nature human behaviour

Article

https://doi.org/10.1038/s41562-024-01914-8

# The neural network RTNet exhibits the signatures of human perceptual decision-making

Received: 18 October 2023

Accepted: 13 May 2024

Published online: 12 July 2024

Check for updates

Farshad Rafiei [1,2] ✉, Medha Shekhar [1,2] & Dobromir Rahnev [1]

Convolutional neural networks show promise as models of biological vision. However, their decision behaviour, including the facts that they are deterministic and use equal numbers of computations for easy and difficult stimuli, differs markedly from human decision-making, thus limiting their applicability as models of human perceptual behaviour. Here we develop a new neural network, RTNet, that generates stochastic decisions and human-like response time (RT) distributions. We further performed comprehensive tests that showed RTNet reproduces all foundational features of human accuracy, RT and confidence and does so better than all current alternatives. To test RTNet's ability to predict human behaviour on novel images, we collected accuracy, RT and confidence data from 60 human participants 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 participants. Critically, human participants who were more similar to the average human performance were also found to be closer to RTNet's predictions, suggesting that RTNet successfully captured average human behaviour. Overall, RTNet is a promising model of human RTs that exhibits the critical signatures of perceptual decision-making.

Traditional cognitive models of perceptual decisions[1–4] are able to account for the major features of human perceptual decision-making but do not operate on the level of images. Recently, convolutional neural networks (CNNs) have reached and sometimes exceeded human-level performance for novel images[5,6]. In addition, these networks naturally handle multi-choice categorization tasks and are promising models of the processing related to object recognition in the ventral visual stream of the human brain[5,7,8]. However, traditional CNNs' decision behaviour differs markedly from human decision behaviour, thus limiting their applicability as models of human perceptual decision-making. Specifically, unlike humans, traditional CNNs are both deterministic (that is, they always give the same response for a given stimulus) and static (that is, they are invariant in the amount of time spent on processing different images and thus always produce the same response time (RT)).

Several lines of work have tried to build mechanisms into neural networks to make them stochastic and dynamic[9–13]. Early research on shallow multi-layer perceptron models was able to create models that were both stochastic and dynamic. These models were able to explain human behaviour on simple cognitive tasks[14–16]. However, these models are not image-computable (that is, they cannot handle complex input such as images). More recent work has produced image-computable dynamic networks capable of generating RTs via mechanisms that allow computational resources used for the decision to increase with time[9–11], thus allowing responses to evolve through each processing step. However, although these networks can mimic the speed–accuracy

[1]School of Psychology, Georgia Institute of Technology, Atlanta, GA, USA. [2]These authors contributed equally: Farshad Rafiei, Medha Shekhar.
✉e-mail: farshadrafiei3@gmail.com

trade-off (SAT) found in humans, they are deterministic, and their internal mechanisms are not well supported by existing models of human perception and cognition. Finally, another class of models generates RTs using the biologically inspired mechanism of recurrent processing[17–21], which allows flexible modulation of a finite network's computational power[10,22]. Nevertheless, these networks are also deterministic and have not been evaluated on the whole range of choice, RT and confidence effects shown by humans.

Here we combine modern CNNs with traditional cognitive models to create a model that is image-computable, stochastic and dynamic and that can reproduce the critical features of perceptual decision-making for novel images. The model, which we call RTNet for its ability to model human RTs, features a deep CNN with noisy weights and processes a given image several times using a different random sample of these weights in each processing step (Fig. 1a). These weights are sampled from a Bayesian neural network (BNN) that estimates a posterior distribution over the best network parameters learned during training. By sampling from these noisy weight distributions at each processing step, the network's units produce variable responses to the same input that mimic the randomness of neural responses. After each processing step, RTNet accumulates the output corresponding to each choice until one of the choices reaches a predefined threshold. The model therefore has a 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[23–25]. By combining the image-computability of CNNs with traditional models of perception, we expect RTNet to be applicable across a wide range of perceptual tasks as well as reproduce the basic features of human perceptual decision-making.

To assess a model's ability to make decisions similar to those of humans, one needs to test whether it produces the foundational features of human decision-making[26]. Human perceptual decision-making has been studied primarily in the context of two-choice tasks using artificial stimuli such as Gabor patches or random dot motion[27] (although notable exceptions exist where *N*-choice tasks are used[28–31]). We therefore first replicated the known decision-making signatures from two-choice tasks using an eight-choice task with meaningful images (handwritten digits taken from the MNIST dataset[32]). We manipulated (1) task difficulty by adding two different levels of noise to the images and (2) the SAT by asking the participants to emphasize either the accuracy or the speed of their responses on different trials.

Critically, we tested RTNet under the same conditions and with the same images seen by the human participants to explore the model's capability to produce behaviour similar to that of human agents. Beyond testing whether RTNet can reproduce the basic features of human perceptual decision-making, we also explored whether the accuracy, RT and confidence produced by RTNet for individual images predict the corresponding quantities for humans on the same images. Finally, throughout the study, we compared the behaviour of RTNet to that of three other popular dynamic CNNs. The first model is Parallel Cascaded Network[9] (CNet; Fig. 1b), which is currently thought to be the best image-computable model that can mimic the SAT characteristics of humans[12]. The second is BLNet[10], which belongs to a class of models that use recurrent processing and has been validated on a range of perceptual tasks involving manipulations beyond the SAT (Fig. 1c). The third is Multi-Scale Dense Networks[13] (MSDNet; Fig. 1d), which implements one of the most common ways for generating RTs in image-computable models. We found that RTNet's behaviour mimics human perceptual decision-making better than all three of these CNNs.

## Results

We collected data from 60 human participants who performed a digit discrimination task (Fig. 2a). The experiment was a 2 × 2 design with factors of task difficulty (easy versus difficult images) and speed pressure (speed versus accuracy focus). Each condition consisted of 120 unique images, and each participant made a decision regarding each image exactly twice, which allowed us to determine the level of stochasticity in human behaviour (Fig. 2b). Overall, each participant completed 960 trials.

Having obtained these human data, we compared human behaviour to that of RTNet, CNet, BLNet and MSDNet. Both RTNet and MSDNet were implemented using the eight-layer AlexNet architecture with five convolutional layers followed by three fully connected layers[33]. CNet was based on the architecture of ResNet18 since the implementation of this model relies on residual blocks and skip connections. Finally, for BLNet, we used the original architecture implemented by Spoerer et al.[10], which consists of seven convolutional layers and a fully connected readout layer. Given that humans and deep learning models are impacted differently by stimulus noise[34,35], we adjusted the noise levels of the images seen by each network to match their overall accuracy to the accuracy produced by the human participants. In addition, to allow the networks to reproduce the SAT observed in the human data, we adjusted the threshold value that triggers a decision for each CNN to match the human accuracy separately in the speed and accuracy focus conditions. To improve the correspondence between the model predictions and the human data, we trained 60 instances of each model (by changing only the random initialization before training began) and analysed the data produced by these 60 instances in an equivalent manner to those from the 60 human participants.

### Signatures of human perceptual decision-making

We examined six foundational signatures of human perceptual decision-making that have already been established in studies of two-choice tasks: (1) human decisions are stochastic, meaning that the same stimulus can elicit different responses on different trials[36,37]; (2) increasing speed stress shortens RTs but decreases accuracy (SAT)[26,38,39]; (3) more difficult decisions lead to reduced accuracy and longer RTs[26,40,41]; (4) RT distributions are right-skewed, and this skew increases with task difficulty[26]; (5) RT is lower for correct trials than for error trials[41–45]; and (6) confidence is higher for correct trials than for error trials[46]. For each of these signatures, we confirmed that the signature also occurs for our eight-choice task with naturalistic images, and we then tested whether RTNet, CNet, BLNet and MSDNet exhibit the same signature.

**Stochasticity of human decisions.** A central feature of human behaviour is that human decisions are stochastic such that the same stimulus can elicit different responses on different trials[36,37,47]. We quantified the level of stochasticity in each condition by presenting each image twice. We first confirmed that our estimates of human stochasticity were robust and reliable by showing that similar estimates are obtained when analysing the odd- versus even-numbered participants (Supplementary Fig. 1). On average across all conditions, 36% of all images received different responses on the two presentations. A one-sided Wilcoxon signed-rank test showed that this observed frequency of stochastic responses is indeed significantly greater than zero ($Z_{59}$ = 32,896; $P$ < 0.001; rank-biserial correlation (effect size), 1) (Fig. 3a). A repeated-measures analysis of variance with the factors stimulus difficulty (easy versus difficult) and SAT (speed versus accuracy stress) revealed that stochasticity increased with both higher task difficulty ($F_{1,63}$ = 871.869, $P$ < 0.001, $\eta_p^2$ = 0.933) and higher speed pressure ($F_{1,63}$ = 9.135, $P$ = 0.004, $\eta_p^2$ = 0.127).

Since 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 (one-sided Wilcoxon signed-rank test: $Z_{59}$ = 2,892; $P$ < 0.001; rank-biserial correlation (effect size), 1; Fig. 3b). This level of stochasticity was lower than for human participants and stems from the fact that the variability in the weights was fixed a priori by training a BNN. However, it is possible for RTNet to match the level

https://doi.org/10.1038/s41562-024-01914-8

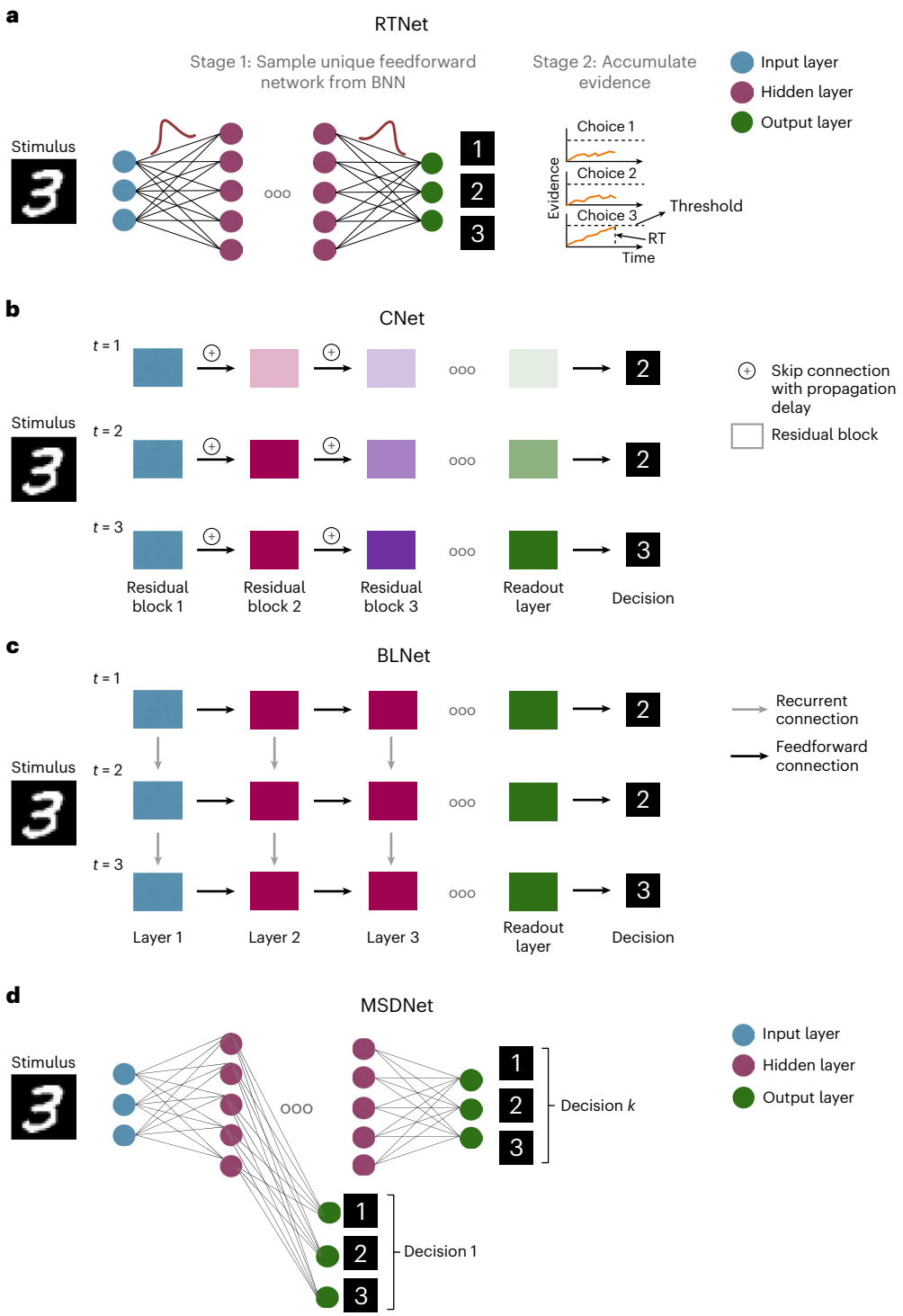


**Fig. 1 | Model architectures. a**, RTNet architecture. Unlike standard CNNs, the connection weights in RTNet are not fixed but chosen from a distribution. A stimulus is processed multiple times by the network, each time using a different set of weights sampled randomly from a BNN. The evidence from each processing step is accumulated, and a decision is made when the evidence for one of the choices reaches a threshold. This architecture results in both stochastic decisions and variable RTs. **b**, CNet architecture[9]. CNet introduces propagation delays between residual blocks (each of which consists of two convolutional layers). At each time step ($t$), all residual blocks parallelly receive inputs from lower blocks, but due to propagation delays, earlier blocks achieve stable activations faster, whereas the later blocks require multiple processing steps to receive complete input and achieve stable activations. The network can generate a decision via the readout layer at any time step, although if the time step is less than the number of residual blocks, the decision will be based on partial input in later blocks. **c**, BLNet architecture[10]. BLNet is an RCNN with bottom-up and lateral recurrent connections. Time steps are defined in terms of the number of feedforward sweeps of the network. At each time step, a layer receives feedforward input from the previous layer as well as recurrent input from its own activations at the previous time step. The readout can be evaluated at each time step to generate a response if it exceeds the threshold. The network can trade off speed and accuracy as higher thresholds require more feedforward and recurrent computations, effectively resulting in a deeper network being unrolled across time. **d**, MSDNet architecture[13]. In this network, each hidden layer features its own classifier, allowing MSDNet 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.

https://doi.org/10.1038/s41562-024-01914-8

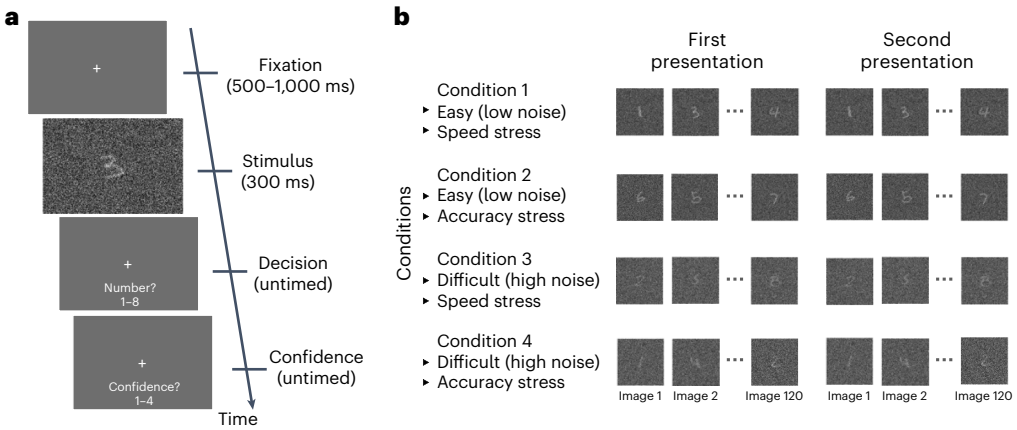


**Fig. 2 | Experiment task. a**, Trial structure. Each trial began with a fixation cross presented for 500 to 1,000 ms, followed by an image of a handwritten digit from the MNIST dataset embedded in noise and presented for 300 ms. Only the digits 1–8 were used. The participants reported their choice and confidence (on a four-point scale) using separate, untimed button presses. Note that the noisy stimulus subtended a visual angle of 6.06° and did not cover the entire screen. **b**, Experimental design. The experiment included four conditions such that the participants 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 participants (for a 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 participant thus completed a total of 960 trials. The images in the first and second sets of presentation were shown in a different random order.

of stochasticity observed in humans by increasing the variability of the network's weights. Indeed, we confirmed that the stochasticity of the decisions made by RTNet can be robustly manipulated by changing the variability of its weight distributions (Supplementary Fig. 2). Furthermore, the stochasticity in human decisions partially stems from factors such as fluctuations in attention, arousal or serial dependence[36,37,47,48], 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 ($F_{1,59} = 120.124$, $P < 0.001$, $\eta_p^2 = 0.671$) and higher speed stress ($F_{1,59} = 87.730$, $P < 0.001$, $\eta_p^2 = 0.598$).

In contrast, for a fixed level of SAT, CNet, BLNet and MSDNet are fully deterministic and do not exhibit any decision stochasticity (Fig. 3c–e). We note that it should be possible to add noise in the weights of these models to induce stochastic decisions, but such noise would decrease their accuracy much more than it affects RTNet given that only RTNet is able to average out the noise over repeated processing steps. Because RTNet is the only model that incorporates a mechanism for generating stochastic responses, these stochasticity analyses a priori favour RTNet over the other models. However, the rest of our analyses compare the behaviour and predictions of the models across a range of stimulus manipulations in which no model is a priori expected to be favoured over the others.

**SAT.** The ability to trade off speed and accuracy against each other is a hallmark of decision-making across humans and many other animal species[38,39]. The human data confirmed that increased speed pressure led to lower accuracy ($F_{1,59} = 4.274$, $P = 0.043$, $\eta_p^2 = 0.068$; Fig. 4a) and shorter RTs ($F_{1,59} = 119.29$, $P < 0.001$, $\eta_p^2 = 0.964$; Fig. 4b). We also found a significant interaction between the SAT and task difficulty for accuracy such that the SAT effect was greater for easy images ($F_{1,59} = 5.71$, $P = 0.020$, $\eta_p^2 = 0.088$). For RTs, however, we observed the opposite pattern, where the SAT effect was heightened for difficult images ($F_{1,59} = 22.423$, $P < 0.001$, $\eta_p^2 = 0.275$). These results replicate findings from a previous study examining the effects of SAT manipulations on accuracy and RT as a function of stimulus contrast[49]. Furthermore, as shown before[49], these findings are also in line with predictions of the drift diffusion model, which is currently the standard model for explaining human RTs[1,2].

All models were able to replicate the SAT observed in humans. Increased speed pressure resulted in lower accuracy for RTNet ($F_{1,59} = 9.683$, $P = 0.003$, $\eta_p^2 = 0.141$), CNet ($F_{1,59} = 50.025$, $P < 0.001$, $\eta_p^2 = 0.459$), BLNet ($F_{1,59} = 11.611$, $P = 0.001$, $\eta_p^2 = 0.164$) and MSDNet ($F_{1,59} = 21.841$, $P < 0.001$, $\eta_p^2 = 0.270$). Increased speed pressure also led to shorter RTs for RTNet ($F_{1,59} = 3{,}362.567$, $P < 0.001$, $\eta_p^2 = 0.983$), CNet ($F_{1,59} = 695.878$, $P < 0.001$, $\eta_p^2 = 0.922$), BLNet ($F_{1,59} = 607.093$, $P < 0.001$, $\eta_p^2 = 0.911$) and MSDNet ($F_{1,59} = 584.081$, $P < 0.001$, $\eta_p^2 = 0.908$). We note that the SAT manipulation had a relatively small effect on accuracy (1.04% for easy and 1.24% for difficult conditions for RTNet; the effects for the rest of the networks were of similar magnitude; Fig. 4a). However, despite the small effect size, these effects were generally consistent across the 60 model instances (for RTNet, 54/60 instances showed the effect for easy images, and 42/60 showed the effect for difficult images).

The SAT manipulation had a much stronger effect on RTs than on accuracy, which may be attributed to the fact that RTs are a more sensitive measure of performance. Furthermore, the SAT effect on RTs was much stronger for humans, RTNet and BLNet than for the other models. The individual RT distributions show a clear separation between the speed and accuracy focus conditions for humans, RTNet and BLNet but not for CNet and MSDNet (Fig. 4c). Nevertheless, these results indicate that the SAT is robustly observed even for a relatively complex task with naturalistic images, and that all models examined here exhibit this foundational phenomenon.

**Difficult decisions lead to reduced accuracy and longer RTs.** Another ubiquitous feature of decision-making is that more difficult stimuli lead to lower accuracy and longer RTs[26,50]. Our human data robustly showed this effect, with more difficult stimuli leading to lower accuracy ($F_{1,59} = 1{,}558.500$, $P < 0.001$, $\eta_p^2 = 0.964$; Fig. 4a) and longer RTs ($F_{1,59} = 411.154$, $P < 0.001$, $\eta_p^2 = 0.875$; Fig. 4b). The same pattern was robustly observed for RTNet and BLNet, where difficult stimuli led to lower accuracy (RTNet: $F_{1,59} = 218.510$, $P < 0.001$, $\eta_p^2 = 0.787$; BLNet: $F_{1,59} = 200.543$, $P < 0.001$, $\eta_p^2 = 0.773$) but longer RTs (RTNet: $F_{1,59} = 233.452$, $P < 0.0001$, $\eta_p^2 = 0.798$; BLNet: $F_{1,59} = 186.604$, $P < 0.001$, $\eta_p^2 = 0.760$). However, while CNet and MSDNet also showed a very robust effect on accuracy (CNet: $F_{1,59} = 1{,}116.800$, $P < 0.001$, $\eta_p^2 = 0.950$; MSDNet: $F_{1,59} = 247.520$, $P < 0.001$, $\eta_p^2 = 0.808$), they exhibited a smaller effect for RT (CNet: $F_{1,59} = 11.070$, $P = 0.016$, $\eta_p^2 = 0.158$; MSDNet: $F_{1,59} = 6.171$, $P = 0.002$, $\eta_p^2 = 0.095$). Indeed, of the 60 model instances, only 23 CNet instances and 36 MSDNet instances exhibited an RT increase for more difficult stimuli, whereas this effect was present

https://doi.org/10.1038/s41562-024-01914-8

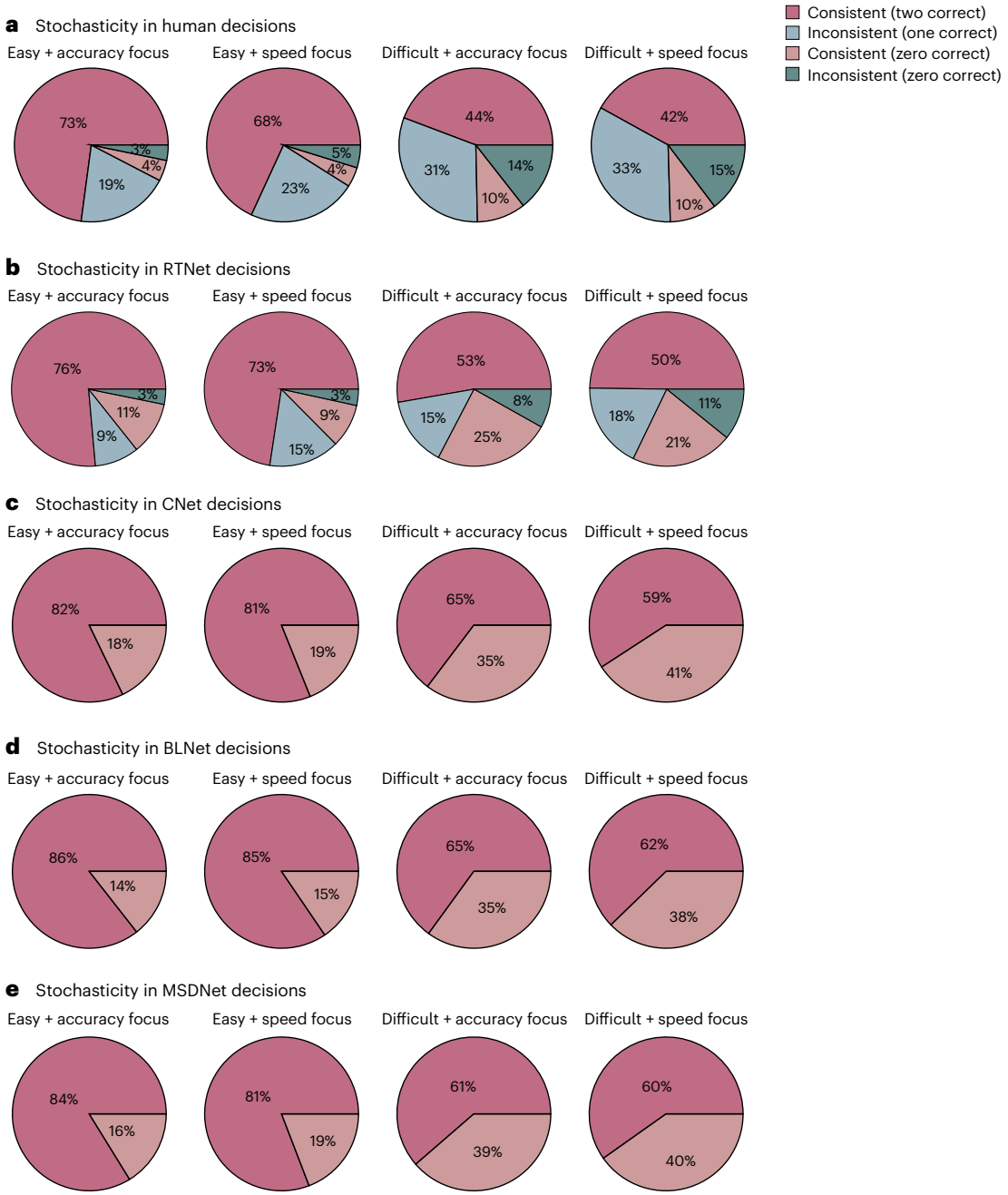


**Fig. 3 | Decision stochasticity in humans and all networks. a**–**e**, Stochasticity of decisions made by humans (**a**), RTNet (**b**), CNet (**c**), BLNet (**d**) and MSDNet (**e**). Warm colours indicate that the same response was given both times an image was presented (whether the response was correct or incorrect), whereas cool colours indicate that different responses were given for the two image presentations (whether or not either of them was correct). Humans and RTNet exhibit stochastic decision-making with stochasticity increasing with task difficulty and speed stress. However, CNet, BLNet and MSDNet in their standard versions are fully deterministic. In the legend, 'Consistent (two correct)' refers to instances when the correct response was given for both presentations of a given image, 'Consistent (zero correct)' refers to instances when the same incorrect choice was made both times, 'Inconsistent (one correct)' refers to instances when only one of the choices was correct and 'Inconsistent (zero correct)' refers to instances when different incorrect choices were made each time.

in 60/60 human participants, 58/60 RTNet instances and 59/60 BLNet instances. These results indicate that the effect of task difficulty on accuracy is exhibited robustly in humans and all networks, but the effect of task difficulty on RT is larger for humans, RTNet and BLNet than for CNet and MSDNet (Discussion).

**Skewness of RT distributions.** For simple two-choice decisions, human RT distributions are generally positively skewed, and the skewness changes as a function of task conditions[2,26]. Our eight-choice task produced RT distributions that closely resemble what is observed in standard two-choice tasks (Fig. 4c). Similar-looking RT distributions were produced by RTNet, but CNet and MSDNet produced RT distributions that, while still right-skewed, exhibited qualitative differences in their shapes (Fig. 4c). BLNet, in contrast, produced RT distributions that were frequently bimodal and left-skewed.

We further assessed how the skewness of the RT distributions changed under different conditions. In humans, we found higher skewness for accuracy than for speed focus ($F_{1,59} = 32.837$, $P < 0.001$, $\eta_p^2 = 0.358$) and higher skewness for easy stimuli than for difficult stimuli ($F_{1,59} = 5.098$, $P = 0.028$, $\eta_p^2 = 0.080$; Fig. 4d). RTNet exhibited

https://doi.org/10.1038/s41562-024-01914-8

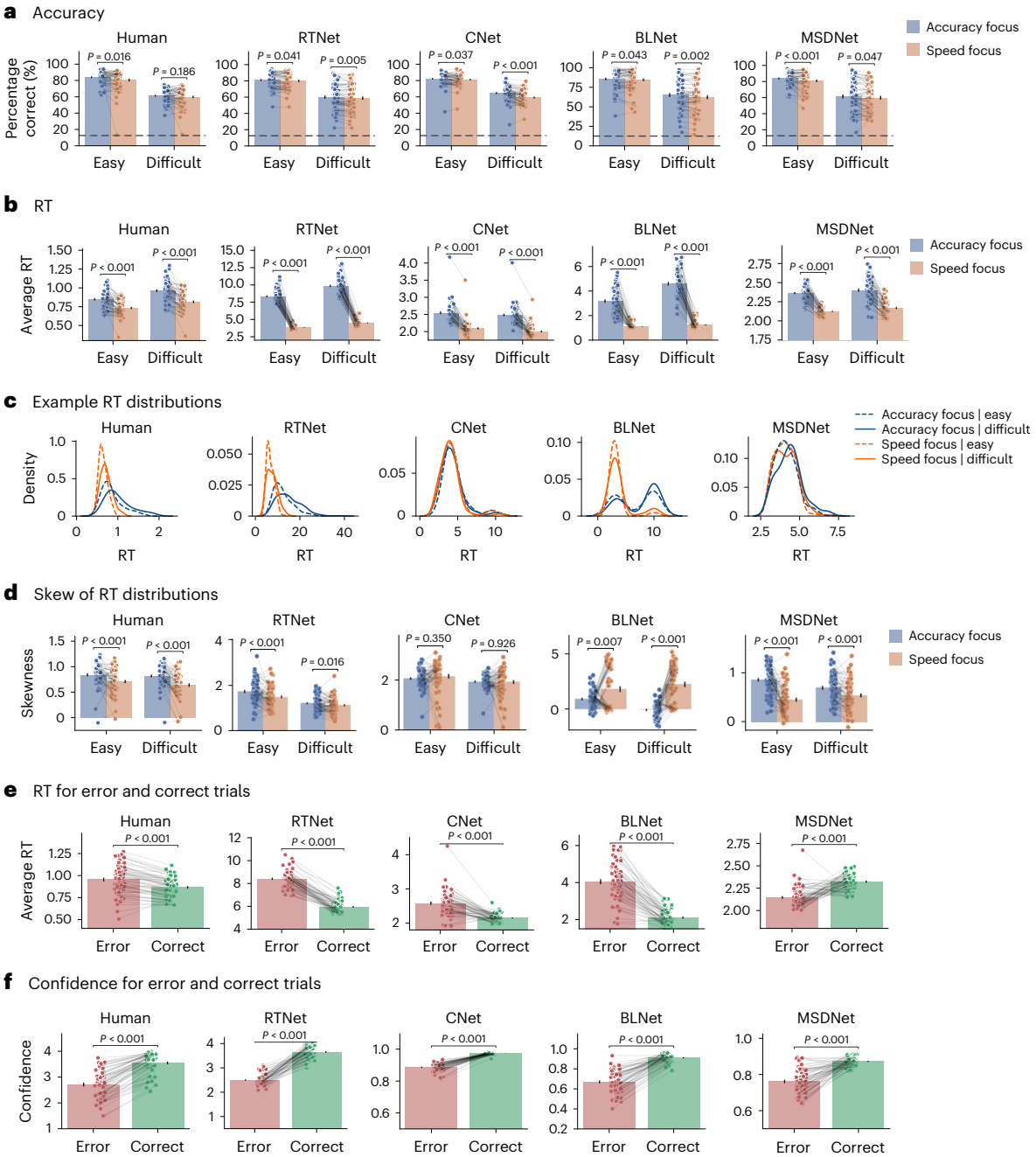


**Fig. 4 | Behavioural effects shown by human participants and the models.** **a**, Accuracy for humans ($n = 60$) decreases when response speed is emphasized as well as for more difficult decisions. Both effects are exhibited by all the networks ($n = 60$ model instances). **b**, RT for humans becomes shorter when response speed is emphasized, as well as for easier decisions. Both effects are also exhibited robustly by RTNet and BLNet. However, while both CNet and MSDNet produced a robust effect for the speed manipulation, they exhibited much smaller effects for the difficulty manipulation. RT for humans is measured in seconds, and RT for the networks is measured in the number of steps over which evidence is accumulated (for RTNet), the number of propagation steps (for CNet), the number of feedforward sweeps (for BLNet) and the number of layers (for MSDNet). **c**, RT distributions for a representative participant/model. **d**, The skewness of RT distributions changes across conditions. For humans and RTNet, the skewness of the RT distributions was higher for easier tasks and for accuracy focus. However, CNet, BLNet and MSDNet showed clear deviations from the human pattern of results. **e**, For humans, RTNet, CNet and BLNet, two-sided paired *t*-tests showed that error trials were associated with higher RT than correct trials. However, MSDNet 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 and all networks. For all panels, the dots represent individual participants, and the error bars show the s.e.m. The *P* values are derived from two-sided Wilcoxon's signed-rank tests (for mean RT comparisons) and two-sided paired *t*-tests (for all other measures).

the same pattern, with skewness increasing with a focus on accuracy ($F_{1,59} = 19.077$, $P < 0.001$, $\eta_p^2 = 0.244$) and with easier stimuli ($F_{1,59} = 93.342$, $P < 0.001$, $\eta_p^2 = 0.613$). For CNet, we found no difference in the skewness of RT distributions between the SAT conditions ($F_{1,59} = 0.428$, $P = 0.515$, $\eta_p^2 = 0.007$), but skewness was higher for easy than for difficult stimuli ($F_{1,59} = 8.612$, $P = 0.005$, $\eta_p^2 = 0.127$). BLNet showed the opposite pattern to CNet, with skewness increasing for the speed focus condition ($F_{1,59} = 39.219$, $P < 0.001$, $\eta_p^2 = 0.399$) and failing

to show a difference in skewness between the easy and difficult stimuli ($F_{1,59}$ = 3.517, $P$ = 0.066, $\eta_p^2$ = 0.056). Finally, while MSDNet showed an increase in skewness with a focus on accuracy ($F_{1,59}$ = 64.866, $P$ < 0.001, $\eta_p^2$ = 0.524), it produced RT distributions that did not significantly differ in skewness between the task difficulty conditions ($F_{1,59}$ = 1.259, $P$ = 0.266, $\eta_p^2$ = 0.021). Overall, RTNet produced RT distributions that reflected the observed patterns in the human data better than all other networks. It should be noted that CNet, BLNet and MSDNet can only produce distinct RTs that are less than or equal to their layer numbers or residual blocks, which may affect their ability to reproduce human RT distributions unless a relatively high number of layers is used. In contrast, RTNet can go through an arbitrary number of samples regardless of the number of layers in its architecture.

**RT is faster for correct trials than for error trials.** Another ubiquitous feature of human behaviour in two-choice tasks is that correct decisions are typically accompanied by faster RTs than incorrect decisions[41–45]. We replicated this effect in our eight-choice task ($F_{1,59}$ = 82.080, $P$ < 0.001, $\eta_p^2$ = 0.582; Fig. 4e). The same difference between correct and error RTs also emerged for RTNet ($F_{1,59}$ = 831.153, $P$ < 0.001, $\eta_p^2$ = 0.934), CNet ($F_{1,59}$ = 83.921, $P$ < 0.001, $\eta_p^2$ = 0.587) and BLNet ($F_{1,59}$ = 286.157, $P$ < 0.001, $\eta_p^2$ = 0.582). However, MSDNet exhibited the opposite pattern such that RTs were faster for error trials than for correct trials ($F_{1,59}$ = 65.696, $P$ < 0.001, $\eta_p^2$ = 0.527). This behaviour is due to the fact that errors produced by MSDNet come mostly from decisions made in earlier layers. It may be possible to reverse this behaviour by using a much more conservative decision threshold in the early than in the late layers of MSDNet, though the effectiveness of this strategy and its effect on all other behavioural signatures examined here would need to be tested. What is clear is that MSDNet in its current form makes a qualitatively wrong prediction regarding the difference between correct and error RTs, whereas RTNet, CNet and BLNet naturally reproduce the empirical effect.

**Confidence is higher for correct trials than for error trials.** Finally, a ubiquitous feature of confidence ratings is that they are higher for correct than for incorrect decisions[46,51]. Our human data replicated this effect ($F_{1,59}$ = 472.172, $P$ < 0.001, $\eta_p^2$ = 0.889; Fig. 4f). The effect was also robustly exhibited by all networks: RTNet ($F_{1,59}$ = 966.796, $P$ < 0.001, $\eta_p^2$ = 0.942), CNet ($F_{1,59}$ = 785.992, $P$ < 0.001, $\eta_p^2$ = 0.930), BLNet ($F_{1,59}$ = 374.031, $P$ < 0.001, $\eta_p^2$ = 0.864) and MSDNet ($F_{1,59}$ = 131.923, $P$ < 0.001, $\eta_p^2$ = 0.691). Therefore, humans and all networks robustly showed higher confidence for correct trials than for incorrect trials.

## Model predictions of responses for individual images

The above results demonstrate that RTNet naturally reproduces all foundational features of human decision-making. In contrast, CNet, BLNet and MSDNet fail to exhibit stochastic decisions and skewness difference in RT distributions between the SAT/difficulty conditions, and MSDNet further fails to account for lower RTs 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[24,42,52]. A critical advantage of RTNet over traditional cognitive models would therefore be the ability to predict human behaviour 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 the networks predict the same quantities in humans.

**Model predictions for individual participants.** In the first set of analyses, we assessed the correlations between the accuracy, RT and confidence for each human participant and the corresponding quantities predicted by RTNet, CNet, BLNet and MSDNet across all four conditions (easy with speed stress, difficult with speed stress, easy with accuracy stress and difficult with accuracy stress). We compared how well the data from individual human participants could be predicted by each model as well as by the data from the 59 remaining human participants. This last quantity, which we call the participant-to-group relationship, provides an estimate of the noise ceiling (that is, the performance that a true model could achieve given inter-participant variability)[10].

We found that all models predicted individual human data much better than chance for accuracy, RT and confidence (two-sided one-sample $t$-tests, all $P$ < 0.001, all Cohen's $d$ > 1.20). The one exception was BLNet, which had a weak negative correlation with human image-by-image accuracy (average $r$ = −0.06; $P$ = 0.002; Cohen's $d$ = 0.410; 95% confidence interval (CI), (−0.09, −0.02)). Critically, RTNet provided substantially better predictions than all other models (Fig. 5). Specifically, two-sided paired $t$-tests showed that RTNet produced better image-by-image predictions about accuracy (RTNet versus CNet: $t_{59}$ = 30.672; $P$ < 0.001; Cohen's $d$ = 4.747; 95% CI, (0.24, 0.27); RTNet versus BLNet: $t_{59}$ = 20.842; $P$ < 0.001; Cohen's $d$ = 3.864; 95% CI, (0.37, 0.44); RTNet versus MSDNet: $t_{59}$ = 30.672; $P$ < 0.001; Cohen's $d$ = 4.747; 95% CI, (0.24, 0.27)), RT (RTNet versus CNet: $t_{59}$ = 18.638; $P$ < 0.001; Cohen's $d$ = 2.370; 95% CI, (0.29, 0.35); RTNet versus BLNet: $t_{59}$ = 13.135; Cohen's $d$ = 0.472; 95% CI, (0.06, 0.08), $P$ < 0.001; RTNet versus MSDNet: $t_{59}$ = 13.318; $P$ < 0.001; Cohen's $d$ = 1.152; 95% CI, (0.13, 0.18)) and confidence (RTNet versus CNet: $t_{59}$ = 8.394; $P$ < 0.001; Cohen's $d$ = 0.936; 95% CI, (0.07, 0.11); RTNet versus BLNet: $t_{59}$ = 6.587; $P$ < 0.001; Cohen's $d$ = 0.391; 95% CI, (0.03, 0.05); RTNet versus MSDNet: $t_{59}$ = 7.68; $P$ < 0.001; Cohen's $d$ = 0.471; 95% CI, (0.04, 0.06)).

RTNet's predictions were reasonably close to the noise ceiling in all cases (calculated as the average participant-to-group correlation in the human data). Specifically, RTNet's predictions were within 62.5%, 79.6% and 64.8% of the noise ceiling for accuracy, RT and confidence, respectively. These numbers were substantially lower for CNet (16.1%, 20.3% and 40.5%), BLNet (0%, 64.4% and 54.1%) and MSDNet (16.1%, 50% and 51.3%). Thus, by reaching between 62.5% and 79.6% of the noise ceiling, RTNet can provide excellent predictions of the accuracy, RT and confidence produced by human participants for images that the model was not trained on. We also derived the model predictions for averages across the 60 participants across all conditions (Supplementary Fig. 3) and found that RTNet still predicts average human accuracy and RT better than the other networks.

**Model predictions within each condition separately.** The above analyses explored the correlations between model predictions and human behaviour 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 performed within each condition separately. We therefore repeated the analyses but within each of the four conditions separately to investigate whether the models can still account for accuracy, RT and confidence on individual images. We found that RTNet, BLNet and MSDNet produced accuracy, RT and confidence predictions that significantly correlate with individual participant data in all conditions (two-sided one-sample $t$-tests, all $P$ < 0.001; Fig. 6). However, while CNet produced accuracy and confidence predictions that significantly correlated with individual participant data in all conditions, the correlations for its RT predictions for all conditions except accuracy focus with difficult images were either zero or negative ($P$ > 0.62).

RTNet predicted the individual data significantly better than the rest of the networks. Specifically, two-sided paired $t$-tests showed that RTNet provided better predictions than CNet in two conditions for accuracy (both $P$ < 0.001), in all four conditions for RT (all $P$ < 0.0001) and in two conditions for confidence (both $P$ < 0.005). Compared with BLNet, RTNet predicted individual data significantly better in three conditions for accuracy (all $P$ < 0.0001) and in all four conditions for RT (all $P$ < 0.025). Compared with MSDNet, RTNet predicted the individual data significantly better in three conditions for accuracy (all $P$ < 0.001) and in all four conditions for RT (all $P$ < 0. 02). There was no

https://doi.org/10.1038/s41562-024-01914-8

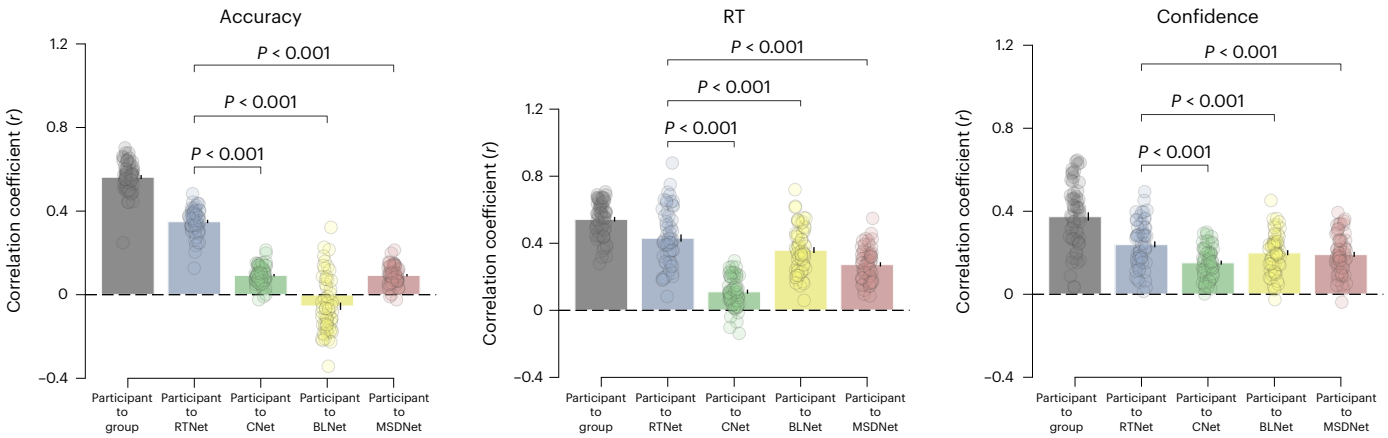


**Fig. 5 | Image-by-image correlation between human data and each model across all experimental conditions for individual participants.** Correlations between data from individual human participants ($n = 60$) and the group average, and correlations between data from individual participants and the average of all 60 instances for RTNet, CNet, BLNet and MSDNet. The correlations were computed separately for accuracy, RT and confidence across all conditions. The correlation is stronger for RTNet than for CNet, BLNet or MSDNet for each measure. The participant-to-group correlation provides an estimate of the noise ceiling for the network correlations. The dots represent individual participants; the error bars show the s.e.m. The *P* values are derived from two-sided paired *t*-tests.

significant difference in confidence predictions between RTNet and BLNet or between RTNet and MSDNet for any of the four conditions (all $P > 0.05$). RTNet was never significantly worse than CNet, BLNet or MSDNet in predicting any of the 12 comparisons. Overall, these results demonstrate that RTNet predicts human behaviour well across all three measures and across different types of analyses (across or within conditions), and it does so better than CNet, BLNet and MSDNet.

### Humans more similar to the group are more similar to RTNet

Our participant-to-group analyses revealed substantial variability in how well individual participants' data corresponded to the group average (Fig. 5). Since the group average constitutes the best model of human behaviour, one would expect that any good, generalizable model of behaviour would also be able to capture this relationship between individual participants and the group average. In other words, the strength of the relationship for an individual participant and the group should be linked to the strength of the relationship of that same participant and the model. Here we tested whether such dependency holds true for RTNet, CNet, BLNet and MSDNet. We found that participants 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 (Pearson's $r = 0.685$; $P < 0.001$; 95% CI, (0.52, 0.80); Fig. 7a). The same correspondence also emerged for RT (Pearson's $r = 0.825$; $P < 0.001$; 95% CI, (0.72, 0.89)) and confidence (Pearson's $r = 0.894$; $P < 0.001$; 95% CI, (0.83, 0.94)). Similar results were obtained for CNet (accuracy: Pearson's $r = 0.389$; $P = 0.002$; 95% CI, (0.15, 0.59); RT: Pearson's $r = 0.432$; $P < 0.001$; 95% CI, (0.20, 0.62); confidence: Pearson's $r = 0.639$; $P < 0.001$; 95% CI, (0.46, 0.77); Fig. 7b) and MSDNet (accuracy: Pearson's $r = 0.389$; $P = 0.002$; 95% CI, (0.15, 0.59); RT: Pearson's $r = 0.80$; $P < 0.0001$; 95% CI, (0.69, 0.88); confidence: Pearson's $r = 0.853$; $P < 0.001$; 95% CI, (0.77, 0.91); Fig. 7d), demonstrating that all three models better predict the data from individuals who behave more similarly to the rest of the group. However, BLNet showed no significant correlation for accuracy predictions (Pearson's $r = -0.029$; $P = 0.828$; 95% CI, (−0.28, 0.23); Fig. 7c), while exhibiting high correlations for RT (Pearson's $r = 0.831$; $P < 0.001$; 95% CI, (0.73, 0.90)) and confidence (Pearson's $r = 0.809$; $P < 0.001$; 95% CI, (0.70, 0.88)). All correlations were higher for RTNet than for the other three networks. These analyses further support the notion that RTNet provides the best model of average human behaviour among the existing alternatives.

To better understand these results, we further examined which participants had the most similar accuracy, RT and confidence to those of the group. We found that different participants had the highest similarity to the group for RT than for accuracy or confidence (Supplementary Fig. 4a–c). RTNet and the other models therefore did not simply provide a good fit to specific participants but instead provided good fits to different groups of participants for different measures. Finally, the individuals closest to the group in their mean accuracy also tended to be those who had the highest task accuracy, suggesting that RTNet and the other models were better at predicting the image-by-image accuracy of participants with higher task performance (Supplementary Fig. 4d).

Given the variability in how similar individual participants' data were to the group data, we also explored how well the models compared with the ability of individual participants to predict the group data. Two-sided paired *t*-tests showed that RTNet outperformed individual human participants in predicting the accuracy ($t_{59} = 4.076$; $P < 0.001$; Cohen's $d = 0.526$; 95% CI, (0.02, 0.06)), RT ($t_{59} = 16.174$; $P < 0.001$; Cohen's $d = 2.088$; 95% CI, (0.2, 0.25)) and confidence ($t_{59} = 10.927$; $P < 0.001$; Cohen's $d = 1.411$; 95% CI, (0.18, 0.26)) of the rest of the group across all conditions (Fig. 8). In fact, RTNet outperformed every individual human participant in predicting the group RT and confidence results, as well as 73.3% of individual participants in predicting accuracy. CNet, by contrast, was significantly worse than individual participants in predicting group accuracy and RT but not confidence (accuracy: $t_{59} = -42.425$; $P < 0.001$; Cohen's $d = 5.477$; 95% CI, (−0.4, −0.39); RT: $t_{59} = -25.439$; $P < 0.001$; Cohen's $d = 3.284$; 95% CI, (−0.38, −0.32); confidence: $t_{59} = -0.361$; $P = 0.719$; Cohen's $d = 0.047$; 95% CI, (−0.05, −0.03)). BLNet was significantly worse than individual participants in predicting group accuracy but predicted group RT and confidence better than individuals (accuracy: $t_{59} = -68.395$; $P < 0.001$; Cohen's $d = 8.830$; 95% CI, (−0.67, −0.63); RT: $t_{59} = 7.018$; $P < 0.001$; Cohen's $d = 0.906$; 95% CI, (0.07, 0.13); confidence: $t_{59} = 6.170$; $P < 0.001$; Cohen's $d = 0.797$; 95% CI, (0.08, 0.16)). Finally, MSDNet's predictions of group accuracy and RT were significantly worse than those of human participants, but its predictions of group confidence were better than those of individual participants (accuracy: $t_{59} = -42.425$; $P < 0.001$; Cohen's $d = 5.477$; 95% CI, (−0.42, −0.39); RT: $t_{59} = -4.019$; $P < 0.001$; Cohen's $d = 0.519$; 95% CI, (−0.08, −0.03); confidence: $t_{59} = 5.266$; $P < 0.001$; Cohen's $d = 0.68$; 95% CI, (0.07, 0.15)). In sum, RTNet was the only network that outperformed

https://doi.org/10.1038/s41562-024-01914-8

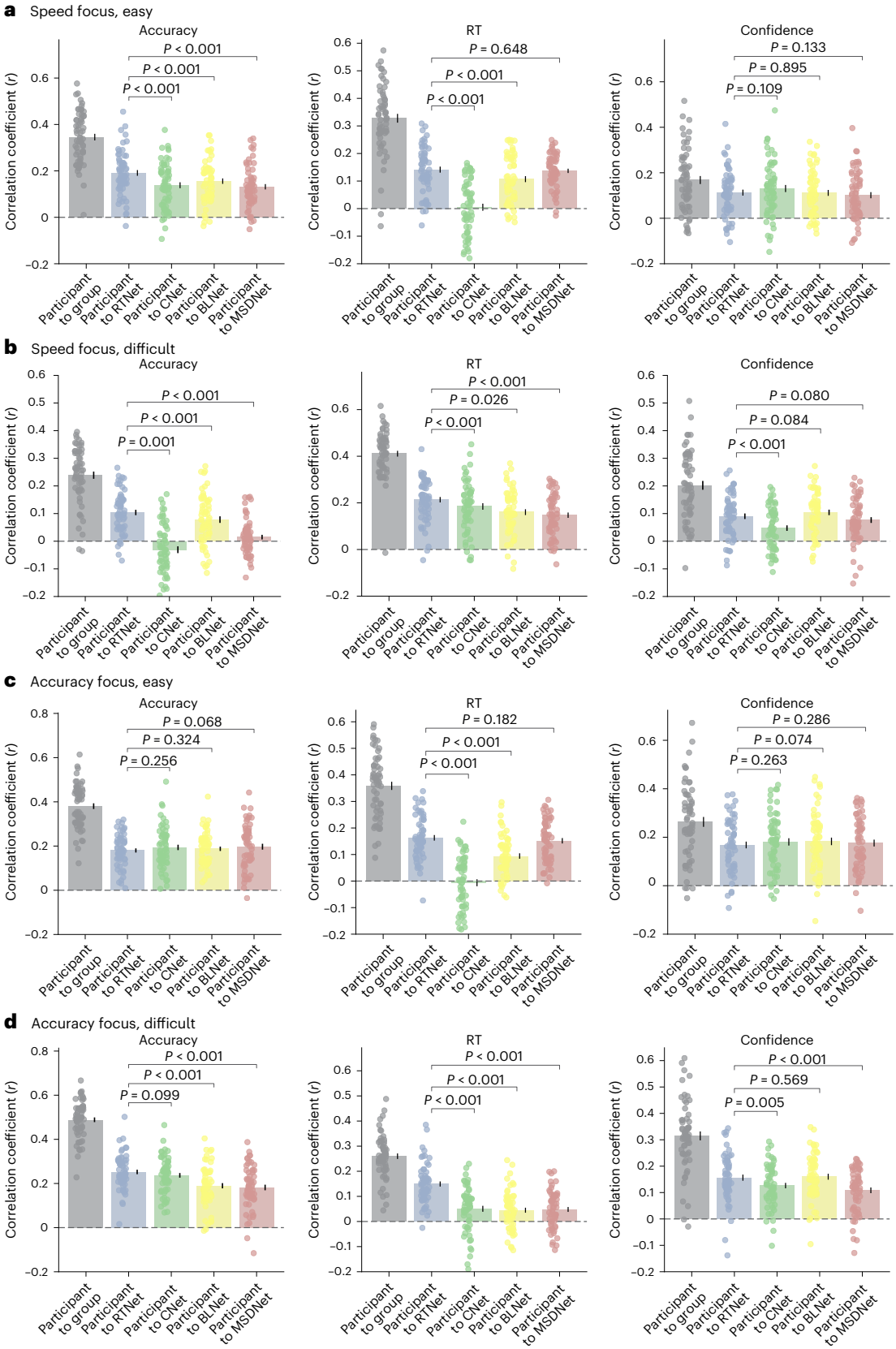


**Fig. 6 | Image-by-image correlation between human data and each network in each experimental condition. a–d**, Correlation between data from individual human participants ($n = 60$) and the group average, as well as the average of all 60 instances for RTNet, CNet, BLNet and MSDNet. The correlations were computed separately for accuracy, RT and confidence in each experimental condition: speed focus, easy (**a**); speed focus, difficult (**b**); accuracy focus, easy (**c**); and accuracy focus, difficult (**d**). The correlation is significantly stronger for RTNet than for CNet (8/12 comparisons), BLNet (7/12 comparisons) and MSDNet (7/12 comparisons). RTNet never exhibits significantly weaker correlations than CNet, BLNet or MSDNet. In all panels, the dots represent individual participants; the error bars show the s.e.m. The *P* values are derived from two-sided paired *t*-tests.

https://doi.org/10.1038/s41562-024-01914-8

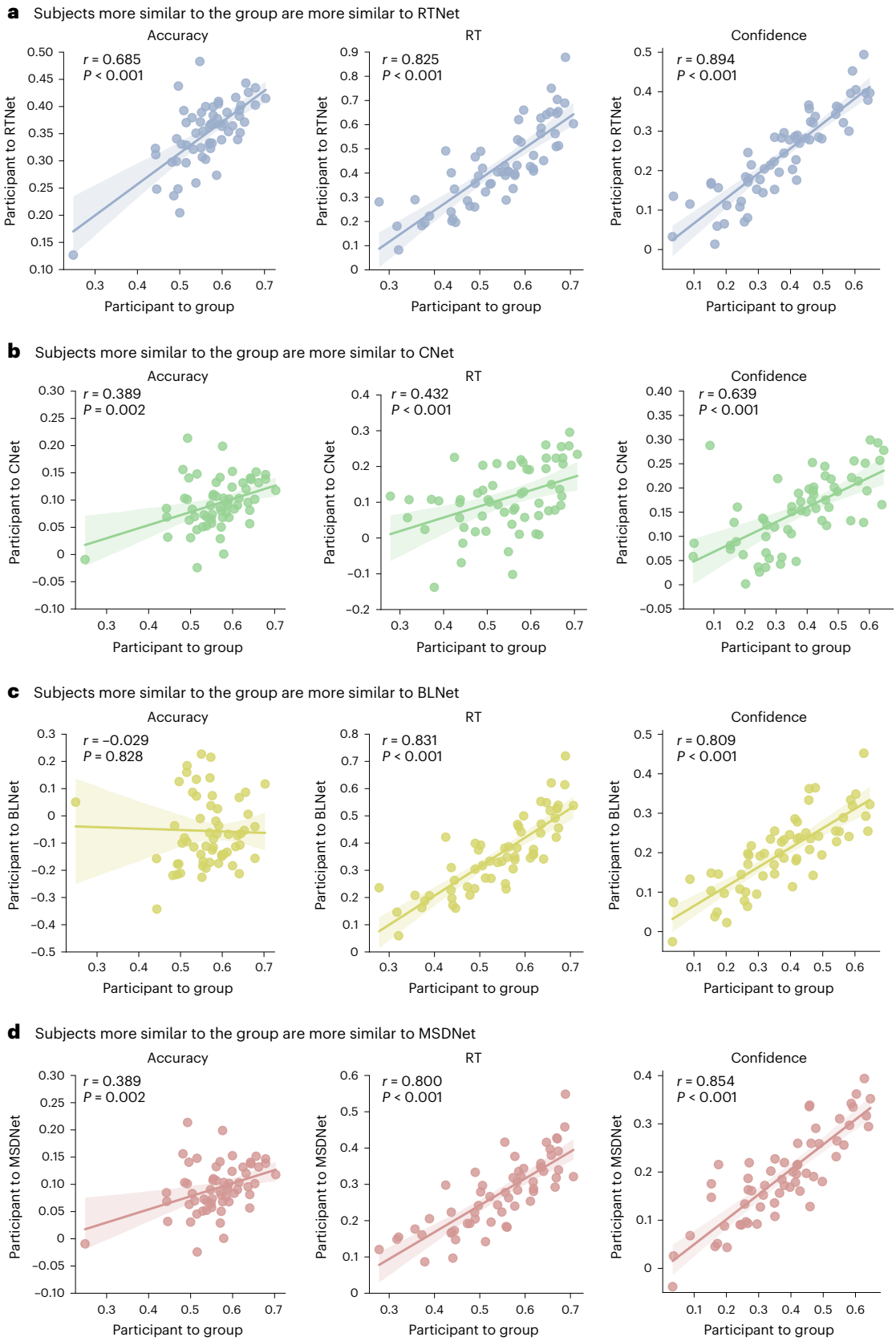


**Fig. 7 | Humans who are more similar to the group average are also more similar to each model. a**, We observed a strong positive correlation between the participant-to-group and participant-to-RTNet similarity values for accuracy, RT and confidence. This finding indicates that individual participants whose behaviour was more similar to the group average on per-image basis were also more similar to the predictions made by RTNet. **b**–**d**, Similar results were also observed for CNet, BLNet (except for accuracy correlations) and MSDNet, although these correlations tended to be lower than those for RTNet. The dots represent individual participants, the lines depict best-fit regressions and the shaded areas depict 95% CIs around the regression estimate.

https://doi.org/10.1038/s41562-024-01914-8

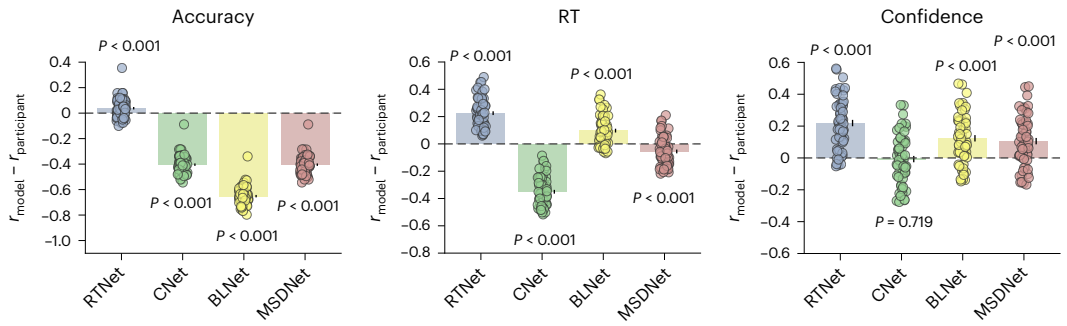


**Fig. 8 | Comparison between individual participants and the models in predicting the group data.** RTNet significantly outperformed individual human participants ($n = 60$) in predicting group accuracy, RT and confidence. In contrast, CNet, BLNet and MSDNet were worse than individual humans in predicting accuracy, and CNet and MSDNet were worse in predicting RT. We note that the effect sizes are very small for RTNet's predictions of accuracy and MSDNet's predictions of RT. However, the effect was sufficiently consistent across participants to make these results statistically significant (RTNet outperformed 44/60 participants in predicting accuracy, and MSDNet did worse than 43/60 participants in predicting RT). In all panels, the dots represent individual participants; the error bars show the s.e.m. The *P* values are derived from two-sided one-sample *t*-tests.

most individual participants in predicting all three measures of human performance (accuracy, RT and confidence).

## Discussion

There is considerable interest in using neural networks as models of human visual processing and behaviour, but relatively little work has been done on testing the extent to which existing image-computable models reproduce the full range of behavioural signatures exhibited by humans. Here we show that the current state-of-the-art neural networks such as CNet, BLNet and MSDNet diverge in several ways from human behaviour. We also develop a new neural network, RTNet, that exhibits all the critical features of human perceptual decision-making, including effects on accuracy, RT and confidence. Furthermore, RTNet predicted human group behaviour for novel images well and did so better than CNet, BLNet and MSDNet, as well as better than individual human participants. Finally, individual humans who were more similar to the group were also more similar to RTNet. Overall, RTNet is a promising image-computable model of human accuracy, RT and confidence.

### Relationship between RTNet and cognitive models of perceptual decision-making

RTNet is a neural network that exhibits all the 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 behaviour[24,26,40,52,53]. 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 two-choice tasks where evidence in favour of one response alternative is also evidence against the other alternative[1,40]. Instead, RTNet is conceptually more similar to another subgroup of sequential sampling models called race models, in which each choice option has its own accumulation system and evidence for each choice is accumulated in parallel[42,54].

Despite their conceptual similarity, RTNet has two important advantages over traditional cognitive models. Most importantly, RTNet is image-computable and can be applied to actual images, whereas traditional models cannot. Traditional models thus 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)[55]. 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 BNN 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.

### Performance differences between RTNet and other networks

RTNet outperformed all other networks we tested (CNet, BLNet and MSDNet) in capturing the signatures of perceptual decision-making. Specifically, while MSDNet and CNet show weaker effects of task difficulty on RTs than humans do, RTNet closely captures the observed magnitude of this effect. RTNet is also the only model that mimics the observed shape and skewness of RT distributions in response to SAT/difficulty manipulations. Finally, RTNet yielded the closest image-by-image predictions of human choice, RT and confidence.

We speculate that RTNet's ability to match observed patterns in human behaviour, particularly RTs, is primarily due to its internal mechanisms being closer to the true mechanisms that give rise to RTs in humans. Specifically, RTNet's core assumption that RTs are generated by a process of sequential sampling and evidence accumulation is inspired from a long tradition of cognitive modelling[1,2]. In fact, these evidence accumulation models have been tested extensively against human data and are currently the best models of human RTs[1,2]. Models such as CNet, BLNet and MSDNet, by contrast, rely on mechanisms that, although they can generate RTs, have not been as extensively validated by empirical tests and are therefore less likely to capture the true mechanisms that generate RTs in humans.

Another reason why CNet and MSDNet may struggle with generating human-like RTs is that the RTs generated by the models are constrained by the number of layers or residual blocks present in the networks. In contrast, RTNet's evidence accumulation mechanism allows flexible generation of RTs across a potentially very large number of steps, thus allowing the RTs to have higher resolution and sensitivity to experimental manipulations.

### Biological plausibility of neural network models of RT

Physiological recordings have uncovered several features of the processing in the human visual system that are relevant to judging the

plausibility of the networks examined here. First, the conduction from one area to another in the visual cortex (roughly corresponding to different layers in neural networks) takes approximately 10 ms[56], with signals from the photoreceptors reaching the top of the visual hierarchy in the inferior temporal cortex in 70–100 ms[57]. A single sweep from input to output in a purely feedforward network should therefore result in decisions with RTs less than a few hundred milliseconds, even though human decisions can range from hundreds of milliseconds to a few seconds. Second, 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[58]. Finally, neuronal processing is known to be noisy such that the same image input generates very different neuronal activations on different trials[37].

MSDNet diverges from these known properties of the human visual cortex in several important ways. To generate meaningful RTs, MSDNet 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 the early visual cortex without further processing in subsequent layers. Moreover, because it 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, MSDNet 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 MSDNet.

In comparison, the dynamics of CNet are closer to those of biological neural networks. Indeed, several of CNet's features—such as parallel and continuous processing of input, and transmission delays between layers—were directly inspired by biology. The transmission delays allow the network to mimic the processing latencies across cortical layers. These features were also found to account for differences in processing efficiency between images such that CNet produced more rapid responses for prototypical images with clear backgrounds than for unusual or cluttered images. However, CNet includes several features that are not biologically plausible, such as its lack of stochasticity of decisions and recurrent processing. It also remains unclear how its cascaded architecture could map onto brain areas[12].

BLNet appears more biologically plausible than both MSDNet and CNet, as it features recurrent visual processing. Lateral connections in recurrent CNNs (RCNNs) enable a layer's activations from previous time steps to feed back into itself, which allows state dependence to naturally emerge in these networks, thus mimicking biological networks[59]. Additionally, RCNNs have been found to generate RTs that align closely with human RTs on a range of complex perceptual tasks involving scene categorization, perceptual grouping and mental simulation[22]. These findings suggest further similarities in perceptual processing between humans and RCNNs. However, in spite of these advantages, RCNNs still lack certain features of biological networks, such as stochasticity of responses.

It is possible to introduce stochasticity in CNet and MSDNet by feeding the outputs of the final softmax layer into a race model. However, such an architecture would imply that response stochasticity arises purely from noise in the decision stage. Although decision noise may exist in humans, contributing to noisy motor responses, stochasticity in human responses is thought to predominantly arise from noisy inference[29] or noisy sensory representations[60–62]. CNNs with additional noise at the decision stage are therefore less biologically plausible than RTNet, which includes noise in the evidence-processing stage.

While also not capturing all properties of visual processing, RTNet appears more biologically plausible than the other models. First, it mimics the noisiness of neuronal responses for repeated presentations of the same stimulus. Second, through the process of evidence accumulation, RTNet naturally generates long-lasting neuronal activations. Third, the network's output is inherently stochastic, unlike that of CNet, BLNet, MSDNet or standard feedforward networks 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[63–66]. Nevertheless, one critical limitation of the biological plausibility of RTNet is its lack of recurrence. That being said, the question of how to train recurrent neural networks on static images remains open[10,57,59,67,68]. Furthermore, while the core of RTNet does not include recurrence, the evidence accumulation system can be thought of as a recurrent network. In fact, several recent studies have 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[69,70]. Future studies should explore how to introduce recurrence into RTNet's mechanisms and whether such modifications can improve its predictions of human behaviour.

### Using noisy weights to generate stochasticity in RTNet's responses

One critical feature of RTNet is that its weights are noisy. Practically, there are many 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[71], 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[72–74]. We therefore chose to obtain the weight variability by training a BNN 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 are currently unaware of any method besides training a BNN that can generate appropriate noise for each weight.

Another alternative to implementing noise in RTNet is to add noise only to the weights in the pre-readout layer (which can mimic noise in the decision process rather than in the sensory processing). As there are many different ways to implement stochasticity in the network, it is important for future studies to test how these differences in implementation affect the model's performance.

RTNet is built such that every time evidence is sampled from a stimulus, the network's weights change randomly (according to the BNN's posterior weight distributions). These random moment-by-moment fluctuations in the network's weights lead to noisy activations. However, in the brain, noisy activations in response to a stimulus are thought to arise from random fluctuations in neuronal activity itself. It can therefore be argued that a more biologically plausible implementation of RTNet would involve noise in unit activations rather than weights[75]. The main reason we chose to add noise in weights rather than activations is the practical ease of implementing BNNs that can naturally generate variability in networks. Mechanistically, however, there may be no meaningful distinction between noisy weights and noisy activations. Indeed, noisy weights lead to noisy activations, which mimic the randomness of neural responses.

### 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[24,42,55], 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 has shown that guessing can be an appropriate behaviour if the observer knows that the task is very difficult[76] or if the observer has been deliberating

for a long time[77]. However, in a race model, guessing can happen at any time point regardless of task difficulty. Nevertheless, human decisions are often suboptimal[78,79], and therefore it is unclear 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, whereas the current state in the human brain is influenced by its previous states[59]. 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, during testing, the connection weight 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 the 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 current state-of-the-art networks for generating RTs. RTNet thus represents an important step in the use of neural networks as models of human decisions.

## Methods

All participants signed informed consent and were compensated for their participation. The protocol was approved by the Georgia Institute of Technology Institutional Review Board (protocol no. H15308). All methods were carried out in accordance with the relevant guidelines and regulations.

### Behavioural experiment

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

**Participants.** Sixty-four participants (31 female; age, 18–32 years) with normal or corrected-to-normal vision were recruited. We had preregistered the collection of only 40 participants, but due to less time restrictions than we had anticipated, and to further increase the statistical power, we collected data from more participants.

**Stimulus, task and procedure.** The participants performed a digit discrimination task where they reported the perceived digit followed by rating their decision confidence. Each trial began with the participants fixating on a small white cross for 500–1,000 ms, followed by a presentation of the stimulus for 300 ms. The stimulus was a digit between 1 and 8 (the digits 0 and 9 were excluded) superimposed on a noisy background. The participants' 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 the participants to respond without looking at the keyboard, thus providing less noisy RTs. Following their categorization response, the participants reported their decision confidence on a four-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 SAT and task difficulty. The SAT was manipulated by asking the participants to emphasize either the speed or the 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 on the basis of pilot testing to produce two different performance levels. Easy and difficult images were randomly interleaved.

The task stimuli were selected from a publicly available dataset of handwritten digits (MNIST)[32]. 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 the MNIST test set to include in our experiment. This ensured that the selected images for the experiment were novel for both the human participants and the trained models. We randomly selected 480 images for the experiment (120 for each condition). The MNIST dataset images are 28 × 28 pixels, which appeared overly small on the computer screens we were using. Therefore, before we added noise, the selected images were first resized to 84 × 84 pixels (using MATLAB's imresize function), and they were padded with the background colour of the MNIST images to size 256 × 256 pixels (visual angle, 6.06°).

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 participants with the experiment. The next two blocks were used to introduce the SAT by asking the participants to focus on accuracy in the first block and on speed in the second. The noise level of the stimuli in these two training blocks was the same as in the main experiment (that is, 0.25 and 0.40 for the easy and difficult stimuli, respectively). During the whole training session, the experimenter was standing beside the participant 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 participant confirmed that he or she understood the task, the experimenter left the room. The participants then completed the main experiment, which consisted of 960 trials organized in four runs, each containing four blocks of 60 trials. Each block consisted of a single SAT condition, and each run included exactly two 'accuracy focus' and two 'speed focus' conditions in a randomized order. At the beginning of each block, the participants 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 the participants 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 the same for all participants, and each image was assigned only to one specific condition.

**Apparatus.** The experiment was designed in the MATLAB v.2020b environment using Psychtoolbox v.3 (ref. 80). The stimuli were presented on a 21.5-inch Dell P2217H monitor (1,920 × 1,080 pixel resolution, 60 Hz refresh rate). The participants were seated 60 cm away from the screen and provided their responses using a keyboard.

### Behavioural analyses

We followed the data analysis steps outlined in our preregistration. All analyses were performed in Python (v.3.10.11) using Google Colab (v.2.0). We first excluded participants who did not follow the speed/accuracy instructions sufficiently well by not providing faster average RTs in the speed focus than in the accuracy focus condition. This resulted in removing two participants (out of 64). We preregistered the exclusion of participants with floor or ceiling effects on accuracy, but no participant met the criteria for exclusion. However, following our preregistration, we excluded two participants because they showed

ceiling effects for confidence. Note that our preregistration document called for excluding participants who provided average confidence of more than 3.7, but because this would have resulted in excluding a much larger number of participants than we had anticipated, we only excluded participants whose average confidence was above 3.85. Therefore, 60 participants 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 participant, 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 $Q_1$ is the RT value at the 25th percentile and $Q_3$ is the RT value at the 75th percentile, we removed values smaller than $Q_1 - 1.5 \times (Q_3 - Q_1)$ and larger than $Q_3 + 1.5 \times (Q_3 - Q_1)$. This step resulted in removing an average of 5.46% of the total trials (with a range of 1.35–8.22% for each participant).

Once these preprocessing steps were completed, we computed the average accuracy, RT, confidence and skewness of the RT distributions separately for each condition. The skewness was computed separately for each individual participant's RT distribution as $\frac{\sum_{i=1}^{N}(x_i-\mu)^3}{(N-1)\sigma^3}$, where $\mu$ and $\sigma$ 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 participants to examine how RT and confidence change as a function of response accuracy. Finally, for visualization purposes, we plotted RT distributions for one participant in Fig. 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[81]. The KDE plots were created using Seaborn's KDE plot with a smoothing bandwidth of 1.2 (ref. 82).

## RTNet

**Network architecture.** The RTNet model consists of two main modules (Fig. 1a). The first module is a BNN, which makes 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[83]. The main difference between a BNN and a standard feedforward neural network is that in a 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. At each processing step, the output of the network (in the form of activations of the final layer) is accumulated towards a predefined threshold. Evidence for each choice option is accumulated separately from the rest, similar to a race model[24].

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 RT produced by RTNet is simply the number of samples used to reach the decision threshold. The confidence of the model was obtained by taking the difference in evidence scores between the chosen response and the second-best choice.

**Implementation.** We implemented RTNet using the AlexNet architecture, which has eight layers with learnable parameters[33]. The AlexNet architecture consists of five convolutional layers with a combination of max pooling followed by three fully connected layers. We chose to implement RTNet within a relatively large-scale CNN such as AlexNet (rather than a shallow network, which may have also been able to learn to classify the MNIST dataset) because our goal was to eventually compare our model to others such as CNet and MSDNet, which are generally based on larger CNNs and work on multiple existing datasets. Additionally, difficulties associated with training BNNs limited us to relatively small network structures (rather than VGG or ResNet models). We found the AlexNet architecture to be a reasonable compromise in this trade-off between model complexity and ease of training BNNs. RTNet was implemented in PyTorch[84], while the Bayesian networks were implemented using Pyro[85], which is a probabilistic programming library built on PyTorch[84].

**Training the BNN module of RTNet.** BNNs are probabilistic models that incorporate uncertainty into their weights and biases, rather than treating them as point estimates. Consider a training dataset, $x$, for which we must predict the class labels, $y$. In traditional neural networks, the predicted class label, $\hat{y}$, is a function of the network's weights, $w$, and these weights are tuned to optimize the correspondence between the predicted ($\hat{y}$) and true class labels ($y$). In BNNs, however, weights are modelled as probability distributions instead of point estimates. Following the rules of Bayesian inference, one can infer the posterior distribution of these weights ($w$) using the formula $p(w|x) = \frac{p(w,x)}{p(x)}$. However, this computation is intractable for large networks since it involves computing the marginal likelihood of the data $p(x)$ across all possible configurations of weights. Therefore, computing this posterior distribution is typically done using a method of approximation called variational inference. A stand-in distribution, $q(w)$, is specified to approximate the posterior, and its parameters are tuned to maximize the similarity between the two distributions. The similarity between the distributions is quantified by the information theoretical measure called Kullback–Liebler (KL) divergence:

$$\mathrm{KL}[q(w)||p(w|x)] = E_q[\log q(w) - \log p(w,x)] + \log p(x) \quad (1)$$

Although $\mathrm{KL}[q(w)||p(w|x)]$ cannot be directly computed since $p(x)$ is intractable, one can side-step this computation by defining a surrogate objective function called the evidence lower bound (ELBO) function as:

$$\mathrm{ELBO}(q) = E_q[\log p(w,x) - \log q(w)] \quad (2)$$

where both $p(w, x)$ and $q(w)$ are tractable, and due to their negative relationship, maximizing $\mathrm{ELBO}(q)$ thus results in the minimization of $\mathrm{KL}[q(w)||p(w|x)]$, allowing one to approximate the true posterior distribution of the network's weights.

We trained the network to achieve classification accuracy higher than 97% on the MNIST test set. We trained the BNN module of RTNet for a total of 15 epochs with a batch size of 500. We used the ELBO loss function[86] and Adam[87] for optimization with a learning rate of 0.001, and the default values for weight decay and epsilon (weight decay, 0; $\varepsilon = 10^{-8}$). To ensure that each BNN performs better than 97% on the MNIST test set, we followed a specific rule for each model instance. When testing an image with the BNN module of RTNet, we sampled ten times from the posterior distributions learned during training and thus obtained ten unique responses for each image. The response with the highest frequency among the ten responses was chosen as the final decision of the BNN module. Note that there were no RTs generated at this step since we only implemented the BNN module of RTNet and generated a set of responses that would allow us to evaluate how well the BNN's posterior distributions had been trained. These trained BNN models were later used to generate variable activations for the evidence accumulation process that resulted in RTs.

We resized the MNIST images to the standard input size to the AlexNet model architecture (227 × 227 pixels). We also normalized the input images to have a mean of 0.1307 and a standard deviation of 0.3081, which is a standard procedure when using AlexNet for classification of the ImageNet dataset[88]. We trained 60 instances of RTNet using the above procedure but with different weight initializations for each network instance. We used a different combination of mean and

standard deviation values for each of the 60 instances to maximize differences in network initializations. Specifically, different network instances 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 standard deviations of the weights and biases were set to a value ranging from 1 to 5 with increments of 1 (for a total of $12 \times 5 = 60$ instances).

**Generating RTNet's responses from the evidence accumulation module.** Sequential sampling models belong to a class of cognitive models that assume that observers make decisions by repeated sampling and accumulation of noisy evidence until a threshold is reached[1,2]. In these models, RT reflects the number of sampling steps required to reach the threshold. RTNet utilizes this evidence accumulation mechanism to generate RTs. To generate noisy evidence, we used the probability distribution of weights in the BNNs to randomly sample one unique feedforward network at each time step. At each time step, $t$, the presented image results in a feedforward sweep of the sampled network and generates a set of activations ($a_t$) where $a_t = [a_{1,t}, a_{2,t}...a_{8,t}]$ are the values obtained in the last layer after the softmax function has been applied. Each unit in the output layer corresponds to the activation for one of the eight choice options, and for each choice, the evidence obtained at the current step is added to the sum of evidence collected from all previous steps. Thus, a running total of accumulated evidence is maintained such that $a_i = \sum_{t=1}^{n} a_{i,t}$, where $n$ refers to the total number of steps over which evidence has been accumulated and $i \in [1,8]$ refers to the response option. When the total evidence in favour of any of the options exceeds a predefined threshold $k$, the corresponding response option is chosen such that the network's response $r = \text{argmax}(a_1, a_2...a_8)$ at the time step when $\max(a_1, a_2...a_8) \geq k$.

What are the properties of evidence accumulation? Everything else being equal, decisions that are based on fewer evidence samples are more likely to be influenced by chance fluctuations in evidence that favour incorrect decisions. However, when the model is allowed to accumulate evidence over a longer period, these random variations are more likely to cancel out, thus increasing the likelihood of a correct response. In turn, because a longer period of accumulation leads, on average, to stronger evidence, this directly results in higher confidence.

### CNet

**Network architecture.** CNet builds on the architecture of residual networks (ResNet) by utilizing skip connections to introduce propagation delays during input processing (Fig. 1b). At each processing step, all units in all layers are updated parallelly. However, due to the propagation delays introduced by each residual block, simpler perceptual features get transmitted faster across blocks. For instance, at the first time step, only the first residual block receives input, and model predictions at this step are based only on the computations of the first residual block. At the second time step, all the other layers receive partial input from the first block. Even though the model prediction at this point will be based on computations from all blocks, only the first block will have received complete input and achieved stable output. The other blocks will only contain partial updates from the lower block, and therefore their output will not be stable. In general, a residual block, $t$, takes $t-1$ time steps to receive complete and stable input. At any point during processing, the network can generate a prediction since all the residual blocks contribute to the computations. However, if the time step ($t$) is less than the number of residual blocks, the responses will be based on unstable representations in the higher blocks. Due to this architecture, the network's responses are subject to a trade-off between speed and complexity of processing. Decision time is indicated by the processing step at which the decision was made, and decision confidence is derived from the softmax value in the final layer, at the time of decision. The softmax values are obtained by transforming the activation scores ($z$) of all nodes in the output layer according to the function: $\frac{e^{z_i}}{\sum_j^n e^{z_j}}$, where $i$ refers to the node whose output is being transformed and $n$ refers to the number of nodes in the output layer (which is equal to the number of classes).

**Implementation.** CNet was implemented using the architecture of ResNet-18 (ref. 9) since it requires networks with skip connections. ResNet-18 architecture consists of 17 convolutional layers, where 16 of these layers are embedded within eight residual blocks (skip connections), followed by a final fully connected layer with softmax activation to generate the decision. The network was implemented in PyTorch[84].

**Network training.** We trained CNet using the same procedure that was used by the original authors since their training protocol was found to yield the best network behaviour and performance. The network achieved an accuracy of >97% with 12 training epochs and a batch size of 500. The models were trained on a temporal-difference learning procedure along with cross-entropy loss. In the original publication, temporal-difference learning was found to perform better than softmax-based cross-entropy loss in encouraging correct responses to emerge faster. The loss function was optimized using an initial learning rate of 0.01, a weight decay of 0.005 and a momentum of 0.9. The images were normalized to a mean of 0.1307 and a standard deviation of 0.3081. We trained 60 instances of CNet using the above procedure but using a different random seed for initializing the network's weights to allow individual differences in the network's learning.

### BLNet

**Network architecture.** BLNet is an RCNN consisting of a standard feedforward CNN with recurrent connections that connect each layer to itself[10] (Fig. 1c). A final readout layer computes the network's output at each time step via a softmax function. Time steps are defined as the number of feedforward sweeps of the network that have occurred until the time at which the readout is evaluated. At each time step, a given layer receives input from two sources—the feedforward input from the previous convolutional layer and recurrent input from itself in the form of activations from the previous time step. The readout is evaluated at each time step such that if it exceeds a predefined threshold, the network generates a response. The response corresponds to the choice that generates the highest softmax value, and the time step at which the response was made indicates the decision time. The softmax value associated with the choice at the time of decision indicates the decision confidence. The network's ability to trade off speed and accuracy comes from the fact that higher softmax thresholds require more feedforward and recurrent computations, which effectively results in a deeper network being unrolled across time, which, in turn, leads to both higher RT and higher accuracy.

**Implementation.** BLNet was implemented as a custom-built network consisting of seven convolutional layers of increasing size and a final readout layer, as defined by the original authors[10]. Each layer consists of two sets of weights—the bottom-up weights that transform the input from the previous layer and the lateral weights that act on recurrent input that the layer receives from itself. The readout layer is a fully connected layer with softmax activation to generate the decision. The network was unrolled across time for eight time steps and was implemented using TensorFlow.

**Network training.** We were able to achieve a test accuracy of >97% with only three epochs with a batch size of 32 and a sparse categorical cross-entropy loss function[89]. Adam[87] was used for optimization with a learning rate of 0.001. For testing, the response at the final time step was taken as the network's decision. We resized the MNIST images to the standard input size of $128 \times 128$ pixels defined for the network. We

trained 60 instances of BLNet using the above procedure but using a different random seed for initializing the network's weights to allow individual differences in the network's learning.

**Testing.** Unlike the other networks, BLNet exhibited an overall accuracy that was about 5% greater for the 120 images used in the easy, speed focus condition than for the 120 images used in the easy, accuracy focus condition. This resulted in a lack of the expected accuracy difference between these two conditions when BLNet was run on all images (Supplementary Fig. 5). On further investigation, we found that for each condition, the image set contained a small subset of images for which the network showed chance-level performance (12.5%). The image set for the easy, accuracy focus condition contained more such images than the image set for the easy, speed focus condition, explaining the observed accuracy differences. Therefore, when testing BLNet on the effects reported in Fig. 4, we excluded this subset of images for all conditions (10 of 480 images). This exclusion led to BLNet showing the expected SAT (Fig. 4a,b).

### MSDNet

**Network architecture.** MSDNet has an architecture similar to a standard feedforward neural network but with early-exit classifiers after each of its layers (Fig. 1d). 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 the decision confidence[13,89].

**Implementation.** We implemented MSDNet using the AlexNet architecture, which has eight layers with learnable parameters[33]. The AlexNet architecture consists of five convolutional layers with a combination of max pooling followed by three fully connected layers. In addition to the standard AlexNet structure, we incorporated additional readout layers located right after each layer of processing. The feature map size of all these readout layers was set to the number of classes. The network was implemented in PyTorch[84].

**Network training.** Due to MSDNet's deterministic nature, only three epochs with a batch size of 500 were enough to achieve a test accuracy of more than 97% with the same batch size and a weighted cumulative loss function[89]. Adam[87] was used for optimization with a learning rate of 0.001. For testing, the response of the last output layer was taken as the network's decision. If a network did not achieve an accuracy greater than 97%, we started the training over with the same initial values. Since MSDNet is also built on the architecture of AlexNet, we resized the MNIST images to the standard input size for AlexNet and normalized the images to have a mean of 0.1307 and a standard deviation of 0.3081. To make the initializations of MSDNet as similar as possible to the initializations of RTNet, for each RTNet instance, we set the initial values for the weights and biases of the MSDNet instance 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 all 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). Lower noise levels lead to higher accuracy, whereas higher threshold parameters lead to longer processing and RTs (and contribute to higher accuracy levels).

The 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 (where the noise amplitude refers to the length of the interval over which the noise values are generated) 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 for easy images and 4.1 for difficult images, and threshold values of 3 for the speed condition and 6 for the accuracy condition. With these threshold and noise parameters, the evidence accumulation process in RTNet executed 6.5 sampling steps on average, although the distributions were wide such that the actual steps varied from 1 to 35. However, the number of processing steps depended on the experimental manipulation, with the number of steps increasing both for difficult images and with stress on accuracy over speed (the average number of steps observed for each condition corresponds to the height of the bars for RTNet in Fig. 4b).

We used a similar procedure to tune the parameters of CNet, BLNet and MSDNet. Note that the threshold value for CNet is the softmax evidence at the output layer. The coarse search was performed using threshold values between 0.5 and 0.9 with increments of 0.04. The results were closest to the human accuracy levels when the threshold was in the range 0.79–0.83 for the speed focus condition and 0.86–0.9 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 0.83 for the speed condition and 0.9 for the accuracy condition. For noise levels, the best match to human accuracy was obtained when the noise levels were set to 1.42 for easy images and 1.83 for difficult images.

For BLNet, like CNet, the threshold value is the softmax evidence at the output layer. The coarse search was performed using threshold values between 0.1 and 0.95 with increments of 0.2. The results were closest to the human accuracy levels when the threshold was in the range 0.4–0.5 for the speed focus condition and 0.9–0.95 for the accuracy focus condition. We then performed a fine search in these ranges by incrementing the threshold by steps of 0.05. The closest match to human accuracy was achieved for a threshold value of 0.4 for the speed condition and 0.95 for the accuracy condition. For noise levels, the best match to human accuracy was obtained when the noise levels were set to 0.55 for easy images and 1.2 for difficult images.

The threshold value for MSDNet is the softmax evidence at each early exit. The coarse search was performed using 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 the 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, the best match was obtained when the noise levels were set to 1.9 for easy images and 3.0 for difficult images.

Although we tried to closely match each network's accuracy with that of humans for each condition, our ability to do this was limited by the fact that a given SAT threshold must predict accuracies for both the easy and difficult conditions and a given noise level must predict accuracies for both the SAT conditions. We therefore obtained parameter estimates that resulted in closely (but not exactly) matched accuracies.

**Reporting summary**
Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability
The behavioural data have been made publicly available at https://osf.io/akwty.

## Code availability
All code and trained models are publicly available at https://osf.io/akwty.

## References

1. Ratcliff, R. A theory of memory retrieval. *Psychol. Rev.* **85**, 59–108 (1978).
2. Ratcliff, R. & McKoon, G. The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput.* **20**, 873–922 (2008).
3. Ashby, F. G. & Townsend, J. T. Varieties of perceptual independence. *Psychol. Rev.* **93**, 154–179 (1986).
4. Green, D. M. & Swets, J. A. *Signal Detection Theory and Psychophysics* (John Wiley, 1966).
5. Kriegeskorte, N. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* **1**, 417–446 (2015).
6. Kriegeskorte, N. & Golan, T. Neural network models and deep learning. *Curr. Biol.* **29**, R231–R236 (2019).
7. Kietzmann, T. C., McClure, P. & Kriegeskorte, N. Deep neural networks in computational neuroscience. *Oxf. Res. Encycl. Neurosci.* https://doi.org/10.1093/ACREFORE/9780190264086.013.46 (2019).
8. Yamins, D. L. K. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
9. Iuzzolino, M. L., Mozer, M. C. & Bengio, S. Improving anytime prediction with parallel cascaded networks and a temporal-difference loss. *Adv. Neural Inf. Process. Syst.* **33**, 27631–27644 (2021).
10. Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I. & Kriegeskorte, N. Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLoS Comput. Biol.* **16**, e1008215 (2020).
11. Zhang, L. et al. SCAN: A Scalable Neural Networks Framework Towards Compact and Efficient Models. in *Advances in Neural Information Processing Systems 32* Vol. 32 (eds Wallach, H. et al.) (Curran Associates, 2019).
12. Subramanian, A., Sizikova, E., Kumbhar, O., Majaj, N. & Pelli, D. G. Benchmarking dynamic neural-network models of the human speed–accuracy trade off. *J. Vis.* **22**, 4359 (2022).
13. Huang, G. et al. Multi-scale dense networks for resource efficient image classification. In *Proc. 6th International Conference on Learning Representations* (ICLR, 2018).
14. Kalanthroff, E., Davelaar, E. J., Henik, A., Goldfarb, L. & Usher, M. Task conflict and proactive control: a computational theory of the Stroop task. *Psychol. Rev.* **125**, 59–82 (2018).
15. Mewhort, D. J. K., Braun, J. G. & Heathcote, A. Response time distributions and the Stroop task: a test of the Cohen, Dunbar, and McClelland (1990) model. *J. Exp. Psychol. Hum. Percept. Perform.* **18**, 872–882 (1992).
16. Cohen, J. D., Dunbar, K. & McClelland, J. L. On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychol. Rev.* **97**, 332–361 (1990).
17. Koivisto, M., Railo, H., Revonsuo, A., Vanni, S. & Salminen-Vaparanta, N. Recurrent processing in V1/V2 contributes to categorization of natural scenes. *J. Neurosci.* **31**, 2488–2492 (2011).
18. Tang, H. et al. Recurrent computations for visual pattern completion. *Proc. Natl Acad. Sci. USA* **115**, 8835–8840 (2017).
19. Lamme, V. A. F. & Roelfsema, P. R. The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* **23**, 571–579 (2000).
20. Kar, K. & DiCarlo, J. J. Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron* **109**, 164–176.e5 (2021).
21. Kar, K., Kubilius, J., Schmidt, K., Issa, E. B. & DiCarlo, J. J. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.* **22**, 974–983 (2019).
22. Goetschalckx, L. et al. Computing a human-like reaction time metric from stable recurrent vision models. In *Advances in Neural Information Processing Systems* (eds Oh, A. et al.) 14338–14365 (Curran Associates, 2023).
23. Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J. D. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* **113**, 700–765 (2006).
24. Heathcote, A. & Matzke, D. Winner takes all! What are race models, and why and how should psychologists use them? *Curr. Dir. Psychol. Sci.* **31**, 383–394 (2022).
25. Vickers, D. Evidence for an accumulator model of psychophysical discrimination. *Ergonomics* **13**, 37–58 (2007).
26. Forstmann, B. U., Ratcliff, R. & Wagenmakers, E.-J. Sequential sampling models in cognitive neuroscience: advantages, applications, and extensions. *Annu. Rev. Psychol.* **67**, 641–666 (2016).
27. Rahnev, D. Confidence in the real world. *Trends Cogn. Sci.* **24**, 590–591 (2020).
28. Yeon, J. & Rahnev, D. The suboptimality of perceptual decision making with multiple alternatives. *Nat. Commun.* **11**, 3857 (2020).
29. Drugowitsch, J., Wyart, V., Devauchelle, A. D. & Koechlin, E. Computational precision of mental inference as critical source of human choice suboptimality. *Neuron* **92**, 1398–1411 (2016).
30. Li, H. H. & Ma, W. J. Confidence reports in decision-making with multiple alternatives violate the Bayesian confidence hypothesis. *Nat. Commun.* **11**, 2004 (2020).
31. Churchland, A. K., Kiani, R. & Shadlen, M. N. Decision-making with multiple alternatives. *Nat. Neurosci.* **11**, 693–702 (2008).
32. Deng, L. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Process. Mag.* **29**, 141–142 (2012).
33. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25 (NIPS 2012)* (eds Pereira, F. et al.) (Curran Associates, 2012); https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html
34. Geirhos, R. et al. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)* (eds Bengio, S. et al.) (Curran Associates, 2018); https://proceedings.neurips.cc/paper/2018/hash/0937fb5864ed06ffb59ae5f9b5ed67a9-Abstract.html
35. Geirhos, R. et al. Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv* https://doi.org/10.48550/arxiv.1706.06969 (2017).
36. Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E. & Pouget, A. Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron* **74**, 30–39 (2012).
37. Renart, A. & Machens, C. K. Variability in neural activity and behavior. *Curr. Opin. Neurobiol.* **25**, 211–220 (2014).

38. Heitz, R. P. The speed–accuracy trade off: history, physiology, methodology, and behavior. *Front. Neurosci.* **8**, 150 (2014).
39. Heitz, R. P. & Schall, J. D. Neural mechanisms of speed–accuracy trade off. *Neuron* **76**, 616–628 (2012).
40. Ratcliff, R. & Rouder, J. N. Modeling response times for two-choice decisions. *Psychol. Sci.* **9**, 347–356 (1998).
41. Wagenmakers, E.-J. & Brown, S. On the linear relation between the mean and the standard deviation of a response time distribution. *Psychol. Rev.* **114**, 830–841 (2007).
42. Brown, S. & Heathcote, A. The simplest complete model of choice response time: linear ballistic accumulation. *Cogn. Psychol.* **57**, 153–178 (2008).
43. Forstmann, B. U. et al. Striatum and pre-SMA facilitate decision-making under time pressure. *Proc. Natl Acad. Sci. USA* **105**, 17538–17542 (2008).
44. Luce, R. D. *Response Times* (Oxford Univ. Press, 1986); https://doi.org/10.1093/acprof:oso/9780195070019.001.0001
45. Ratcliff, R. 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. *Psychon. Bull. Rev.* **9**, 278–291 (2002).
46. Rahnev, D. Visual metacognition: measures, models, and neural correlates. *Am. Psychol.* **76**, 1445–1453 (2021).
47. Wyart, V. & Koechlin, E. Choice variability and suboptimality in uncertain environments. *Curr. Opin. Behav. Sci.* **11**, 109–115 (2016).
48. Findling, C. & Wyart, V. Computation noise in human learning and decision-making: origin, impact, function. *Curr. Opin. Behav. Sci.* **38**, 124–132 (2021).
49. Rafiei, F. & Rahnev, D. Qualitative speed–accuracy tradeoff effects that cannot be explained by the diffusion model under the selective influence assumption. *Sci. Rep.* **11**, 45 (2021).
50. Gold, J. I. & Shadlen, M. N. The neural basis of decision making. *Annu. Rev. Neurosci.* https://doi.org/10.1146/annurev.neuro.29.051605.113038 (2007).
51. Yeung, N. & Summerfield, C. Metacognition in human decision-making: confidence and error monitoring. *Phil. Trans. R. Soc. B* **367**, 1310–1321 (2012).
52. Heathcote, A. & Love, J. Linear deterministic accumulator models of simple choice. *Front. Psychol.* **3**, 292 (2012).
53. Ratcliff, R. & Smith, P. L. A comparison of sequential sampling models for two-choice reaction time. *Psychol. Rev.* **111**, 333–367 (2004).
54. Brown, S. & Heathcote, A. A ballistic model of choice response time. *Psychol. Rev.* **112**, 117–128 (2005).
55. Tillman, G., Van Zandt, T. & Logan, G. D. Sequential sampling models without random between-trial variability: the racing diffusion model of speeded decision making. *Psychon. Bull. Rev.* **27**, 911–936 (2020).
56. Mizuseki, K., Sirota, A., Pastalkova, E. & Buzsáki, G. Theta oscillations provide temporal windows for local circuit computation in the entorhinal–hippocampal loop. *Neuron* **64**, 267–280 (2009).
57. Nayebi, A. et al. Task-driven convolutional recurrent models of the visual system. *Adv. Neural Inf. Process. Syst.* **31**, 5290–5301 (2018).
58. Issa, E. B., Cadieu, C. F. & Dicarlo, J. J. Neural dynamics at successive stages of the ventral visual stream are consistent with hierarchical error signals. *eLife* **7**, e42870 (2018).
59. van Bergen, R. S. & Kriegeskorte, N. Going in circles is the way forward: the role of recurrence in visual inference. *Curr. Opin. Neurobiol.* **65**, 176–193 (2020).
60. Kaufman, M. T. & Churchland, A. K. Sensory noise drives bad decisions. *Nature* **496**, 172–173 (2013).
61. Brunton, B. W., Botvinick, M. M. & Brody, C. D. Rats and humans can optimally accumulate evidence for decision-making. *Science* **340**, 95–98 (2013).
62. Osborne, L. C., Lisberger, S. G. & Bialek, W. A sensory source for motor variation. *Nature* **437**, 412–416 (2005).
63. Huk, A. C. & Shadlen, M. N. Neural activity in macaque parietal cortex reflects temporal integration of visual motion signals during perceptual decision making. *J. Neurosci.* **25**, 10420–10436 (2005).
64. Huk, A. C., Katz, L. N. & Yates, J. L. The role of the lateral intraparietal area in (the study of) decision making. *Annu. Rev. Neurosci.* **40**, 349–372 (2017).
65. Bahl, A. & Engert, F. Neural circuits for evidence accumulation and decision making in larval zebrafish. *Nat. Neurosci.* **23**, 94–102 (2019).
66. Hanks, T. D., Kiani, R. & Shadlen, M. N. A neural mechanism of speed–accuracy tradeoff in macaque area LIP. *eLife* **3**, e02260 (2014).
67. Spoerer, C. J., McClure, P. & Kriegeskorte, N. Recurrent convolutional neural networks: a better model of biological object recognition. *Front. Psychol.* **8**, 1551 (2017).
68. Kietzmann, T. C. et al. Recurrence is required to capture the representational dynamics of the human visual system. *Proc. Natl Acad. Sci. USA* **116**, 21854–21863 (2019).
69. Schwarzschild, A. et al. Can you learn an algorithm? Generalizing from easy to hard problems with recurrent networks. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)* (eds Ranzato, M. et al.) (Curran Associates, 2021); https://proceedings.neurips.cc/paper/2021/hash/3501672ebc68a5524629080e3ef60aef-Abstract.html
70. Zhou, D. et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv* https://doi.org/10.48550/arxiv.2205.10625 (2022).
71. Saltelli, A. et al. Sensitivity analysis for neural networks: natural computing. *Risk Anal.* **159**, 179–201 (2009).
72. Ko, J. H., Kim, D., Na, T., Kung, J. & Mukhopadhyay, S. Adaptive weight compression for memory-efficient neural networks. In *Proc. 2017 Design, Automation and Test in Europe* 199–204 (IEEE, 2017); https://doi.org/10.23919/DATE.2017.7926982
73. Koutník, J., Gomez, F. & Schmidhuber, J. Evolving neural networks in compressed weight space. In *Proc. 12th Annual Genetic and Evolutionary Computation Conference* 619–625 (Association for Computing Machinery, 2010); https://doi.org/10.1145/1830483.1830596
74. Kung, J., Kim, D. & Mukhopadhyay, S. A power-aware digital feedforward neural network platform with backpropagation driven approximate synapses. In *Proc. 2015 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)* 85–90 (IEEE, 2015).
75. Tsvetkov, C., Malhotra, G., Evans, B. D. & Bowers, J. S. The role of capacity constraints in convolutional neural networks for learning random versus natural data. *Neural Netw.* **161**, 515–524 (2023).
76. Malhotra, G., Leslie, D. S., Ludwig, C. J. H. & Bogacz, R. Overcoming indecision by changing the decision boundary. *J. Exp. Psychol. Gen.* **146**, 776–805 (2017).
77. Drugowitsch, J., Moreno-Bote, R. N., Churchland, A. K., Shadlen, M. N. & Pouget, A. The cost of accumulating evidence in perceptual decision making. *J. Neurosci.* **32**, 3612–3628 (2012).
78. Rahnev, D. & Denison, R. N. Suboptimality in perceptual decision making. *Behav. Brain Sci.* **41**, e223 (2018).
79. Evans, N. J., Bennett, A. J. & Brown, S. D. Optimal or not; depends on the task. *Psychon. Bull. Rev.* **26**, 1027–1034 (2019).
80. Brainard, D. H. The Psychophysics Toolbox. *Spat. Vis.* **10**, 433–436 (1997).
81. Chen, Y. C. A tutorial on kernel density estimation and recent advances. *Biostat. Epidemiol.* https://doi.org/10.1080/24709360.2017.1396742 (2017).

82. Waskom, M. L. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
83. Jospin, L. V., Buntine, W., Boussaid, F., Laga, H. & Bennamoun, M. Hands-on Bayesian neural networks—a tutorial for deep learning users. *IEEE Comput. Intell. Mag.* **17**, 29–48 (2020).
84. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* (eds Wallach, H. et al.) (Curran Associates, 2019); https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html
85. Bingham, E. et al. Pyro: deep universal probabilistic programming. *J. Mach. Learn. Res.* **20**, 1–6 (2019).
86. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. In *Proc. 2nd International Conference on Learning Representations* (ICLR, 2013). Preprint at https://arxiv.org/abs/1312.6114 (2022).
87. Kingma, D. P. & Ba, J. L. Adam: a method for stochastic optimization. In *Proc. 3rd International Conference on Learning Representations* (ICLR, 2014). Preprint at https://arxiv.org/abs/1412.6980 (2017).
88. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2010); https://doi.org/10.1109/CVPR.2009.5206848
89. Kumbhar, O., Sizikova, E., Majaj, N. & Pelli, D. G. Anytime prediction as a model of human reaction time. Preprint at https://arxiv.org/abs/2011.12859 (2020).

## Acknowledgements

This work was supported by the National Institutes of Health (award no. R01MH119189) and the Office of Naval Research (award no. N00014-20-1-2622), both awarded to D.R. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank S. Varma and P. Verhaeghen for helpful suggestions about the analyses, as well as A. Shin and H. S. Pandi for assistance with data collection.

## Author contributions

F.R. and M.S. performed the research and analysed the data. F.R. collected the data and wrote the first draft of the paper. M.S. and D.R. edited the paper. All authors designed the research.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41562-024-01914-8.

**Correspondence and requests for materials** should be addressed to Farshad Rafiei.

**Peer review information** *Nature Human Behaviour* thanks Sushrut Thorat and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Farshad Rafiei

Last updated by author(s): May 3, 2024

# Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed |
|---|---|
| ☐ | ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☒ | ☐ A description of all covariates tested |
| ☒ | ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | The experiment was designed in MATLAB_2020b environment using Psychtoolbox 3. This study's sample size, experiment design, variables, hypothesis, and planned analyses were pre-registered on Open Science Framework (https://osf.io/kmraq) prior to any data being collected. |
| Data analysis | All data analyses were done in Python (version 3.10.11) using Google Colab (version 2.0). All the behavioral data, as well as the models and codes to generate the simulations are all available at https://osf.io/akwty. |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data collected are available at https://osf.io/akwty

## Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

| | |
|---|---|
| Reporting on sex and gender | Sixty-four subjects (31 females, age=18-32) |
| Reporting on race, ethnicity, or other socially relevant groupings | N/A |
| Population characteristics | Data from 64 healthy participants was collected. Participants were undergraduate students of Georgia Tech, with 31 female subjects and the age range was 18-32 years. |
| Recruitment | Participants were recruited through SONA Experiment Management System of School of Psychology at Georgia Tech. Participants voluntarily signed up for the study and were not individually selected by the experimenters. They were only excluded from participation if they did not have normal or corrected-to-normal vision. |
| Ethics oversight | The protocol was approved by the Georgia Institute of Technology Institutional Review Board. |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Study description | In this study, 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. The study is a quantitative experimental study. |
| Research sample | Sixty-four subjects (31 female, age=18-32) with normal or corrected to normal vision were recruited. We used convenience sampling and our final sample is likely to be representative of the college student population but not of the population as a whole. |
| Sampling strategy | We used convenience sampling because the goal of our study was to compare the human subjects behavioral data to model generated data. We didn't perform sample size calculation. We aimed to include a large number of participants with large number of trials per participant. For behavioral studies, we hypothesized that a sample size of 60 can allow us to reliably measure behavioral effects. This allowed us to have a good representation of data both in group level and individual level. |
| Data collection | 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. Two undergraduate research assistants helped the data collection but the participants completed the studies alone in the testing room. The study hypothesis was not blinded to the researcher and all participants were exposed to all experimental conditions. |
| Timing | The data collection started on February 1st, 2022 and completed on March 7, 2022. |
| Data exclusions | 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. |
| Non-participation | No participant dropped out. |
| Randomization | We only have one experimental group in our experiment. The order of stimulus presentation for each subject was pseudo-randomized. |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |
| ☒ | ☐ Plants |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

## Plants

Seed stocks

*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*

Novel plant genotypes

*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*

Authentication

*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*