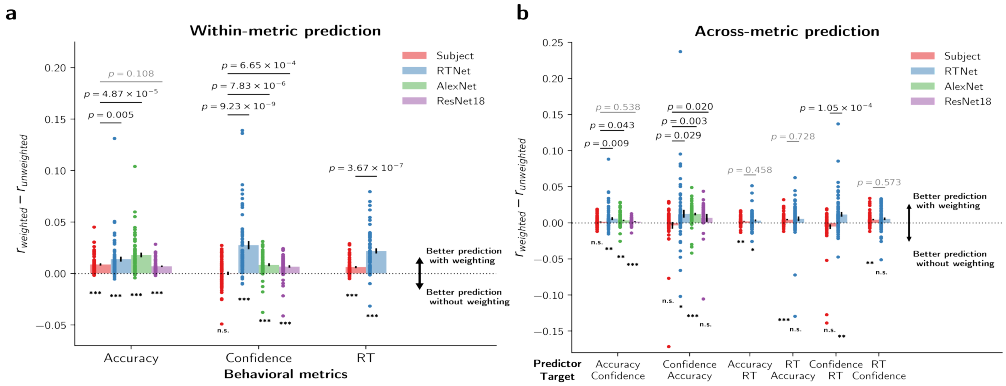

Figure 3: **Leveraging individual differences in ANN instances to improve single-subject response prediction. a.** Within-metric prediction. **b.** Across-metric prediction. Error bars: SEM. Dots: Individual human subjects. P-values: Significance of paired t-tests. Asterisks: One-sample t-test vs. zero *n.s.*: $p > 0.05$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

4 Discussion

We investigated the extent to which ANN instances that only differ in their random initializations exhibit individual differences and whether individual differences in ANN instances mimic individual differences in humans. We found that ANN instances demonstrated substantial behavioral variability. Critically, these individual differences closely resembled those observed in humans: the alignment between individual human subjects and individual ANN instances was often comparable to, and in some cases exceeded, the consistency among humans themselves (Figure 1). These alignment consistencies held across different data subsets and even behavioral metrics (Figure 2), suggesting that ANN instances can capture individual human behavior at a deeper level. Finally, we demonstrated that these human-ANN alignments could be leveraged to improve predictions of individual responses in held-out data (Figure 3).

Our findings connect to a broader trend in AI research, which has shifted from merely achieving human-level performance toward crafting models that behave more like humans. For instance, aligning the visual diet of neural networks with the developmental trajectory of the human visual system has been shown to reduce texture bias and improve robustness to different forms of visual noise [16–19], thereby bringing model perception closer to human perception. Efforts in this space have primarily evaluated success using behavioral similarity [11, 12, 16, 18], behavioral and neural predictivity (e.g. BrainScore [20]), and representational similarity [1, 21–23]. Our work introduces a complementary and largely overlooked dimension by examining the alignment between individual variability in ANNs and humans. Specifically, future work should focus not only on mimicking and explaining the average human behavior [12, 13], but also on capturing the structured idiosyncrasies in human behavior. As we show, not all network architectures capture these individual differences in human behavior. ResNet18, for instance, underperforms in mimcking human behavioral variability despite matched accuracy and variance, illustrating that not all variability is equivalent and humans have a distinctive structure to their variability. Future work should also examine how other factors,

i.e. training image order, stimulus frequency, category distribution [15], and network architecture can better capture human-like variability.

Acknowledgments and Disclosure of Funding

The authors declared no competing interests. This work was supported by the National Institute of Health (award: R01MH119189) and the Office of Naval Research (award: N00014-20-1-2622). We thank Alish Dipani for helpful suggestions and discussions.

References

[1] Nikolaus Kriegeskorte. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, 1(1):417–446, November 2015. ISSN 2374-4642, 2374-4650. doi: 10.1146/annurev-vision-082114-035447.

[2] Sam Whitman McGrath, Jacob Russin, Ellie Pavlick, and Roman Feiman. How Can Deep Neural Networks Inform Theory in Psychological Science? *Current Directions in Psychological Science*, 2024.

[3] Felix A. Wichmann and Robert Geirhos. Are Deep Neural Networks Adequate Behavioral Models of Human Visual Perception? *Annual Review of Vision Science*, 9(1):501–524, September 2023. ISSN 2374-4642, 2374-4650. doi: 10.1146/annurev-vision-120522-031739.

[4] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, June 2014. doi: 10.1073/pnas.1403112111.

[5] Daniel L K Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, March 2016. ISSN 1097-6256, 1546-1726. doi: 10.1038/nn.4244.

[6] N Apurva Ratan Murty, Pouya Bashivan, Alex Abate, James J DiCarlo, and Nancy Kanwisher. Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature communications*, 12(1):5540, 2021.

[7] Martin Schrimpf, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian, and James J DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423, 2020.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015.

[10] Rishi Rajalingham, Elias B. Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J. DiCarlo. Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *The Journal of Neuroscience*, 38(33):7255–7269, August 2018. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.0388-18.2018.

[11] Farshad Rafiei, Medha Shekhar, and Dobromir Rahnev. The neural network RTNet exhibits the signatures of human perceptual decision-making. *Nature Human Behaviour*, July 2024. ISSN 2397-3374. doi: 10.1038/s41562-024-01914-8.

[12] Alan M. Turing. Computing machinery and intelligence. *Mind*, 59:433–460, 1950.

[13] Jenelle Feather, Meenakshi Khosla, N. Apurva Ratan Murty, and Aran Nayebi. Brain-Model Evaluations Need the NeuroAI Turing Test, February 2025.

[14] Johannes Mehrer, Courtney J. Spoerer, Nikolaus Kriegeskorte, and Tim C. Kietzmann. Individual differences among deep neural network models. *Nature communications*, 11(1):5725, 2020.

[15] Jason K. Chow and Thomas J. Palmeri. Manipulating and measuring variation in deep neural network (DNN) representations of objects. *Cognition*, 252:105920, November 2024. ISSN 00100277. doi: 10.1016/j.cognition.2024.105920.

[16] Hojin Jang and Frank Tong. Improved modeling of human vision by incorporating robustness to blur in convolutional neural networks. *Nature Communications*, 15(1):1989, March 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-45679-0.

[17] Omisa Jinsi, Margaret M. Henderson, and Michael J. Tarr. Early experience with low-pass filtered images facilitates visual category learning in a neural network model. *PLOS ONE*, 18 (1):e0280145, January 2023. ISSN 1932-6203. doi: 10.1371/journal.pone.0280145.

[18] Marin Vogelsang, Lukas Vogelsang, Priti Gupta, Tapan K. Gandhi, Pragya Shah, Piyush Swami, Sharon Gilad-Gutnick, Shlomit Ben-Ami, Sidney Diamond, Suma Ganesh, and Pawan Sinha. Impact of early visual experience on later usage of color cues. *Science*, 384(6698):907–912, May 2024. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adk9587.

[19] Zejin Lu, Sushrut Thorat, Radoslaw M. Cichy, and Tim C. Kietzmann. Adopting a human developmental visual diet yields robust, shape-based AI vision. URL http://arxiv.org/abs/2507.03168.

[20] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?, January 2020.

[21] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability, November 2017.

[22] Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

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

[24] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[25] Medha Shekhar, Farshad Rafiei, and Dobromir Rahnev. Using artificial neural networks to reveal the human confidence computation. URL https://osf.io/ad8nx.

5 Supplementary Information

5.1 Human data

We reanalyzed publicly available data from an existing study [11]. In the original experiment, 60 human subjects performed an 8-choice noisy digit recognition task. The stimulus set consisted of a total of 480 unique images randomly selected from the MNIST validation dataset [24]. Each image was tested twice, producing a total of 960 trials. Each trial began with subject fixating on a white fixation cross for 500 to 1000 ms, followed by an image of a handwritten digit infused with two levels of random uniform noise, shown for 300 ms. Subjects then sequentially reported their choice and confidence (4-point rating scale) with no time constraints. The data contained accuracy, confidence, and response time (RT) on each trial.

5.2 Neural network data

We trained ANNs of three different architectures—RTNet [11], AlexNet [8], and ResNet18 [9]. For each architecture, we trained 60 unique instances (matching the number of human subjects) by changing only the random weight initialization. All models were trained on the MNIST training dataset to reach at least 97% accuracy on the validation dataset. Models were then tested on the same 480 images that were presented to the human subjects in the experiment. We matched the accuracy of all models at the group level by adjusting the noise level in the image. Confidence for all model architectures was defined as the logit margin between the first and second highest predicted classes, a method shown in a recent work to better predict human confidence than alternative approaches [25]. Below, we detail the specifics for each architecture.

RTNet: RTNet is a recently developed neural network designed to capture key features of human perceptual decision-making and predict RT [11]. Unlike standard feedforward CNNs, 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. RT can thus be defined by the number of repetitions needed to reach this threshold. We used the published dataset of 60 instances of the RTNet that differ only in their random initialization during training [11]. 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.

AlexNet & ResNet18: 60 instances of the AlexNet [8] and 60 instances of the ResNet18 [9] were trained for 15 epochs with a batch size of 128, using the cross-entropy loss function and the Adam optimizer with the default parameters.