RTNet: A neural network that exhibits the signatures of human perceptual decision making

Farshad Rafiei & Dobromir Rahnev

School of Psychology, Georgia Institute of Technology, Atlanta, GA

Keywords: Deep neural networks, reaction time, perceptual decision making, sequential sampling, confidence

Acknowledgments

This work was supported by the National Institute of Health (award: R01MH119189) and Office of Naval Research (award: N00014-20-1-2622). We thank Sashank Varma and Paul Verhaeghen for helpful suggestions about the analyses, as well as Ana Shin and Himanaga Sahithi Pandi for assistance with data collection.

Competing interests:

None

Correspondence

Farshad Rafiei

Georgia Institute of Technology

654 Cherry Str. NW

Atlanta, GA 30332

E-mail: farshad@gatech.edu

Abstract

Convolutional neural networks currently provide the best models of biological vision. However, their decision behavior, including the facts that they are deterministic and use equal number of computations for easy and difficult stimuli, differs markedly from human decision-making, thus limiting their applicability as models of human perceptual behavior. Here we develop a new neural network, RTNet, that generates stochastic decisions and human-like response time (RT) distributions, and also reproduces all foundational features of human accuracy, RT, and confidence. To test RTNet's ability to predict human behavior on novel images, we collected accuracy, RT, and confidence data from 60 human subjects performing a digit discrimination task. We found that the accuracy, RT, and confidence produced by RTNet for individual novel images correlated with the same quantities produced by human subjects. Critically, human subjects who were more similar to the average human performance were also found to be closer to RTNet's predictions. Overall, RTNet is the first neural network that exhibits all basic signatures of perceptual decision making, and therefore provides the most detailed model of all critical features of human behavior for novel images.

Introduction

Traditional cognitive models of perceptual decisions (Ratcliff, 1978; Ratcliff & McKoon, 2008) are able to account for the major features of human perceptual decision making, but do not operate on the level of images. Further, these models have been designed primarily in the context of 2-choice tasks and extending them to multi-choice decisions has often been challenging (Ratcliff, Smith, Brown, & McKoon, 2016; Trueblood, Brown, & Heathcote, 2014). Recently, convolutional neural networks (CNNs) have reached and sometimes exceeded human-level performance for novel images (Kriegeskorte, 2015; Kriegeskorte & Golan, 2019). In addition, these networks naturally handle multi-choice categorization tasks and currently provide the best models of the processing related to object recognition in the ventral visual stream of the human brain (Kietzmann, McClure, & Kriegeskorte, 2019; Kriegeskorte, 2015; Yamins & DiCarlo, 2016). However, traditional CNNs' decision behavior differs markedly from human decision behavior, thus limiting their applicability as models of human perceptual decision making. Specifically, unlike humans, traditional CNNs are both deterministic (they always give the same response for a given stimulus) and invariant in the amount of time spent on processing different images (**Figure 1A**).

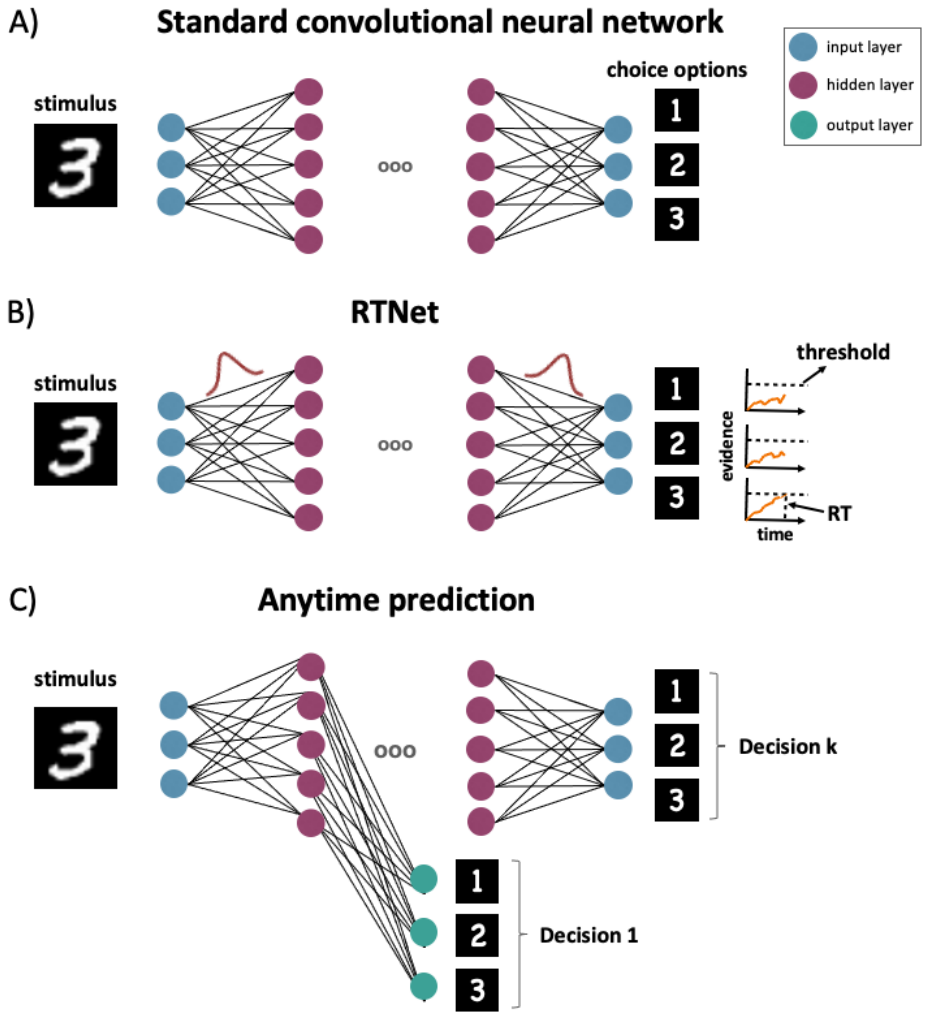

Figure 1. Model architectures. (A) A standard feedforward CNN architecture that consists of an input layer, several hidden layers, and an output layer. All images receive the same amount of processing and therefore the network cannot account for variable RT. Because all weights are fixed, the network is deterministic (i.e., it always arrives at the same response for a given stimulus). (B) RTNet architecture. Unlike standard CNNs, the connection weights in RTNet are not fixed but are instead each chosen from a distribution. A stimulus is processed multiple times by the network, each time using a different set of randomly chosen weights. The evidence from each processing step is accumulated and a decision is made when the evidence for one of the choice options reaches a predefined threshold. This architecture results in both stochastic decisions and variable RT. (C) Anytime Prediction architecture. Similar to a standard feedforward CNN, Anytime Prediction has a single input layer and several hidden layers. However, in this network, each hidden layer features its own classifier (i.e., its own output layer) allowing Anytime Prediction to make a separate decision after the processing in each layer is completed. This allows the network to stop processing an image early if that image can already be decoded from earlier layers of the network, thus resulting in different RTs for different images (though a given image still always produces the same response and RT).

Here we combine modern CNNs with traditional cognitive models to create a model that can reproduce all basic features of perceptual decision making for novel images. The model, which we call RTNet for its ability to model human RTs, features noisy weights and processes a given image several times using a different random sample of these weights in each processing step (**Figure 1B**). RTNet then accumulates the output from each processing step until a predefined threshold is reached. The model therefore has strong conceptual relationship to race models from the cognitive literature on decision-making, which postulate a noisy accumulation process with separate accumulators for each choice (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Heathcote & Matzke, 2022; Vickers, 2007).

To assess a model's ability to make decisions similar to humans, one needs to test whether it produces the foundational features of human decision-making (Forstmann, Ratcliff, & Wagenmakers, 2016). However, human perceptual decision making has been studied primarily in the context of 2-choice tasks using artificial stimuli such as Gabor patches or random dot motion (Rahnev, 2020). Therefore, we first sought to replicate the known decision-making signatures from such tasks using an 8-choice task with meaningful images (hand-written digits taken from the MNIST dataset (L. Deng, 2012)). We manipulated 1) task difficulty by adding two different levels of noise to the images, and 2) speed-accuracy trade-off (SAT) by asking subjects to emphasize either the accuracy or speed of their responses on different trials.

Critically, we tested RTNet under same conditions and with the same images seen by the human subjects to explore the model's capability to produce behavior similar to human agents. Beyond testing whether RTNet can reproduce the foundational features of human perceptual decision making, we also explored whether the accuracy, RT, and confidence produced by RTNet for individual images predicted the corresponding quantities for humans on the same images. Finally, throughout the paper, we compare the behavior of RTNet to that of Anytime Prediction (see **Figure 1C** for a depiction of the model's structure), which is currently the best model of human RT (Subramanian et al., 2022).

Results

We collected data from 60 human subjects who performed a digit discrimination task (**Figure 2A**). The experiment was a 2 x 2 design with factors of task difficulty (easy vs. difficult images) and speed pressure (speed vs. accuracy focus). Each condition consisted of 120 unique images, and each subject made a decision regarding each image exactly twice, which allowed us to determine the level of stochasticity in human behavior (**Figure 2B**). Overall, each subject completed 960 trials in total.

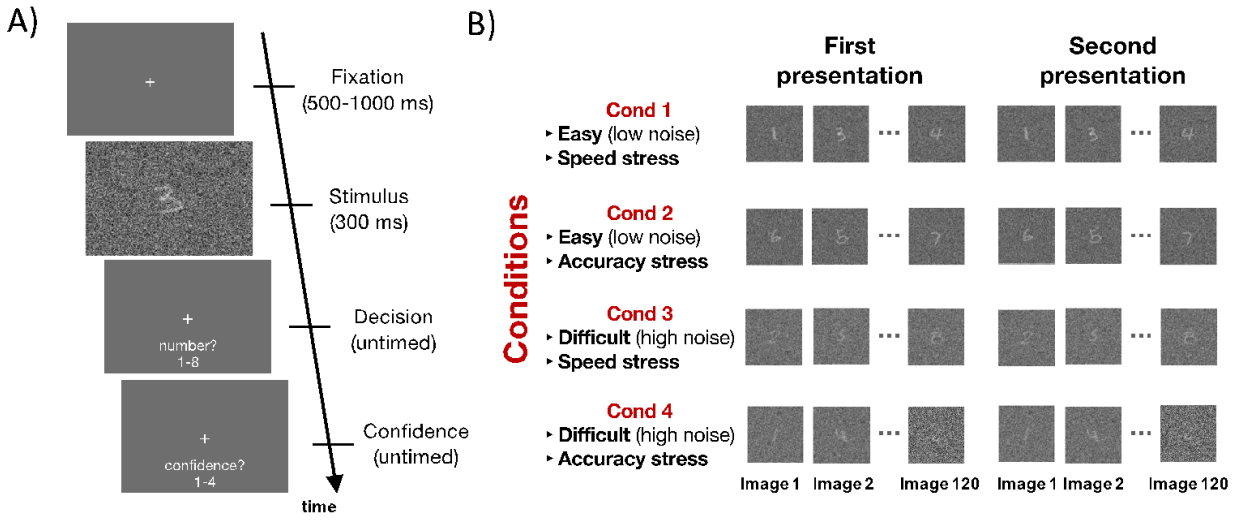

Figure 2. Experiment task. (A) Trial structure. Each trial began with a fixation cross presented for 500 to 1000 ms, followed by an image of a hand-written digit from the MNIST dataset embedded in noise and presented for 300 ms. Only the digits 1-8 were used. Subjects reported their choice and confidence (on a 4-point scale) using separate, untimed button presses. (B) Experimental design. The experiment included four conditions such that subjects judged easy (low noise) or difficult (high noise) images while emphasizing either speed or accuracy. Each condition featured 120 unique images that were the same across all subjects (total of 480 unique images in the experiment). In addition, each image was presented twice to allow the estimation of the stochasticity of human perceptual choices. Each subject thus completed a total of 960 trials. The images within the first and second sets of presentation were shown in a different random order.

Having obtained these human data, we compared the human behavior to that of RTNet and Anytime Prediction. Both networks were implemented using the eight-layer Alexnet architecture with five convolutional layers followed by three fully connected layers (Krizhevsky, Sutskever, & Hinton, 2012). Given that humans and deep learning models are impacted differently by stimulus noise (Geirhos et al., 2017, 2018), we adjusted the noise levels of the images seen by each network to match their overall accuracy to the accuracy produced by the human subjects. In addition, to allow the networks to reproduce the speed-accuracy trade-off observed in the human data, we adjusted the threshold value that triggers a decision for each model as to match the human accuracy separately in the speed- and accuracy-focused conditions. To improve the correspondence between the model predictions and the human data, we trained 60 instantiations of each model (with only changing the initial parameters before training began) and analyzed the data produced by these 60 instantiations in equivalent manner to the 60 human subjects.

Signatures of human decision-making

We examined six foundational signatures of human perceptual decision making that have already been established in studies of 2-choice tasks: 1) Human decisions are stochastic, meaning that the same stimulus can elicit different responses on different trials (Beck, Ma, Pitkow, Latham, & Pouget, 2012; Renart & Machens, 2014), 2) increasing speed stress shortens RT but decreases accuracy (speed-accuracy trade-off) (Forstmann et al., 2016; Heitz, 2014; Heitz & Schall, 2012), 3) more difficult decisions lead to reduced accuracy and longer RT (Forstmann et al., 2016; Ratcliff & Rouder, 1998; Wagenmakers & Brown, 2007), 4) RT distributions are

right-skewed, and this skew increases with task difficulty (Forstmann et al., 2016), 5) RT is lower for correct than for error trials (Brown & Heathcote, 2008; Forstmann et al., 2008; Luce, 1986; Ratcliff, 2002; Wagenmakers & Brown, 2007), and 6) confidence is higher for correct than for error trials (Rahnev, 2021). For each of these signatures, we confirmed that the signature also occurs for our 8-choice task with naturalistic images, and then tested whether RTNet and Anytime Prediction exhibit the same signature.

Stochasticity of human decisions

A central feature of human behavior is that human decisions are stochastic such that the same stimulus can elicit different responses on different trials (Beck et al., 2012; Renart & Machens, 2014; Wyart & Koechlin, 2016). We quantified the level of stochasticity in each condition by presenting each image exactly twice. On average across all conditions, 36% of all images received different responses on the two presentations (one-sample t-test: $t(59) = 36.78$, $p < 0.0001$) (**Figure 3A**). A repeated measures ANOVA with factors stimulus difficulty (easy vs. difficult) and SAT (speed vs. accuracy stress) revealed that stochasticity increased with both higher task difficulty ($F(1,63) = 871.87$, $p < 0.0001$) and higher speed pressure ($F(1,63) = 9.14$, $p = 0.0036$).

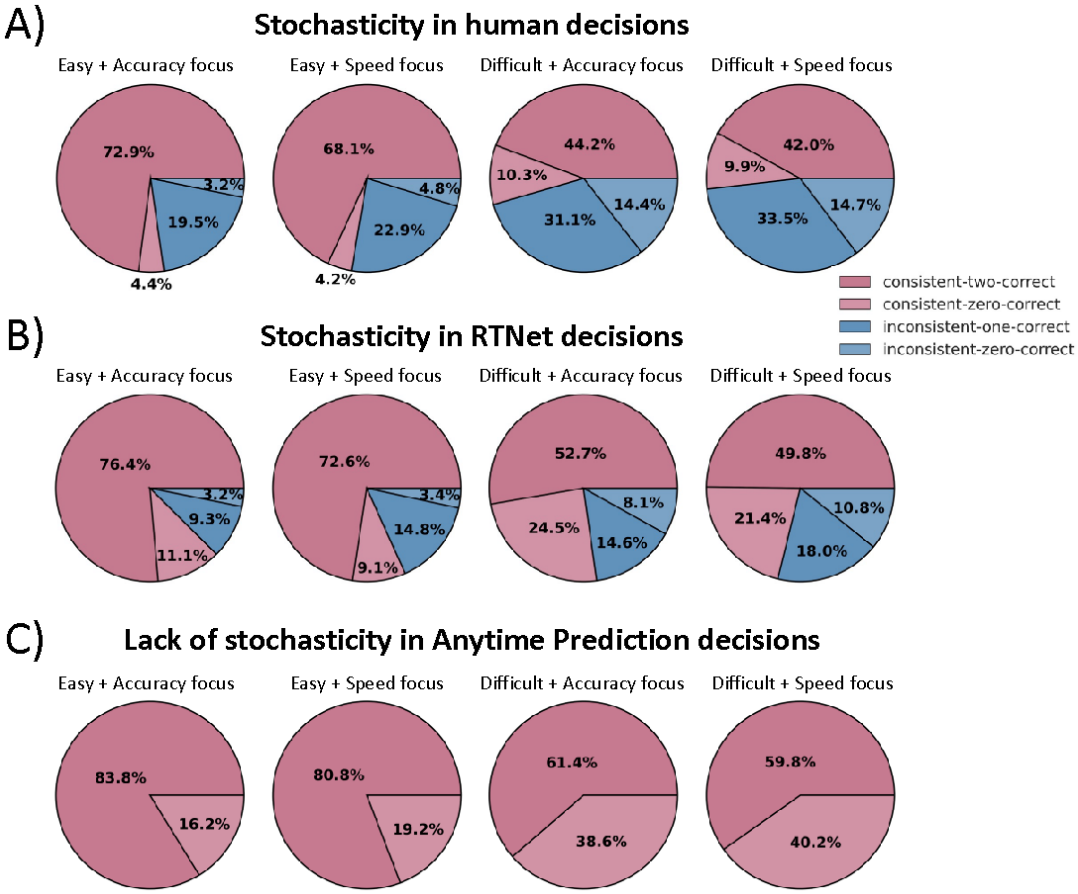

Figure 3. Decision stochasticity in humans, RTNet, and Anytime Prediction. Stochasticity of decisions made by (A) humans, (B) RTNet, and (C) Anytime Prediction. Warm colors indicate that the same response was given both times an image was presented (whether the response was correct or incorrect), whereas cool colors indicate that different responses were given for the two image presentations (whether or not any of them was correct). Humans and RTNet exhibit stochastic decision-making with stochasticity increasing with task difficulty and speed stress. However, the standard version of Anytime Prediction is fully deterministic.

Due to the fact that RTNet uses a random sample of weights for each processing step, it naturally produces stochastic decisions too. On average across all conditions, RTNet produced different responses on the two image presentations on 20% of trials ($t(59) = 32.65$, $p < 0.0001$; **Figure 3B**). This level of stochasticity was lower than for human subjects and stems from the fact that the variability in the weights was fixed a priori by training a Bayesian neural network.

However, increasing the variability of the weights can increase the stochasticity of the decisions made by RTNet. Further, the stochasticity in human decisions partially stems from factors such as fluctuations in attention, arousal, or serial dependence (Beck et al., 2012; Findling & Wyart, 2021; Renart & Machens, 2014; Wyart & Koechlin, 2016), which we did not attempt to model. Because of these considerations, we did not try to match RTNet to the exact level of human decision stochasticity observed in the data. Critically, however, RTNet exhibited the same features such that stochasticity increased with higher task difficulty ($F(1,59) = 120.12$, $p < 0.0001$) and higher speed stress ($F(1,59) = 87.73$, $p < 0.0001$). On the other hand, for a fixed level of speed-accuracy trade-off, Anytime Prediction is fully deterministic and does not exhibit any decision stochasticity, which we confirmed in our simulations (**Figure 3C**). We note that it should be possible to add noise in the weights of Anytime Prediction to induce stochastic decisions, but such noise would decrease the accuracy exhibited by Anytime Prediction much more than it affects RTNet given that only RTNet is able to average out the noise over repeated processing steps.

Speed-accuracy trade-off

The ability to trade off speed and accuracy against each other is a hallmark of decision-making across humans and many other animal species (Heitz, 2014; Heitz & Schall, 2012). The human data confirmed that increased speed pressure led to lower accuracy ($F(1,59) = 4.27$, $p = 0.0431$; **Figure 4A**) and shorter RTs ($F(1,59) = 119.29$, $p < 0.0001$; **Figure 4B**). Both models were able to replicate this pattern. Specifically, increased speed pressure resulted in lower accuracy both for RTNet ($F(1,59) = 11.93$, $p = 0.0010$) and Anytime Prediction ($F(1,59) = 21.84$, $p < 0.0001$), as well

as in shorter RT both for RTNet ($F(1,59) = 2249.86$, $p < 0.0001$) and Anytime Prediction ($F(1,59) = 584.08$, $p < 0.0001$). These results indicate that speed-accuracy trade-off is robustly observed even for relatively complex task with naturalistic images, and that both RTNet and Anytime Prediction exhibit this foundational phenomenon.

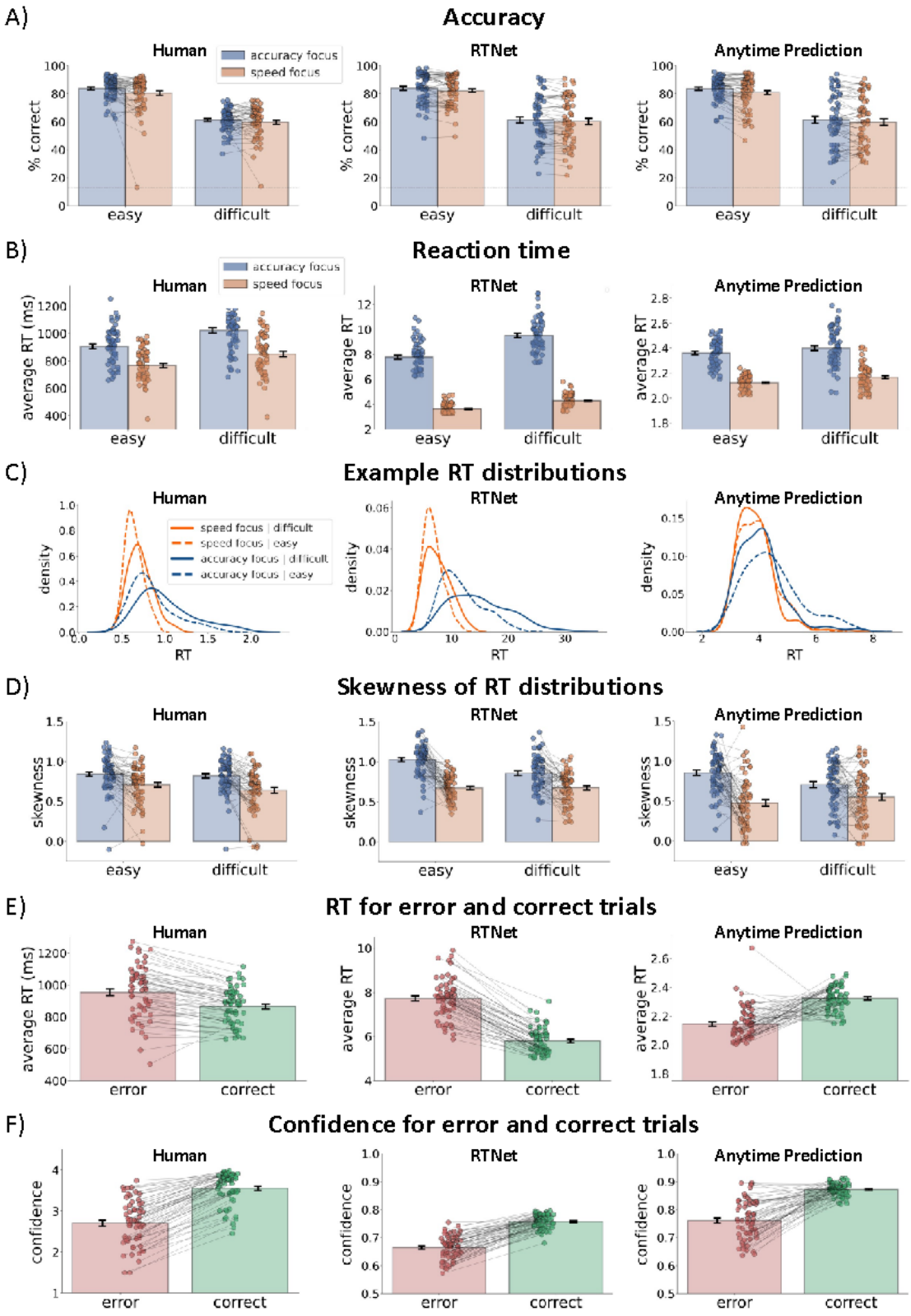

Figure 4. Behavioral effects shown by human subjects and the models. (A) Accuracy decreases when response speed is emphasized as well as for more difficult decisions. Both effects are also

exhibited by both RTNet and Anytime Prediction. (B) RT becomes shorter when response speed is emphasized, as well as for easier decisions. Both effects are also exhibited robustly by RTNet. However, while Anytime Prediction produced a robust effect for the speed manipulation, it exhibited a much smaller effects for the difficulty manipulation. (C) RT distributions for a representative subject/model. (D) The skewness of RT distributions change across conditions. For humans and RTNet, the skewness of the RT distributions was higher for easier tasks and for accuracy focus. However, while Anytime Prediction showed the same effect for accuracy focus, it failed to exhibit skewness differences between easy and difficult decisions. (E) For humans and RTNet, error trials were associated with higher RT than correct trials. However, Anytime Prediction showed the opposite pattern such that correct trials were associated with longer processing time. (F) Confidence for correct trials was higher than confidence for error trials for humans, RTNet, and Anytime Prediction. For all panels, dots represent individual subjects or different instantiations of each model; error bars show SEM.

More difficult decisions lead to reduced accuracy and longer RT

Another ubiquitous feature of decision-making is that more difficult stimuli lead to lower accuracy and longer RT (Forstmann et al., 2016; Gold & Shadlen, 2007). Our human data robustly showed this effect with more difficult stimuli leading to lower accuracy ($F(1,59) = 1558.50$, $p < 0.0001$; **Figure 4A**) and longer RT ($F(1,59) = 411.15$, $p < 0.0001$; **Figure 4B**). The same pattern was robustly observed for RTNet where difficult stimuli led to lower accuracy ($F(1,59) = 229.46$, $p < 0.0001$) but longer RT ($F(1,59) = 223.97$, $p < 0.0001$). While Anytime Prediction also showed a very robust effect on accuracy ($F(1,59) = 247.52$, $p < 0.0001$), it exhibited a much weaker effect for RT ($F(1,59) = 6.17$, $p = 0.0158$). Indeed, out of the 60 Anytime Prediction model instantiations, only 36 exhibited an RT increase for more difficult stimuli, while this effect was present in 60/60 human subjects and 57/60 RTNet instantiations. These results indicate that the effect of task difficulty on accuracy is exhibited robustly in humans, RTNet, and Anytime Prediction, but the effect of task difficulty on RT is more robust for humans and RTNet than for Anytime Prediction.

Skewness of RT distributions

For simple 2-choice decisions, human RT distributions are generally positively skewed and the skewness changes as a function of task conditions (Forstmann et al., 2016; Ratcliff & McKoon, 2008). Our 8-choice task produced RT distributions that closely resemble what is observed in standard 2-choice tasks (**Figure 4C**). Similar-looking RT distributions were produced by RTNet but Anytime Prediction produced RT distributions that, while still right-skewed, exhibited a much sharper drop-off after their peak (**Figure 4C**). We further assessed how the skewness of the RT distributions changed under different conditions. We found higher skewness for accuracy compared to speed focus ($F(1,59) = 32.84$, $p < 0.0001$), as well as for easy compared to difficult stimuli ($F(1,59) = 5.10$, $p = 0.0277$; **Figure 4D**). RTNet exhibited the same pattern with skewness increasing with a focus on accuracy ($F(1,59) = 156.71$, $p < 0.0001$) and with easier stimuli ($F(1,59) = 13.19$, $p = 0.0006$). However, while Anytime Prediction showed a similar increase in skewness with a focus on accuracy ($F(1,59) = 52.75$, $p < 0.0001$), it produced RT distributions whose skewness was unaffected by task difficulty ($F(1,59) = 2.31$, $p = 0.1336$). Overall, RTNet produced RT distributions which reflected the observed patterns in human data better than Anytime Prediction. It should be noted that Anytime Prediction can only produce distinct RTs that are less than or equal to its layer numbers, which may affect its ability to reproduce human RT distributions unless a relatively high number of layers is used. On the other hand, RTNet is capable of going through arbitrary number of samples regardless of the number of layers in its architecture.

RT is faster for correct compared to error trials

Another ubiquitous feature of human behavior in 2-choice tasks is that correct decisions are typically accompanied by faster RTs than incorrect decisions (Brown & Heathcote, 2008; Forstmann et al., 2008; Luce, 1986; Ratcliff, 2002; Wagenmakers & Brown, 2007). We replicated this effect in our 8-choice task ($F(1,59) = 82.08$, $p < 0.0001$; **Figure 4E**). The same difference between correct and error RTs also emerged for RTNet ($F(1,59) = 638.78$, $p < 0.0001$). However, Anytime Prediction exhibited the opposite pattern such that RTs were faster for error compared to correct trials ($F(1,59) = 65.70$, $p < 0.0001$). This behavior is due to the fact that errors produced by Anytime Predictions come mostly from decisions made in earlier layers. It may be possible to reverse this behavior by using a much more conservative decision threshold in the early compared to the late layers of Anytime Prediction, though the effectiveness of this strategy and its effect on all other behavioral signatures examined here would need to be tested. What is clear is that Anytime Prediction in its current form makes a qualitatively wrong prediction regarding the difference between correct and error RT, whereas RTNet naturally reproduces the empirical effect.

Confidence is higher for correct than error trials

Finally, a ubiquitous feature of confidence ratings is that they are higher for correct compared to incorrect decisions (Rahnev, 2021; Yeung & Summerfield, 2012). Our human data replicated this effect ($F(1,59) = 472.17$, $p < 0.0001$; **Figure 4F**). The effect was also robustly exhibited by both RTNet ($F(1,59) = 1021.53$, $p < 0.0001$) and Anytime Prediction ($F(1,59) = 131.92$, $p < 0.0001$). Therefore, humans, RTNet and Anytime Prediction robustly showed higher confidence for correct trials compared to incorrect trials.

Model predictions for accuracy, RT, and confidence for individual images

The results above demonstrate that RTNet is able to reproduce all foundational features of human decision-making, whereas Anytime Prediction fails to exhibit stochastic decisions, RT distribution skewness difference between easy and difficult decisions, or lower RT for correct decisions. However, RTNet's ability in those respects can easily be matched by traditional cognitive models that do not work on image-level data (Brown & Heathcote, 2008; Heathcote & Love, 2012; Heathcote & Matzke, 2022). Therefore, a critical advantage of RTNet over traditional cognitive models would be the ability to predict human behavior for individual, unseen images because traditional models cannot do that. Here we tested specifically whether the accuracy, RT, and confidence for unseen images produced by RTNet and Anytime Prediction predict the same quantities in humans.

Model predictions across all conditions

In a first set of analyses, we assessed the correlations between human accuracy, RT, and confidence and the corresponding quantities predicted by RTNet and Anytime Prediction across all four conditions (easy with speed stress, difficult with speed stress, easy with accuracy stress, difficult with accuracy stress). Because each image appeared in only one condition across all subjects, this analysis allowed us to measure the ability of the two models to predict behavioral outcomes when there is a wider range in the expected behavior, thus maximizing predictive power. In each case the model prediction was derived by averaging the behavior of the 60 instantiations of RTNet and Anytime Predictions.

We found that RTNet predicted average human behavior on individual images very well (**Figure 5A**). Specifically, we observed very high correlation between the RTNet predictions and average human accuracy ($r = 0.61$, $p < 0.0001$), average human RT ($r = 0.77$, $p < 0.0001$), and average human confidence ($r = 0.61$, $p < 0.0001$). In comparison, Anytime Prediction also predicted these quantities well (Accuracy: $r = 0.52$, $p < 0.0001$; RT: $r = 0.49$, $p < 0.0001$; Confidence: $r = 0.5$, $p < 0.0001$), but significantly worse than RTNet (Accuracy: $z(480) = 2.05$, $p = 0.0406$; RT: $z(480) = 7.48$, $p < 0.0001$; Confidence: $z(480) = 2.47$, $p = 0.0137$).

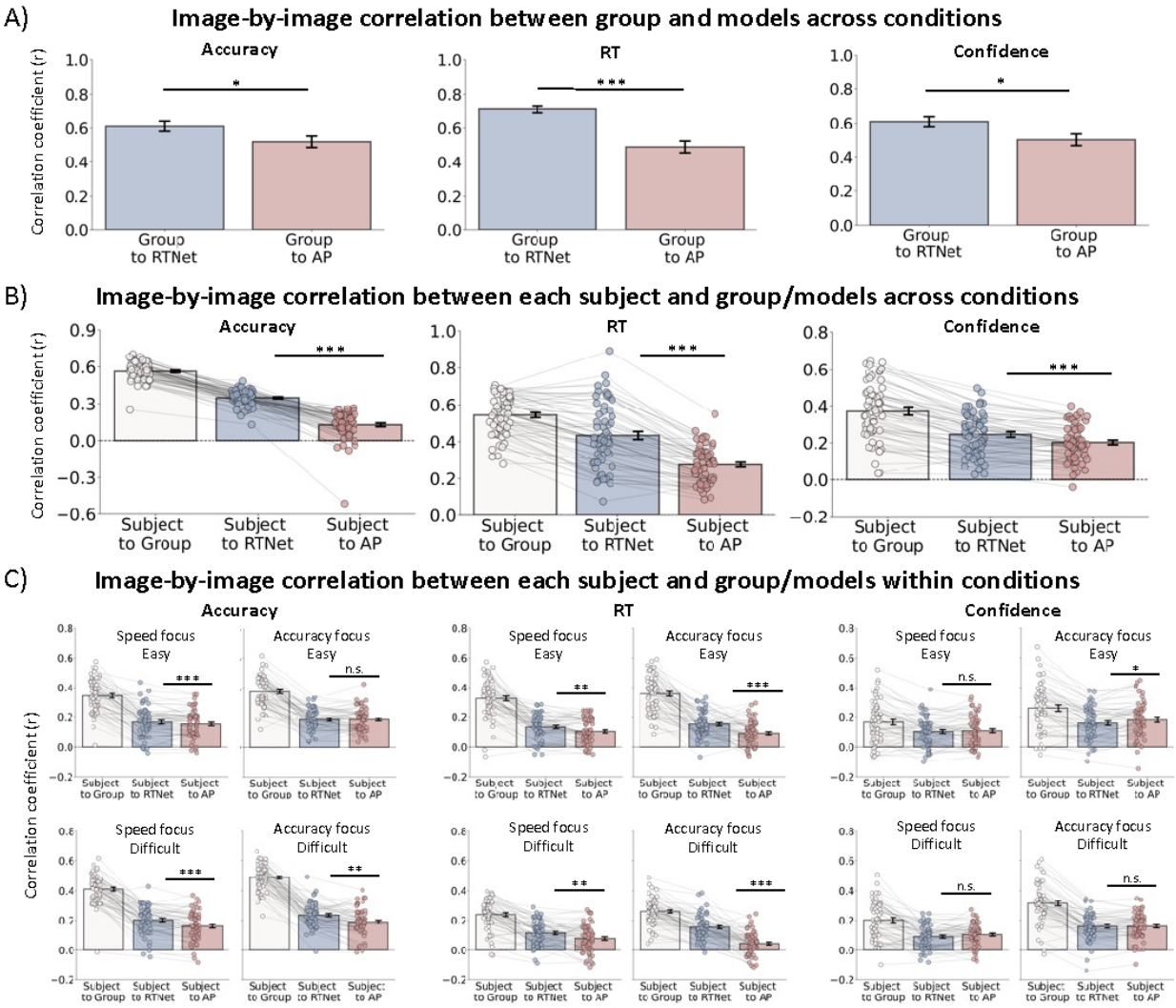

Figure 5. Image-by-image correlation between human data and each model. (A) Correlation between the average data across human subjects and RTNet or Anytime Prediction for accuracy, RT, and confidence across all conditions. The strength of correlation is stronger for RTNet for each measure. (B) Correlation for data from individual human subjects with the group average, RTNet, and Anytime Prediction for accuracy, RT, and confidence across all conditions. The strength of correlation is stronger for RTNet than Anytime Prediction for each measure. The subject-to-group correlation provides an estimate of the noise ceiling for the correlations. (C) Correlation for data from individual human subjects with the group average, RTNet, and Anytime Prediction for accuracy, RT, and confidence within each individual condition. The strength of correlation is stronger for RTNet than Anytime Prediction in seven of the 12 comparisons. For all panels, dots represent individual subjects; error bars show SEM; *p<0.05; **p<0.01; ***p<0.001; n.s., not significant; AP, Anytime Prediction.

These results demonstrate that RTNet predicts average human data better than Anytime Prediction, but they do not reveal how close RTNet comes to the ceiling for predicting human data. To explore this issue, we compared how well data from individual human subjects could be predicted by RTNet, Anytime Prediction, as well as from the data from the 59 remaining human subjects. This last quantity, which we call subject-to-group relationship, provides an estimate of the noise ceiling (i.e., the performance that a true model could achieve given inter-subject variability) (Spoerer, Kietzmann, Mehrer, Charest, & Kriegeskorte, 2020).

We found that, similar to correlating the models' predictions with average human data, both models predicted individual human data much better than chance, but with RTNet providing substantially better predictions than Anytime Prediction (**Figure 5B**). This was true for accuracy (RTNet: average r = 0.35, t(59) = 43.82, $p < 0.0001$; Anytime Prediction: average r = 0.13; t(59) = 8.52, $p < 0.0001$; Difference: t(59) = 18.07, $p < 0.0001$), RT (RTNet: average r = 0.43, t(59) = 19.60, $p < 0.0001$; Anytime Prediction: average r = 0.27; t(59) = 21.72, $p < 0.0001$; Difference: t(59) = 12.80, $p < 0.0001$), and confidence (RTNet: average r = 0.25, t(59) = 16.74, $p < 0.0001$; Anytime Prediction: average r = 0.20; t(59) = 15.84, $p < 0.0001$; Difference: t(59) = 7.48, $p < 0.0001$).

Critically, RTNet's predictions were reasonably close to the noise ceiling in all cases. Specifically, the average subject-to-group correlation for accuracy was 0.56, which means that RTNet's predictions were within 62.5% of the noise ceiling (Anytime Prediction was at 23.2%). Average subject-to-group correlations for RT and confidence were 0.54 and 0.37, meaning that RTNet's

predictions were at 79.6% and 67.6% of the noise ceiling, respectively (the same quantities for Anytime Prediction were 50% and 54.1%, respectively). Thus, by reaching to between 62.5% and 79.6% of the noise ceiling, RTNet can provide excellent predictions for the accuracy, RT, and confidence produced by human subjects for images that the model was not trained on.

Model predictions within each condition separately

The analyses above explored the correlations between model predictions and human behavior across all experimental conditions. Because different conditions vary in their average accuracy, RT, and confidence, analyses across conditions are likely to produce higher correlations than if the same analyses are to be performed within each condition separately. Therefore, we repeated the analyses above but within each of the four conditions separately to investigate if the two models can still account for accuracy, RT, and confidence on individual images. We found that both RTNet or Anytime Prediction produced accuracy, RT, and confidence predictions that significantly correlate with individual subject data in all conditions (all p's < 0.0001; **Figure 5C**). Critically, however, RTNet predicted the individual data better than Anytime Prediction in three of the four conditions for accuracy (all three p's < 0.001; only exception was accuracy focus condition with easy images, p = 0.9447) and in all four conditions for RT (all p's < 0.004). Nevertheless, Anytime Prediction performed slightly better for confidence outperforming RTNet in one of the four conditions (accuracy focus condition with easy images: p = 0.0113, all other p's > 0.05). Overall, these results demonstrate that RTNet predicts human behavior well across all three measures and across different types of analyses (across- or within-condition), and does so better than Anytime Prediction for both accuracy and RT but not for confidence.

Humans who are more similar to the group average are also more similar to RTNet

Our subject-to-group analyses revealed substantial variability in how well individual subjects' data corresponded to the group average (see **Figure 5B**). Since the group average constitutes the best model of human behavior, this variability indicates that different individuals deviate differently from the best model. Therefore, one would expect that the strength of the relationship for an individual subject and the group would be linked to the strength of the relationship of that same subject and any good model of behavior. Here we tested if such dependency holds true for RTNet and Anytime Prediction. We found that subjects who exhibited greater correlation in image-by-image accuracy across all conditions with rest of the group also exhibited greater correlation with the RTNet predictions ($r = 0.67$, $p < 0.0001$; **Figure 6A**). The same correspondence also emerged for RT ($r = 0.81$, $p < 0.0001$) and confidence ($r = 0.91$, $p < 0.0001$). Similar results were obtained for Anytime Prediction too (Accuracy: $r = 0.71$, $p < 0.0001$; RT: $r = 0.79$, $p < 0.0001$; Confidence: $r = 0.85$, $p < 0.0001$; **Figure 6B**), demonstrating that both models predict better the data from individuals who behave more similarly to the rest of the group.

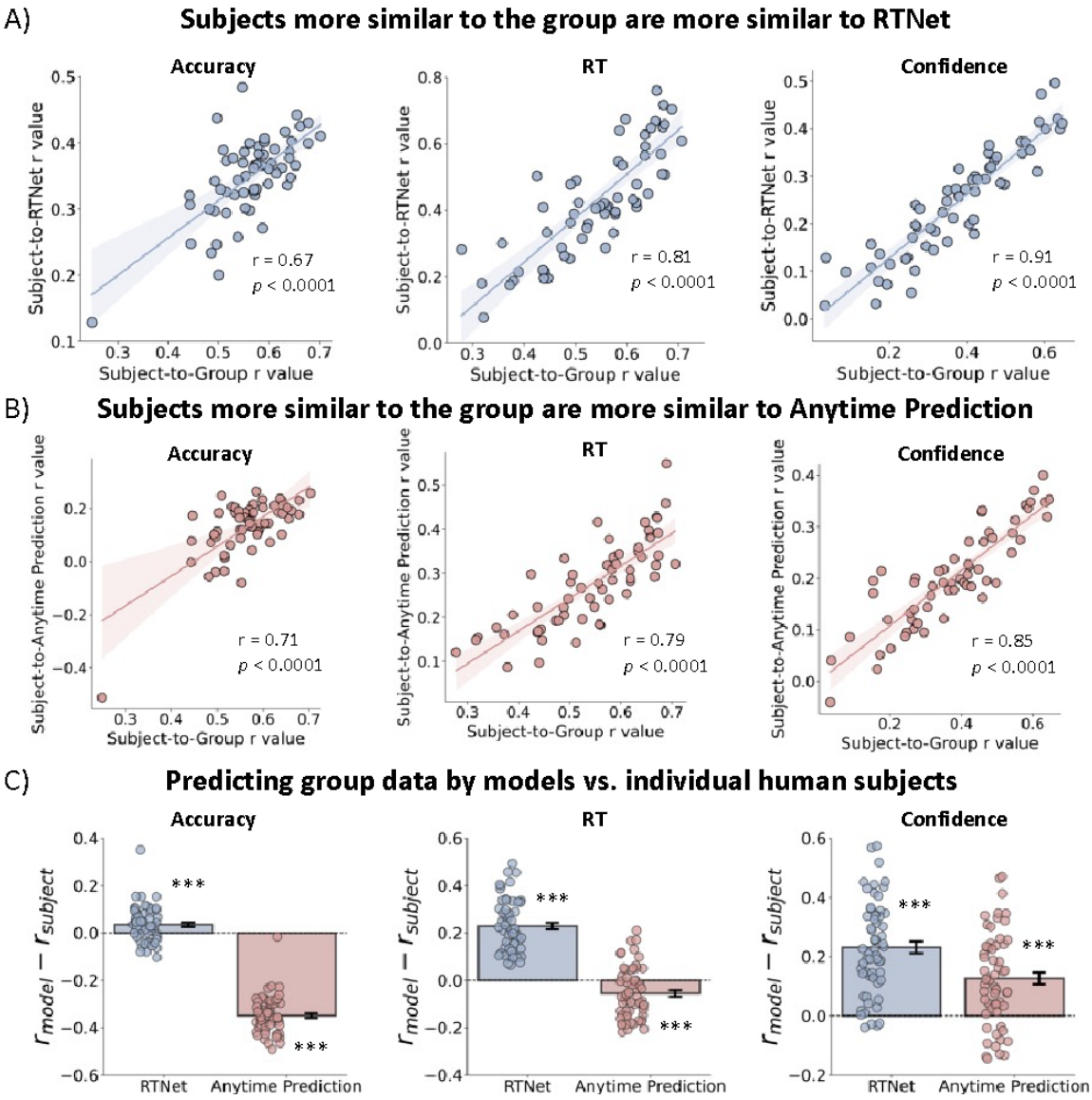

Figure 6. Humans who are more similar to the group average are also more similar to each model. (A) We observed a strong positive correlation between the subject-to-group and subject-to-RTNet similarity values for accuracy, RT, and confidence. This indicates that individual subjects whose behavior was more similar to the group average on per image basis were also more similar to the predictions made by RTNet. (B) Similar results were also observed for Anytime Prediction. (C) Comparison between individual subjects and the two models in predicting the group data. RTNet significantly outperformed individual human subjects in predicting group accuracy, RT, and confidence. On the other hand, Anytime Prediction was worse than individual humans in predicting accuracy and RT, but better in predicting confidence. Dots represent individual subjects; error bars show SEM; ***p<0.001.

Given the variability in how similar individual subjects were to the group data, we also explored how well RTNet and Anytime Prediction compare to the ability of individual subjects to predict the group data. We found that RTNet outperformed individual human subjects in predicting the accuracy ($t(59) = 3.72$, $p = 0.0004$), RT ($t(59) = 16.46$, $p < 0.0001$), and confidence ($t(59) = 11.54$, $p < 0.0001$) of the rest of group across all conditions (**Figure 6C**). Impressively, RTNet outperformed every individual human subject in predicting the group RT results, as well as 71.67% and 93.33% of individual subjects in predicting the accuracy and confidence, respectively. On the other hand, Anytime Prediction performed significantly worse than individual humans in predicting accuracy ($t(59) = -35.67$, $p < 0.0001$; outperformed by 100% of human subjects) and RT ($t(59) = -4.02$, $p < 0.0001$; outperformed by 71.67% of human subjects) of the group, but nevertheless predicted the group confidence results better than most humans ($t(59) = 6.32$, $p < 0.0001$; outperforming 80% of human subjects). Therefore, RTNet outperforms most individual subjects in predicting the group data for accuracy, RT, and confidence, but Anytime Prediction only outperforms most individual subjects in predicting confidence.

Discussion

There is considerable interest in using neural networks as models of human visual processing and behavior, but relatively little work has been done on testing the extent to which existing models reproduce the full range of behavioral signatures exhibited by humans. Here we show that the current state-of-the-art neural network Anytime Prediction diverges in several ways from human behavior. Further, we develop a new neural network, RTNet, that exhibits all critical features of human perceptual decision making, including effects on accuracy, RT, and confidence. Further, RTNet predicted well human group behavior for novel images and did so better than Anytime Prediction or than individual human subjects. Finally, individual humans who were more similar to the group were also more similar to RTNet. Overall, RTNet provides the best current model of human accuracy, RT, and confidence for unseen images.

Relationship between RTNet and cognitive models of perceptual decision making

RTNet is the first neural network to exhibit all critical signatures of human perceptual decision making. This success, however, is hardly surprising given the strong conceptual similarity between RTNet and traditional cognitive models of decision-making that also exhibit the signatures of human behavior (Forstmann et al., 2016; Heathcote & Love, 2012; Heathcote & Matzke, 2022; Ratcliff & Rouder, 1998; Ratcliff & Smith, 2004). These models are often referred to as sequential sampling models where (usually noisy) evidence is accumulated over time until a threshold is reached. The most common sequential sampling models are diffusion models which are typically only applied to 2-choice tasks where evidence in favor of one response alternative is also evidence against the other alternative (Ratcliff, 1978; Ratcliff & Rouder, 1998).

Instead, RTNet is conceptually more similar to another subgroup of sequential sampling models called race models where each choice option has its own accumulation system and evidence for each choice is accumulated in parallel (Brown & Heathcote, 2005, 2008).

Despite their conceptual similarity, RTNet has two important advantages over traditional cognitive models. Most importantly, RTNet can be applied to actual images, whereas traditional models cannot. As such, traditional models cannot replicate RTNet's ability to make accurate predictions regarding human accuracy, RT, and confidence for individual unseen images. The second advantage stems from the inability of traditional cognitive models to naturally capture the relationships between the different choice options. Specifically, to maintain a low number of free parameters, cognitive models are often fit with the assumption that evidence accumulates at the same rate for all incorrect choice options (but accumulates faster for the correct choice) (Tillman, Van Zandt, & Logan, 2020). However, this assumption ignores the fact that some incorrect options may be more similar to the correct option and thus are more likely than other options to be chosen. While dependencies between the choices can easily be incorporated in cognitive models, that would result in a large number of free parameters that would make fitting to data difficult. Conversely, RTNet inherently learns all relationships between the choice options during the training of the Bayesian neural network that forms its core. RTNet still requires the fitting of the overall signal strength (which we accomplish by adjusting the noise level of the images fed to RTNet), but this single free parameter allows it to capture all choice option dependencies, something that traditional models cannot achieve.

Biological plausibility of RTNet and Anytime Prediction

Beyond their behavior, neural networks that aspire to be models of human perceptual decision making should be biologically plausible. Evidence from physiological recordings demonstrates that conduction from one area to another in visual cortex (roughly corresponding to different layers in neural networks) takes approximately 10 ms (Mizuseki, Sirota, Pastalkova, & Buzsáki, 2009), with signal from the photoreceptors reaching the top of the visual hierarchy in inferior temporal cortex in 70-100 ms (Nayebi et al., 2018). Therefore, a single sweep from input to output in a purely feedforward network should result in decisions with RT less than a few hundred milliseconds even though human decisions can range from a hundred of milliseconds to a few seconds. Further, neurons in each layer of the visual cortex continue to fire action potentials for hundreds of milliseconds after the stimulus onset and receive strong recurrent input from later layers of processing (Issa, Cadieu, & Dicarlo, 2018). Finally, neuronal processing is known to be noisy such that the same image input generates very different neuronal activations on different trials (Renart & Machens, 2014).

Anytime Prediction diverges from these known properties of the human visual cortex in several important ways. First, to generate meaningful RTs, Anytime Prediction assumes that classification decisions are made after each layer of processing, though there is no evidence that decisions in the brain can be directly based on information in early visual cortex without further processing in subsequent layers. Moreover, because Anytime Prediction assumes the existence of a single feedforward sweep through the network, it cannot naturally capture large RT variability between stimuli given the short latencies of processing between different layers.

Finally, Anytime Prediction does not incorporate any recurrent processing, capture the noisiness of the responses in the visual cortex, or replicate the long periods of activity of the neurons in each processing area. These properties strongly limit the biological plausibility of Anytime Prediction.

On the other hand, while also not capturing all properties of visual processing, RTNet is much more biologically plausible. First, it naturally mimics the noisiness of neuronal responses for repeated presentations of the same stimulus. Second, because RTNet processes each stimulus multiple times, it naturally generates long-lasting neuronal activations and RTs on the order of many hundreds of milliseconds (or even seconds). Third, the network's output is inherently stochastic, unlike feedforward networks or Anytime Prediction that are inherently deterministic. Finally, the accumulation process implemented in RTNet has been observed in multiple regions in the human parietal cortex, frontal cortex, and subcortical areas (Bahl & Engert, 2019; Hanks, Kiani, & Shadlen, 2014; Huk, Katz, & Yates, 2017; Huk & Shadlen, 2005). Nevertheless, one critical limitation of the biological plausibility of RTNet is its lack of recurrency. That being said, it is currently unknown how to appropriately train recurrent neural networks on static images (Kietzmann, Spoerer, et al., 2019; Nayebi et al., 2018; Spoerer et al., 2020; Spoerer, McClure, & Kriegeskorte, 2017; van Bergen & Kriegeskorte, 2020). Further, while the core of RTNet does not include recurrency, the evidence accumulation system can be thought of as a recurrent network. In fact, several recent studies demonstrated the advantages of combining a standard feedforward network with a recurrent network in performing a range of tasks and extrapolating to solve problems of greater complexity than they were trained on (Schwarzschild et al., 2021;

Zhou et al., 2022). Thus, while RTNet remains less biologically plausible than a true recurrent network, we argue that it is as biologically plausible as current methods of training neural networks permit.

Generating weight distributions for RTNet

One critical feature of RTNet is that its weights are noisy. Practically, there are many different ways of generating noise in the weights. In early iterations of RTNet, we attempted to create variability by training a feedforward network and then adding the same amount of variability to each connection. This approach resulted in variability that was too small for some weights and too large for others (Saltelli et al., 2009), often leading to no accuracy gains from the process of evidence accumulation. Indeed, a given amount of noise over a specific weight may not change the performance of a network at all, but the same disturbance over another weight may have destructive effects (Ko, Kim, Na, Kung, & Mukhopadhyay, 2017; Koutník, Gomez, & Schmidhuber, 2010; Kung, Kim, & Mukhopadhyay, 2015). We therefore chose to obtain the weight variability by training a Bayesian neural network so that each weight has an appropriate amount of noise. In the future, it may be possible to use other methods for setting the noise level for each connection, but we currently unaware of any method besides training a Bayesian neural network that can generate appropriate noise for each weight.

Stochastic vs. deterministic networks

RTNet is inherently a stochastic network. In contrast, most work in machine learning and artificial intelligence has focused on creating deterministic networks. The advantage of

deterministic networks is that they always arrive at the same answer and can be optimized to achieve the best possible performance using the least number of computations. However, such networks are not good models of biological brains that constantly change due to learning and need to be robust to cell death, fluctuations in nutrients, and noise in the information coming from the sensory organs. On a more practical level, deterministic networks may be more vulnerable to adversarial attacks (Chakraborty, Alam, Dey, Chattopadhyay, & Mukhopadhyay, 2018; Xu et al., 2020) compared to, for example, Bayesian neural networks (Uchendu, Campoy, Menart, & Hildenbrandt, 2021; Ye & Zhu, 2018). So, while it remains an open question as to whether and when network stochasticity may be a desirable feature for practical applications, it is critical when trying to model the processing in the human brain.

Limitations

One limitation of RTNet is that its mechanism for stopping the accumulation process is non-optimal. Following a large literature of race models in cognitive psychology (Brown & Heathcote, 2008; Heathcote & Matzke, 2022; Tillman et al., 2020), RTNet makes a decision when any one choice option receives sufficient evidence to exceed a threshold. However, if another choice option has almost same amount of evidence, the observer has little ability to differentiate between the two choices and is essentially guessing between them. Previous research showed that guessing can be an appropriate behavior if the observer knows that the task is very difficult (Malhotra, Leslie, Ludwig, & Bogacz, 2017) or if the observer has been deliberating for a long time (Drugowitsch, Moreno-Bote, Churchland, Shadlen, & Pouget, 2012). However, in a race model, guessing can happen at any time point regardless of task difficulty.

Nevertheless, human decisions are often suboptimal (Evans, Bennett, & Brown, 2019; Rahnev & Denison, 2018), and therefore it is unclear as to whether this suboptimal decision-making mechanism should be seen as a drawback if the goal is to model human decision-making.

Another limitation of RTNet is that each sweep of the feedforward path is independent of the previous states. However, the current state in the human brain is influenced by its previous states (van Bergen & Kriegeskorte, 2020). To address this limitation, the sampling process in RTNet can be modified such that the current state of the network depends on the previous states. For example, a weight over an edge at a specific moment can be made a function of its previous values, which would make the sequential samples dependent on each other. Additional studies are needed to investigate the effect of such state dependence on model performance.

Conclusion

We developed a new neural network, RTNet, which exhibits all basic features of human perceptual decision making and predicts human accuracy, RT, and confidence on an image-by-image basis. The network provides a better model of human perceptual decisions than the state-of-the-art Anytime Prediction network. RTNet thus represents an important step in the use of neural networks as models of human decisions and may even feature properties that are ultimately useful for purely practical applications.

Methods

Behavioral experiment

Pre-registration

This study's sample size, experiment design, included variables, hypothesis, and planned analyses were pre-registered on Open Science Framework (https://osf.io/kmraq) prior to any data being collected.

Subjects

Sixty-four subjects (31 female, age=18-32) with normal or corrected to normal vision were recruited. We had pre-registered the collection of only 40 subjects, but due to less time restrictions than we had anticipated, and to further increase the statistical power, we collected data from more subjects. All subjects signed informed consent and were compensated for their participation. The protocol was approved by the Georgia Institute of Technology Institutional Review Board. All methods were carried out in accordance with relevant guidelines and regulations.

Stimulus, task, and procedure

Subjects performed a digit discrimination task where they reported their perceived digit followed by rating their decision confidence. Each trial began with subjects fixating on a small white cross for 500-1000 ms, followed by a presentation of the stimulus for 300 ms (**Figure 2**). The stimulus was a digit between 1 and 8 (the digits 0 and 9 were excluded) superimposed on a noisy background. Subjects' task was to report the perceived digit using a computer keyboard

by placing four fingers of their left hand on numbers 1-4 and placing four fingers of their right hand on numbers 5-8. This setup allowed subjects to respond without looking at the keyboard, thus providing less noisy response times. Following their categorization response, subjects reported their decision confidence on a 4-point scale (where 1 corresponds to the lowest confidence and 4 corresponds to the highest confidence). There was no deadline on the response or confidence rating.

The experiment included manipulations of speed-accuracy trade-off and task difficulty. Speed-accuracy trade-off was manipulated by asking subjects to emphasize either the speed or accuracy of their responses. To facilitate proper responding, we organized the experiment into alternating blocks of speed and accuracy focus. Task difficulty was manipulated by adding different levels of uniform noise to the stimuli. Specifically, “easy” stimuli included average uniform noise of 0.25 (range = 0-0.5), whereas “difficult” stimuli included average uniform noise of 0.4 (range = 0-0.8). To add the noise, the pixel values were first transformed to be between 0 and 1 and random numbers drawn from the corresponding noise distributions were added separately to each pixel. We scaled the resulting image to be between 0 and 1 again, and finally converted the image to a uint8 format (scaled between 0 and 255). The noise levels were chosen based on the pilot testing to produce two different performance levels. Easy and difficult images were randomly interleaved.

The task stimuli were selected from a publicly available handwritten digits (MNIST) dataset (L. Deng, 2012). This dataset contains 60,000 training images and 10,000 testing images. Since the

training images were used to train the models in this study, we randomly selected images from MNIST test set to include in our experiment. This ensures that the selected images for the experiment are novel both for the human subjects and for the trained models. We randomly selected 480 images for the experiment (120 for each condition). The MNIST dataset images are of size 28 x 28 pixels which appeared overly small on the computer screens we were using. Therefore, before adding noise, the selected images were first resized to 84 x 84 pixels (using MATLAB's *imresize* function), and they were padded with the background color of MNIST images to size 256 x 256 pixels.

The experiment started with three blocks of training each containing 50 trials. The first block contained images from the MNIST dataset without any noise. This was done to familiarize the subjects with the experiment. The next two blocks were used to introduce the speed-accuracy trade-off by asking subjects to focus on accuracy in the first block and on speed in the second. The difficulty level of the stimuli in these two training blocks was same as in the main experiment. During the whole training session, the experimenter was standing beside the subject quietly and was available to answer any questions. None of the images used in the training session was used in the main experiment.

Once the subject confirmed that he or she understands the task, the experimenter left the room and subjects completed the main experiment that consisted of 960 trials organized in four runs each containing four blocks of 60 trials. Each block consisted of a single speed-accuracy trade-off condition, and each run included exactly two "accuracy focus" and two "speed focus"

conditions in a randomized order. At the beginning of each block, subjects were given the name of the condition for that block ("accuracy focus" or "speed focus") and asked to adjust their responding policy accordingly. In each block, we pseudo-randomly interleaved trials from the two difficulty levels such that each was presented exactly 30 times. All 480 images were shown to subjects in first two runs and the procedure was repeated with a new random ordering of the stimuli in the last two runs. All images were same for all subjects, and each image was assigned only to one specific condition.

Apparatus

The experiment was designed in MATLAB 2020b environment using Psychtoolbox 3 (Brainard, 1997). The stimuli were presented on a 21.5-inch Dell P2217H monitor (1920 x 1080 pixel resolution, 60 Hz refresh rate). Subjects were seated 60 cm away from the screen and provided their responses using a keyboard.

Behavioral analyses

We followed the data analyses steps outlined in our preregistration. We first excluded subjects who did not follow sufficiently well the speed/accuracy instructions by not providing faster average RT in the "speed focus" compared to the "accuracy focus" condition. This resulted in removing two subjects (out of 64). We preregistered the exclusion of subjects with floor or ceiling effects on accuracy but no subject met the criteria for exclusion. However, following our preregistration, we excluded two subjects because they showed ceiling effects for confidence. Note that our preregistration document called for excluding subjects who provided average

confidence of more than 3.7 but because this would have resulted in excluding a much larger number of subjects than we had anticipated, we only excluded subjects whose average confidence was above 3.85. Therefore, 60 subjects were used in all subsequent analyses.

We additionally excluded individual trials with extreme RT values using preregistered criteria based on Tukey's interquartile criterion. Specifically, for each subject, we computed the 25th and 75th percentiles of the RT distributions in each condition. We then removed all RTs with values more than 1.5 times the interquartile range such that if $Q1$ is the RT value at the 25th percentile and $Q3$ is the RT value at the 75th percentile, we removed values smaller than $Q1 - 1.5\times(Q3 - Q1)$ and larger than $Q3 + 1.5\times(Q3 - Q1)$. This step resulted in removing an average of 5.46% of total trials (range of 1.35-8.22% for each subject).

Once these preprocessing steps were completed, we computed average accuracy, RT, confidence, and skewness of the RT distributions separately for each condition. The skewness was computed as $\frac{\sum_{i=1}^{N}(x_i-\mu)^3}{(N-1)\sigma^3}$ where μ and σ are the mean and standard deviation of the sample distribution, respectively. We also computed average RT and average confidence scores for error and correct trials across subjects to examine how RT and confidence change as a function of response accuracy. Finally, for visualization purposes, we plotted RT distributions for one subject in Figure 4C. The RT distributions were generated using kernel density estimation (KDE), which approximates the underlying probability density function that generated the data by smoothing

the observations with a Gaussian kernel (Chen, 2017). The KDE plots were created using Seaborn's KDE plot with a smoothing bandwidth of 1.2 (Waskom, 2021).

Model specifications

Network architecture

The RTNet model consists of two main modules (**Figure 1B**). The first module is a Bayesian neural network (BNN) which is capable of making predictions regarding an image. BNNs are a type of artificial neural network built by introducing stochastic components into the network to simulate multiple possible models with their associated probability distribution (Jospin, Buntine, Boussaid, Laga, & Bennamoun, 2020). The main difference between a BNN and standard feedforward neural network is that in BNN the weights are distributions instead of point estimates. A random sample from these distributions results in a unique feedforward network. This random sampling enables variability in the output of the network, which in turn can be fed into an accumulation process that drives a decision. The second module of our model consists of exactly such accumulation of the evidence produced on each step by the first module. Evidence for each choice option was accumulated separately from the rest, similar to a race model (Heathcote & Matzke, 2022). The accumulation process continues until the total amount of accumulated evidence for one of the alternatives reaches a predefined threshold. The alternative for which the threshold was reached then becomes the response of the model. The response time produced by RTNet is simply the number of samples used to reach the decision threshold. The confidence of the model was obtained by applying softmax to the resulting evidence scores.

The Anytime Prediction model has an architecture similar to a standard feedforward neural network (**Figure 1A**) but with early-exit classifiers after each of its layers (**Figure 1C**). At each output layer, the evidence for each choice is computed using a softmax function and if the evidence for any alternative exceeds a predefined value the network stops processing and immediately produces a response. The layer at which the response was made is indicative of the decision time, and the softmax value at that layer is indicative of decision confidence (Huang et al., 2017; Kumbhar, Sizikova, Majaj, & Pelli, 2020).

We implemented both RTNet and Anytime Prediction using the Alexnet architecture, which has eight layers with learnable parameters (Krizhevsky et al., 2012). The Alexnet architecture consists of five convolutional layers with a combination of max pooling followed by three fully connected layers. For Anytime Prediction, in addition to the standard Alexnet structure, we incorporated additional readout layers located right after each layer of processing (**Figure 1C**). The feature map size of all these readout layers were set to the number of classes. All neural networks were implemented in PyTorch (Paszke et al., 2019). Bayesian networks were implemented using Pyro (Bingham et al., 2019), which is a probabilistic programming library built on PyTorch. In addition to the standard Alexnet structure, we incorporated additional readout layers in the Anytime Prediction models. These redout layers were located right after each layer of processing (**Figure 1C**). Specifically, for first and second layer of processing, they were considered immediately after Max pooling layers, whereas for third and fourth layers, they were located right after convolutional layers. The feature map size of all these readout layers

were set to the number of classes. All neural networks in these experiments were implemented in PyTorch (Paszke et al., 2019). Bayesian networks were implemented using Pyro (Bingham et al., 2019) which is a probabilistic programming library built on PyTorch.

Network training

We trained the models to achieve classification accuracy higher than 97% on the MNIST test set. To achieve this, we trained the BNN module of RTNet for a total of 15 epochs with a batch size of 500. Evidence lower bound (ELBO) loss function was used for training these networks (Kingma & Welling, 2013). Due to its deterministic nature, for Anytime Prediction, only three epochs were enough to achieve test accuracy of more than 97% with the same batch size and a weighted cumulative loss function (Kumbhar et al., 2020). For all networks, Adam (Kingma & Ba, 2014) was used for optimization with a learning rate of 0.001. To ensure that each network performs greater than 97% on MNIST test set, we followed a specific rule for each model. When testing an image with the BNN module of RTNet, we sampled 10 times from the posterior distributions learned during the training and thus obtained 10 unique responses for each image. The response with highest frequency among 10 responses was chosen as the final decision of the BNN module. For Anytime Prediction, we considered the response of the last output layer as the network's decision. If a network did not achieve accuracy greater than 97%, we started the training over with same initial values. Because the standard input size to Alexnet model architecture is 227 x 227 pixels, we resized the MNIST images to this size. We also normalized the input images to have a mean of 0.1307 and standard deviation of 0.3081, which is a

standard procedure when using Alexnet for classification of the ImageNet dataset (J. Deng et al., 2010). The training used gradient descent and backpropagation (Rawat & Wang, 2017).

We trained sixty instantiations of both RTNet and Anytime Prediction using the above procedure but with different initializations for each network instantiation. For RTNet, we used a different combination of mean and standard deviation (SD) values for each of the 60 instantiations. Specifically, different network instantiations of RTNet were initialized such that all means of the weights and biases were set to a value between 0.1 and 1.2 with 0.1 increments, and all SDs of weights and biases were set to a value ranging from 1 to 5 with increments of 1 (for a total of $12 \times 5 = 60$ instantiations). To make the initializations of Anytime Prediction as similar as possible to the initializations of RTNet, for each RTNet instantiation, we set the initial values for the weights and biases of the Anytime Prediction instantiation by randomly sampling from the Gaussian distribution used in the corresponding RTNet initialization.

Choosing parameters that allow the models to mimic human accuracy

Because the goal of our study was to examine whether the models exhibit the signatures of human perceptual decision making, we matched the accuracy of the models across the four experimental conditions to the average accuracy in the human data. For both models, this was achieved by adjusting the noise level in the images (separately for the “easy” and “difficult” images) and the threshold parameter (separately for the speed and accuracy conditions).

Parameter values were adjusted using a coarse search followed by a fine search. In the coarse search for RTNet, we varied the amplitude of uniform noise from 1 to 10 with increments of 1, and the threshold value from 2 to 12 with increments of 2. The results were closest to the human accuracy levels when the noise was in the range 2-3 for easy images and 4-5 for difficult images, and the threshold was set to 2-4 for the speed focus condition and 6-8 for the accuracy focus condition. We then conducted a fine search near those values by changing the noise level from 2 to 5 with 0.1 increments and changing the threshold values from 2 to 8 with 0.5 increments. The closest match to human accuracy was achieved for noise levels of 2.1 and 4.1 for easy and difficult images, respectively, and a threshold value of 3 for the speed condition and 6 for the accuracy condition.

We used a similar procedure to tune the parameters of Anytime Prediction. Note that the threshold value for Anytime Prediction is the softmax evidence at each early exit. The coarse search for Anytime Prediction was performed using the threshold values between 0.5 and 0.95 with increments of 0.05. The results were closest to the human accuracy levels when the threshold was in range 0.55-0.65 for the speed focus condition, and 0.8-0.9 for the accuracy focus condition. We then performed a fine search in these ranges by incrementing the threshold by steps of 0.01. The closest match to human accuracy was achieved for a threshold value of 0.58 for the speed condition and 0.82 for the accuracy condition. For finding the optimal noise levels, we followed the same procedure that we used for RTNet. The best match was obtained when the noise levels were set to 1.9 and 3.0 for easy and difficult images, respectively.

Data and code availability

Behavioral data, as well as all codes and trained models are publicly available at:

https://osf.io/akwty.

References

Bahl, A., & Engert, F. (2019). Neural circuits for evidence accumulation and decision making in larval zebrafish. *Nature Neuroscience 2019 23:1*, *23*(1), 94–102. https://doi.org/10.1038/s41593-019-0534-9

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., & Pouget, A. (2012). Not Noisy, Just Wrong: The Role of Suboptimal Inference in Behavioral Variability. *Neuron*, *74*(1), 30–39. https://doi.org/10.1016/J.NEURON.2012.03.016

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., … Goodman, N. D. (2019). Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*, *20*, 1–6. Retrieved from http://jmlr.org/papers/v20/18-403.html.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765. https://doi.org/10.1037/0033-295X.113.4.700

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Brown, S., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review*, *112*(1), 117–128. https://doi.org/10.1037/0033-295X.112.1.117

Brown, S., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. https://doi.org/10.1016/J.COGPSYCH.2007.12.002

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2018). *Adversarial Attacks and Defences: A Survey*. https://doi.org/10.48550/arxiv.1810.00069

Chen, Y. C. (2017). A tutorial on kernel density estimation and recent advances. *Https://Doi.Org/10.1080/24709360.2017.1396742*, *1*(1), 161–187. https://doi.org/10.1080/24709360.2017.1396742

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2010). *ImageNet: A large-scale hierarchical image database*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Deng, L. (2012). The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, *29*(6), 141–142. https://doi.org/10.1109/MSP.2012.2211477

Drugowitsch, J., Moreno-Bote, R. N., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The Cost of Accumulating Evidence in Perceptual Decision Making. *Journal of Neuroscience*, *32*(11), 3612–3628. https://doi.org/10.1523/JNEUROSCI.4010-11.2012

Evans, N. J., Bennett, A. J., & Brown, S. D. (2019). Optimal or not; depends on the task. *Psychonomic Bulletin and Review*, *26*(3), 1027–1034. https://doi.org/10.3758/S13423-018-1536-4/FIGURES/2

Findling, C., & Wyart, V. (2021). Computation noise in human learning and decision-making: origin, impact, function. *Current Opinion in Behavioral Sciences*, *38*, 124–132. https://doi.org/10.1016/J.COBEHA.2021.02.018

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., Cramon, D. Y. von, Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences*, *105*(45), 17538–17542. https://doi.org/10.1073/PNAS.0805903105

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential Sampling Models in

Cognitive Neuroscience: Advantages, Applications, and Extensions. *Annual Review of Psychology*, *67*, 641–666. https://doi.org/10.1146/annurev-psych-122414-033645

Geirhos, R., Janssen, D. H. J., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). *Comparing deep neural networks against humans: object recognition when the signal gets weaker*. https://doi.org/10.48550/arxiv.1706.06969

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. Retrieved June 6, 2022, from Advances in Neural Information Processing Systems 31 website: https://proceedings.neurips.cc/paper/2018/hash/0937fb5864ed06ffb59ae5f9b5ed67a9-Abstract.html

Gold, J. I., & Shadlen, M. N. (2007). *The Neural Basis of Decision Making*. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Hanks, T. D., Kiani, R., & Shadlen, M. N. (2014). A neural mechanism of speed-accuracy tradeoff in macaque area LIP. *ELife*, *2014*(3). https://doi.org/10.7554/ELIFE.02260

Heathcote, A., & Love, J. (2012). Linear deterministic accumulator models of simple choice. *Frontiers in Psychology*, *3*(AUG), 292. https://doi.org/10.3389/FPSYG.2012.00292/BIBTEX

Heathcote, A., & Matzke, D. (2022). Winner takes all! What are race models, and why and how should psychologists use them? *Current Directions in Psychological Science*. Retrieved from https://www.ampl-psych.com/wp-content/uploads/2022/03/Heathcote-Matzke_2022_CDPS.pdf

Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, 150. https://doi.org/10.3389/fnins.2014.00150

Heitz, R. P., & Schall, J. D. (2012). Neural mechanisms of speed-accuracy tradeoff. *Neuron*, *76*(3), 616–628. https://doi.org/10.1016/j.neuron.2012.08.030

Huang, G., Chen, D., Li, T., Wu, F., van der Maaten, L., & Weinberger, K. Q. (2017). Multi-Scale Dense Networks for Resource Efficient Image Classification. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. Retrieved from https://arxiv.org/abs/1703.09844v5

Huk, A. C., Katz, L. N., & Yates, J. L. (2017). The Role of the Lateral Intraparietal Area in (the Study of) Decision Making. *Annual Review of Neuroscience*, *40*, 349. https://doi.org/10.1146/ANNUREV-NEURO-072116-031508

Huk, A. C., & Shadlen, M. N. (2005). Neural Activity in Macaque Parietal Cortex Reflects Temporal Integration of Visual Motion Signals during Perceptual Decision Making. *Journal of Neuroscience*, *25*(45), 10420–10436. https://doi.org/10.1523/JNEUROSCI.4684-04.2005

Issa, E. B., Cadieu, C. F., & Dicarlo, J. J. (2018). Neural dynamics at successive stages of the ventral visual stream are consistent with hierarchical error signals. *ELife*, *7*. https://doi.org/10.7554/ELIFE.42870

Jospin, L. V., Buntine, W., Boussaid, F., Laga, H., & Bennamoun, M. (2020). *Hands-on Bayesian Neural Networks -- a Tutorial for Deep Learning Users*. Retrieved from https://arxiv.org/abs/2007.06823v1

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019). Deep Neural Networks in Computational Neuroscience. *Oxford Research Encyclopedia of Neuroscience*. https://doi.org/10.1093/ACREFORE/9780190264086.013.46

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N.

(2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854–21863. https://doi.org/10.1073/PNAS.1905544116

Kingma, D. P., & Ba, J. L. (2014). Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. https://doi.org/10.48550/arxiv.1412.6980

Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*. https://doi.org/10.48550/arxiv.1312.6114

Ko, J. H., Kim, D., Na, T., Kung, J., & Mukhopadhyay, S. (2017). Adaptive weight compression for memory-efficient neural networks. *Proceedings of the 2017 Design, Automation and Test in Europe, DATE 2017*, 199–204. https://doi.org/10.23919/DATE.2017.7926982

Koutník, J., Gomez, F., & Schmidhuber, J. (2010). Evolving neural networks in compressed weight space. *Proceedings of the 12th Annual Genetic and Evolutionary Computation Conference, GECCO '10*, 619–625. https://doi.org/10.1145/1830483.1830596

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Http://Dx.Doi.Org/10.1146/Annurev-Vision-082114-035447*, *1*(1), 417–446. https://doi.org/10.1146/ANNUREV-VISION-082114-035447

Kriegeskorte, N., & Golan, T. (2019). Neural network models and deep learning. *Current Biology*, *29*(7), R231–R236. https://doi.org/10.1016/J.CUB.2019.02.034

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep

Convolutional Neural Networks. Retrieved April 15, 2022, from Advances in Neural Information Processing Systems 25 (NIPS 2012) website: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

Kumbhar, O., Sizikova, E., Majaj, N., & Pelli, D. G. (2020). *Anytime Prediction as a Model of Human Reaction Time*. Retrieved from https://arxiv.org/abs/2011.12859v1

Kung, J., Kim, D., & Mukhopadhyay, S. (2015). A power-aware digital feedforward neural network platform with backpropagation driven approximate synapses. *Proceedings of the International Symposium on Low Power Electronics and Design*, *2015-September*, 85–90. https://doi.org/10.1109/ISLPED.2015.7273495

Luce, R. D. (1986). *Response Times*. https://doi.org/10.1093/acprof:oso/9780195070019.001.0001

Malhotra, G., Leslie, D. S., Ludwig, C. J. H., & Bogacz, R. (2017). Overcoming indecision by changing the decision boundary. *Journal of Experimental Psychology: General*, *146*(6), 776. https://doi.org/10.1037/XGE0000286

Mizuseki, K., Sirota, A., Pastalkova, E., & Buzsáki, G. (2009). Theta Oscillations Provide Temporal Windows for Local Circuit Computation in the Entorhinal-Hippocampal Loop. *Neuron*, *64*(2), 267–280. https://doi.org/10.1016/J.NEURON.2009.08.037

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., … Yamins, D. L. K. (2018). Task-Driven Convolutional Recurrent Models of the Visual System. *Advances in Neural Information Processing Systems*, *2018-December*, 5290–5301. Retrieved from https://arxiv.org/abs/1807.00053v2

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. Retrieved April 18, 2022, from Advances in Neural Information Processing Systems 32 (NeurIPS) website: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

Rahnev, D. (2020). Confidence in the Real World. *Trends in Cognitive Sciences*, *24*(8), 590–591. https://doi.org/10.1016/J.TICS.2020.05.005

Rahnev, D. (2021). Visual metacognition: Measures, models, and neural correlates. *The American Psychologist*, *76*(9), 1445–1453. https://doi.org/10.1037/AMP0000937

Rahnev, D., & Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences*, *41*. https://doi.org/10.1017/S0140525X18000936

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review*, *9*(2), 278–291. https://doi.org/10.3758/BF03196283

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Ratcliff, R., & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions. *Psychological Science*, *9*(5), 347–356.

Ratcliff, R., & Smith, P. L. (2004). A Comparison of Sequential Sampling Models for Two-Choice

Reaction Time. *Psychological Review*, *111*(2), 333–367. https://doi.org/10.1037/0033-295X.111.2.333

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260–281. https://doi.org/10.1016/J.TICS.2016.01.007

Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, *29*(9), 2352–2449. https://doi.org/10.1162/NECO_A_00990

Renart, A., & Machens, C. K. (2014). Variability in neural activity and behavior. *Current Opinion in Neurobiology*, *25*, 211–220. https://doi.org/10.1016/J.CONB.2014.02.013

Saltelli, A., Yeung, D. S., Cloete, I., Shi, D., Ng, W. W. Y., Lee, J. H. W., … Jayawardena, A. W. (2009). Sensitivity Analysis for Neural Networks. Natural Computing. *Risk Analysis*, *159*(2–3), 179–201. Retrieved from http://ebooks.ciando.com/book/index.cfm/bok_id/43309%5Cnhttp://www.gbv.de/dms/bowker/toc/9783642025310.pdf%5Cnhttp://www.ciando.com/img/books/width167/3642025323_k.jpg%5Cnhttp://www.ciando.com/pictures/bib/3642025323bib_t_1.jpg

Schwarzschild, A., Borgnia, E., Gupta, A., Huang, F., Vishkin, U., Goldblum, M., & Goldstein, T. (2021). Can You Learn an Algorithm? Generalizing from Easy to Hard Problems with Recurrent Networks. Retrieved June 16, 2022, from Advances in Neural Information Processing Systems 34 website: https://proceedings.neurips.cc/paper/2021/hash/3501672ebc68a5524629080e3ef60aef-Abstract.html

Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I., & Kriegeskorte, N. (2020). Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLOS Computational Biology*, *16*(10), e1008215. https://doi.org/10.1371/JOURNAL.PCBI.1008215

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent Convolutional Neural Networks: A Better Model of Biological Object Recognition. *Frontiers in Psychology*, *0*(SEP), 1551. https://doi.org/10.3389/FPSYG.2017.01551

Subramanian, A., Price, S., Kumbhar, O., Sizikova, E., Majaj, N. J., & Pelli, D. G. (2022). *SATBench: Benchmarking the speed-accuracy tradeoff in object recognition by humans and dynamic neural networks*. https://doi.org/10.48550/arxiv.2206.08427

Tillman, G., Van Zandt, T., & Logan, G. D. (2020). Sequential sampling models without random between-trial variability: the racing diffusion model of speeded decision making. *Psychonomic Bulletin and Review*, *27*(5), 911–936. https://doi.org/10.3758/S13423-020-01719-6/FIGURES/16

Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological Review*, *121*(2), 179–205. https://doi.org/10.1037/A0036137

Uchendu, A., Campoy, D., Menart, C., & Hildenbrandt, A. (2021). Robustness of Bayesian Neural Networks to White-Box Adversarial Attacks. *Proceedings - 2021 IEEE 4th International Conference on Artificial Intelligence and Knowledge Engineering, AIKE 2021*, 72–80. https://doi.org/10.1109/AIKE52691.2021.00017

van Bergen, R. S., & Kriegeskorte, N. (2020, December 1). Going in circles is the way forward: the

role of recurrence in visual inference. *Current Opinion in Neurobiology*, Vol. 65, pp. 176–193. https://doi.org/10.1016/j.conb.2020.11.009

Vickers, D. (2007). Evidence for an Accumulator Model of Psychophysical Discrimination. *Http://Dx.Doi.Org/10.1080/00140137008931117*, *13*(1), 37–58. https://doi.org/10.1080/00140137008931117

Wagenmakers, E.-J., & Brown, S. (2007). On the Linear Relation Between the Mean and the Standard Deviation of a Response Time Distribution. *Psychological Review*, *114*(3), 830–841. https://doi.org/10.1037/0033-295X.114.3.830

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/JOSS.03021

Wyart, V., & Koechlin, E. (2016). Choice variability and suboptimality in uncertain environments. *Current Opinion in Behavioral Sciences*, *11*, 109–115. https://doi.org/10.1016/J.COBEHA.2016.07.003

Xu, H., Ma, Y., Liu, H. C., Deb, D., Liu, H., Tang, J. L., & Jain, A. K. (2020). Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. *International Journal of Automation and Computing 2020 17:2*, *17*(2), 151–178. https://doi.org/10.1007/S11633-019-1211-X

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience 2016 19:3*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244

Ye, N., & Zhu, Z. (2018). Bayesian Adversarial Learning. Retrieved June 15, 2022, from Advances in Neural Information Processing Systems 31 website: https://papers.nips.cc/paper/2018/hash/586f9b4035e5997f77635b13cc04984c-Abstract.h

tml

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310–1321. https://doi.org/10.1098/RSTB.2011.0416

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., … Chi, E. (2022). *Least-to-Most Prompting Enables Complex Reasoning in Large Language Models*. https://doi.org/10.48550/arxiv.2205.10625