EXPLAINING FEATURES OF SIMPLE HUMAN DECISIONS
USING BAYESIAN NEURAL NETWORKS

A Dissertation
Presented to
The Academic Faculty

by

Farshad Rafiei

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Psychology, College of Sciences

Georgia Institute of Technology
August 2022

COPYRIGHT © 2022 BY FARSHAD RAFIEI
EXPLAINING FEATURES OF SIMPLE HUMAN DECISIONS USING BAYESIAN NEURAL NETWORKS

Approved by:

Dr. Dobromir Rahnev, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Daniel Spieler
School of Psychology
Georgia Institute of Technology

Dr. Mark Wheeler
School of Psychology
Georgia Institute of Technology

Dr. Sashank Varma
School of Interactive Computing
Georgia Institute of Technology

Dr. Paul Verhaeghen
School of Psychology
Georgia Institute of Technology

Date Approved: July 14, 2022
To my beloved grandparents
ACKNOWLEDGEMENTS

I am obliged to several people without whom this work would have not been possible. First, and foremost, my advisor, Doby, who gave me this chance to pursue my PhD under his guidance. I will always be grateful to him for teaching me how to be an independent researcher, a critical thinker, and even a better person. I cannot forget all his kindness, patience and support and I am extremely grateful to have had a chapter of my life under his supervision.

Secondly, but just as importantly, I must acknowledge my family – my parents and my brother. They always supported me in every single step of my life and were beside me whenever I needed them. Making important decisions was the easiest by seeing their faith in me and all the emotional supports from their side. I cannot express in words that how thankful I am because of all the sacrifices they made for me.

Next, I want to thank all my friends who were always like a big family for me and each touched a chapter of my life to make me who I am. Sara Honarvar, Mir Salim Ghoddousifar, Armin Rasoulipour, Amin Adineh, Forough Irannezhad, Shervin Parvini Ahmadi, Rasa Rikhtegaran, Amin Ashraf Gandomi, Armin Tahmasbi, Sasan Jalili, Amir Seyed Salehi, Farnoush Saeedi Nejad, Ehsan Sadeghian, our friendship is timeless and the memories we have created together are priceless for me. Although we are spread out all over the world and cannot see each other as often as before, it does not change the fact you will always have special places in my heart.
I want to also thank my wonderful lab mates, especially Jiwon, Medha, Nadia, Kai, Sam and Johan, who have also become my close friends over the years. Thanks for the great times we shared and memorable moments we had during the last years. I also want to thank the RAs, Sienna Tetali, Ana Shin and Sahithi Pandi, who helped me accelerate my research by devoting their time at data collection.

Finally, I am indebted to the school of psychology and its faculty for having allowed me to be part of their graduate program and for the training I received here. I also want to express my gratitude to my committee members or giving me access to their invaluable time and expertise, through their feedback and suggestions.
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS  iv

LIST OF TABLES  viii

LIST OF FIGURES  ix

LIST OF SYMBOLS AND ABBREVIATIONS  xv

SUMMARY  xvi

CHAPTER 1. Introduction  1

CHAPTER 2. Methods  10

2.1 Behavioral Experiment  10

2.1.1 Pre-registration  10

2.1.2 Subjects  10

2.1.3 Stimulus, task, and procedure  10

2.1.4 Apparatus  14

2.2 Behavioral analyses  14

2.3 Model specifications  16

2.3.1 RTNet  16

2.3.2 Anytime Prediction  18

2.4 Deep neural network implementation  19

2.4.1 Network architecture  19

2.4.2 Network training  20

2.4.3 Model fitting  22

2.5 Model data analyses  23

CHAPTER 3. Results  25

3.1 Signatures of human decisions  25

3.1.1 Variability in human responses  27

3.1.2 Speed-accuracy trade-off  30

3.1.3 RT is shorter for easy stimuli than difficult stimuli  31

3.1.4 Skewness of RT distributions  32

3.1.5 Error trials RT is slower than correct trials RT  34

3.1.6 Confidence is higher for correct decisions than error decisions  35

3.2 Similarities between model and human data  36
3.2.1  Correlation between the model and average human data  36
3.2.2  Correlation between the model and individual subjects  37

3.3  **Similarities between model and human in each experimental condition**  43
3.3.1  Correlation between the model and average human data  43
3.3.2  Correlation between the model and individual subjects  45

**CHAPTER 4. Discussion**  51
4.1  Summary of aim and results  51
4.2  How RTNet is similar to/different from race models of decision-making?  53
4.3  **Biological plausibility of RTNet vs. Anytime Prediction**  58
4.3.1  How decision-making in brain is different from the Anytime Prediction model?  59
4.3.2  Advantages of using RTNet-like structure in modelling human decision-making  61
4.4  Limitations of using RTNet-like structure  62

**CHAPTER 5. Conclusion**  64

**REFERENCES**  66
Table 1  **Alexnet structure.** The backbone architecture of deep neural network used for all of the trained instantiations. For RTNet instantiations, we trained the network using Bayesian inference and hence the network learnt distributions over weights instead of point estimates. For Anytime Prediction instantiations, additional early exits were applied immediately after each convolutional layer of processing. ReLU: Rectified Linear Unit.

Table 2  **Correlation between the average model performance and average human data in each condition.** The RTNet model generally performed better than the Anytime Prediction model for RT and decision accuracy measures. Whereas Anytime Prediction was more correlated with average human data for confidence measure in different conditions. All correlation values are significantly greater than zero.
LIST OF FIGURES

Figure 1  Model architectures. (A) Feedforward neural network (FNN). A standard feedforward network which is a combination of linear non-linear transformations that can explain a function. A standard FNN consists of an input layer, hidden layers, and an output layer. The weights and biases in the network are point estimates in a standard FNN. The softmax function at output layer converts the scores into probability distributions, and the choice option with highest probability assigned, becomes the model’s decision. This highest probability is often used as FNN’s decision confidence. (B) RTNet. The model contains two modules. The neural network module and evidence accumulation module. Each decision process begins from a starting point and sample a random weight and bias for each edge to process the input image. These steps result in different evidence in favor of each alternative which is accumulated over time in evidence accumulation module. The decision process stops when a predefined threshold level is hit for one of the choice options. Number of samples required to hit the decision boundary is considered as model RT and the confidence is computed by subtracting the second highest accumulated evidence from the highest accumulated evidence. (C) Anytime Prediction. The network is similar to a FNN in that it uses point estimates as weights and biases, but different in that it has multiple output layers. In fact, Anytime Prediction consists of early exits after specific layers of processing, and it stops the decision process if the confidence at an output layer exceeds a pre-defined threshold (Kumbhar et al., 2020). Similar to FNN, the decision confidence is equal to the highest probability computed via softmax function at the exit layer.

Figure 2  Experiment task. Each trial starts with a fixation cross that takes 500-1000 ms. Followed by that, a digit between 1-8 with a noisy background appears on screen and remains for 300 ms. The stimulus is either easy or difficult, depending on the
level of background noise. After 300 ms, the stimulus disappears, and subjects are asked to report their perceived digit and confidence in their decision. Both the decision and the confidence reports are untimed. Each block contains 60 of these trials: 30 easy images and 30 difficult images pseudo-randomly incorporated in a block. At the beginning of each block, subjects were asked to adjust their responding policy based on the instructions given to them. They were asked to either focus on accuracy of their responses or speed of their decisions throughout the block. Each subject underwent four runs each included four of these blocks. In each run, subjects were instructed exactly two times to respond to stimuli by focusing on accuracy and to respond exactly two times by focusing on speed, in a randomized manner. The images for each condition (e.g. accuracy focus with easy images) are unique and no image appears in more than one condition. Each experiment contains 960 trials in total.

Figure 3  Variability in human, RTNet and Anytime Prediction responses for same images in different conditions. We measured the consistency of responses in different experimental conditions for (A) humans, (B) RTNet and (C) Anytime Prediction. In each condition, we compared the first and second responses to the same stimulus and computed the proportion of 1) consistent responses that both were correct (consistent-two-correct), 2) inconsistent responses where only one of them was correct (inconsistent-one-correct), 3) consistent responses where both of them were incorrect (consistent-zero-correct), 4) inconsistent responses where both of them were incorrect. Humans and RTNet provided inconsistent responses for same stimuli whereas Anytime Prediction always produced consistent responses due to its deterministic nature.

Figure 4  Speed-accuracy trade-off. Accuracy of responses decreases when decision speed is emphasized. Two-way repeated measures ANOVA with factors of task difficulty and speed pressure showed main effect for both factors among (A) humans. (B)
RTNet and (C) Anytime Prediction. Error bars show standard error of the mean (SEM).

**Figure 5** Shorter RT for easy stimuli than difficulty stimuli. (A) In human data, difficult trials are associated with longer RTs. Both (B) RTNet and (C) Anytime Prediction showed the same effect. Error bars show standard error of the mean (SEM).

**Figure 6** Skewness of RT distributions for a representative subject/model. RT distributions for a representative subject/model change as a function of experimental condition. For (A) humans and (B) RTNet, RT distributions in accuracy focus condition is more skewed to the right than speed focus condition. Within each speed pressure condition, RT distributions associated to difficult condition is more skewed to right than easy condition. (C) Anytime Prediction does not provide patterns observed in human RT data.

**Figure 7** Skewness of RT distributions across subjects/models. The skewness of RT distributions change across conditions. For (A) human data, skewness of RT distributions changed by task difficulty and speed pressure. The same trend was observed for (B) RTNet, whereas (C) Anytime Prediction showed main effect of speed pressure and an interaction effect, but not main effect of task difficulty. Error bars show standard error of the mean (SEM).

**Figure 8** Error trials are slower than correct trials. For (A) humans and (B) RTNet error trials are associated with higher RT than correct trials. However, (C) Anytime Prediction showed an opposite pattern and correct trials took a longer processing time. Error bars show standard error of the mean (SEM).

**Figure 9** Higher confidence for correct trials than error trials. The confidence scores for correct trials in (A) human data was higher, on average, when compared to error...
trials. The same effect was observed both for (B) RTNet and (C) Anytime Prediction. Error bars show standard error of the mean (SEM).

**Figure 10** Correlation between average model RT and individual subjects’ RT. (A) There is a RT associated with each image provided by individual subjects or model representatives. Each point shows the correlation between an individual subject RT and model representative RT across images. Human consistency model is provided as a lower bound on noise ceiling. The average correlation for RTNet is significantly higher than that for Anytime Prediction. Error bars show standard error of the mean (SEM). (B) The RT correlation between individual subject and a model representative (i.e. RTNet or Anytime Prediction) is plotted against its corresponding correlation value in human consistency model. The regression line with positive slope indicates that subjects which are more correlated with average of other subjects, they are also more correlated with average of the examined model.

**Figure 11** Correlation between average model accuracy and individual subjects’ accuracy. (A) There is an accuracy associated to each image provided by individual subjects or model representatives. Each point shows the correlation between an individual subject accuracy and model representative accuracy across images. Human consistency model is provided as a lower bound on noise ceiling. The average correlation for RTNet is significantly higher than that for Anytime Prediction. Error bars show standard error of the mean (SEM). (B) The accuracy correlation between individual subject and a model representative (i.e. RTNet or Anytime Prediction) is plotted against its corresponding correlation value in human consistency model. The regression line with positive slope indicates that subjects which are more correlated with average of other subjects, they are also more correlated with average of the examined model.
Figure 12  **Correlation between average model confidence and individual subjects’ confidence.** (A) There is a confidence score associated to each image provided by individual subjects or model representatives. Each point shows the correlation between an individual subject confidence and model representative confidence across images. Human consistency model is provided as a lower bound on noise ceiling. The average correlation for RTNet is significantly higher than that for Anytime Prediction. *Error bars show standard error of the mean (SEM).*  (B) The confidence correlation between individual subject and a model representative (i.e. RTNet or Anytime Prediction) is plotted against its corresponding correlation value in human consistency model. The regression line with positive slope indicates that subjects which are more correlated with average of other subjects, they are also more correlated with average of the examined model.

Figure 13  **Correlation between average model performance and individual subjects within each condition.** (A) shows the correlation between an individual subject and average of other subjects (left), RTNet (middle), Anytime Prediction (right) for RT measure. All average values are significantly greater than zero. RTNet predicts individual RT data significantly better than Anytime Prediction in all experimental conditions. Also, RTNet preserves the patterns observed in human data, but Anytime Prediction only preserves the patterns across task difficulty conditions and fails to preserve the patterns across speed pressure conditions. (B) shows the correlation between an individual subject and average of other subjects (left), RTNet (middle), Anytime Prediction (right) for accuracy measure. All average values are significantly greater than zero. RTNet predicts individual accuracy data significantly better than Anytime Prediction in all experimental conditions except accuracy focus with easy images where both models perform equally well. Also, RTNet showed correlation patterns similar to human data across conditions but Anytime Prediction only succeeded to preserve patterns across speed pressure conditions.
conditions but not across task difficulty. (C) shows the correlation between an individual subject and average of other subjects (left), RTNet (middle), Anytime Prediction (right) for confidence measure. All average values are significantly greater than zero. Anytime Prediction predicts individual confidence data significantly better than RTNet in accuracy focus condition with easy images, but in other conditions the models perform equally well. both models preserved the correlation patterns across speed pressure conditions but failed to preserve it across task difficulty conditions. Error bars show standard error of mean (SEM).
LIST OF SYMBOLS AND ABBREVIATIONS

RT   Reaction time

FNN  Feedforward neural networks

RNN  Recurrent neural network

RDM  Random diffusion model

LBA  Linear ballistic accumulation

KDE  Kernel density estimation
SUMMARY

Feedforward neural networks exhibit excellent object recognition performance and currently provide the best models of biological vision. However, despite their remarkable performance in recognizing unseen images, their decision behavior differs markedly from human decision-making. Standard feedforward neural networks perform an identical number of computations to process a given stimulus and always land on the same response for that stimulus. Human decisions, in contrast, take variable amount of time and are stochastic (i.e., the same stimulus elicits different reaction time, RT, and sometimes different responses on different trials). Here we develop a new neural network, RTNet, that closely approximates all basic features of perceptual decision making. RTNet has noisy weights and processes the same stimulus multiple times until the accumulated evidence reaches a threshold, thus producing both variable RT and stochastic decisions. In addition, RTNet exhibits several features of human perceptual decision-making including speed-accuracy tradeoff, right-skewed RT distributions, lower accuracy and confidence for harder decisions, etc. Finally, data from 60 human subjects on a digit discrimination task demonstrates that RT, accuracy, and confidence produced by RTNet for individual novel images correlate with the same quantities produced by human subjects. Overall, RTNet is the first neural network that exhibits all basic signatures of perceptual decision making.
CHAPTER 1. INTRODUCTION

Feedforward neural networks (FNNs) are the dominant models of visual object recognition in the human brain (Hasson, Nastase, & Goldstein, 2020; Kietzmann, McClure, & Kriegeskorte, 2019; Kriegeskorte, 2015; Lindsay, 2021; Yamins & DiCarlo, 2016). Among these models deep neural networks perform better in computer vision tasks and also provide a more accurate predictions of neural and behavioral data than their counterparts (Güçlü & Gerven, 2015; Kell, Yamins, Shook, Norman-Haignere, & McDermott, 2018; Khaligh-Razavi & Kriegeskorte, 2014; Rajalingham, Schmidt, & DiCarlo, 2015; Yamins et al., 2014; Zhuang et al., 2021). These kinds of neural networks, utilize a deep hierarchy of linear-nonlinear filters with local receptive fields which is inspired by the architecture of the human brain, and they are shown to perform as well as humans in real life tasks (Voulodimos, Doulamis, Doulamis, & Protopapadakis, 2018) such as object detection (Gupta, Seal, Prasad, & Khanna, 2020; L. Liu et al., 2020; Zou, Shi, Guo, & Ye, 2019), face recognition (Lal et al., 2018; Singh & Prasad, 2018), motion detection (Mathis, Schneider, Lauer, & Mathis, 2020; Yang & Ismail, 2022), action recognition (Kamel et al., 2019; L. Wang, Huynh, & Koniusz, 2020), etc.

However, despite achieving human-like performance, their behavior in making decisions is different from a human agent. Human decisions are highly variable even for the same stimulus (Beck, Ma, Pitkow, Latham, & Pouget, 2012; Renart & Machens, 2014), while FNNs are deterministic models without any response variability for the same images. An FNN takes identical number of computations (i.e., time) to process a stimulus and it always lands on the same response when that stimulus is shown. Human decisions,
however, change as a function of instructions and task difficulty (Renart & Machens, 2014). Under different conditions, reaction time (RT) and accuracy of responses can vary, even for the same stimulus (Rafiei & Rahnev, 2021). For example, the relationship between reaction times and accuracy of responses is not fixed and it varies according to whether speed or accuracy is emphasized (Ratcliff & Rouder, 1998). This is called speed-accuracy trade-off (Heitz, 2014; Heitz & Schall, 2012) which is the humans’ ability of trading speed for accuracy when making decisions. On contrary, FNNs cannot account for speed-accuracy trade-off phenomenon since they are not inherently sensitive to instructions or task difficulty (Spoerer, McClure, & Kriegeskorte, 2017). In fact, FNNs used in computer vision do not have meaningful dynamics since they instantaneously transform the input into an output regardless of the task difficulty or response urgency (van Bergen & Kriegeskorte, 2020).

Recently several studies tried to explain speed-accuracy trade-off effect by making some structural changes to FNN architecture (Figure 1A). In the first group of studies, the goal is to adapt network structure or parameters to different inputs, leading to advantages in terms of accuracy and computational efficiency. These are called dynamic neural networks (Han et al., 2021). The idea behind these networks is that, when there is no need for high decision accuracy or when fast responses are desirable, then there is no need to go through all the computations in the networks’ structure. A variable subset of convolutional filters (Z. Chen, Li, Bengio, & Si, 2019; Gao, Zhao, Dudziak, Mullins, & Xu, 2018; Hua, Zhou, De Sa, Zhang, & Suh, 2018; C. Li et al., 2021) or a variable number of layers (Bolukbasi, Saligrama, Wang, & Dekel, 2017; Huang et al., 2017; Veit & Belongie, 2018; X. Wang, Yu, Dou, Darrell, & Gonzalez, 2018) can be used depending on the urgency of
decision-making. Recently, a study compared the RT vs. accuracy relationship in humans against a model from dynamic neural network classes of models and found that both humans and “Anytime Prediction” (Figure 1C) provide patterns that trades accuracy for speed (Kumbhar, Sizikova, Majaj, & Pelli, 2020). An anytime algorithm assumes that the decision process can occur at intermediate layers of the architecture in expense of decision accuracy (Iuzzolino, Mozer, & Bengio, 2021; Kumbhar et al., 2020; Z. Liu, Xu, Wang, Darrell, & Shelhamer, 2022). As more computation time is given to anytime algorithm, the results gradually improve. Depending on user requirements, the decision process can be interrupted at any point during its computation. Therefore, it can account for speed-accuracy trade-off effect observed in human behavior while making decisions.
Figure 1. Model architectures. (A) Feedforward neural network (FNN). A standard feedforward network which is a combination of linear non-linear transformations that can explain a function. A standard FNN consists of an input layer, hidden layers, and an output
layer. The weights and biases in the network are point estimates in a standard FNN. The softmax function at output layer converts the scores into probability distributions, and the choice option with highest probability assigned, becomes the model’s decision. This highest probability is often used as FNN’s decision confidence. (B) RTNet. The model contains two modules. The neural network module and evidence accumulation module. Each decision process begins from a starting point and sample a random weight and bias for each edge to process the input image. These steps result in different evidence in favor of each alternative which is accumulated over time in evidence accumulation module. The decision process stops when a predefined threshold level is hit for one of the choice options. Number of samples required to hit the decision boundary is considered as model RT and the confidence is computed by subtracting the second highest accumulated evidence from the highest accumulated evidence. (C) Anytime Prediction. The network is similar to a FNN in that it uses point estimates as weights and biases, but different in that it has multiple output layers. In fact, Anytime Prediction consists of early exits after specific layers of processing, and it stops the decision process if the confidence at an output layer exceeds a pre-defined threshold (Kumbhar et al., 2020). Similar to FNN, the decision confidence is equal to the highest probability computed via softmax function at the exit layer.

A second group of studies are the ones which changed the structure of an FNN by adding recurrent connections to its structure and showed that these models better explain neural activity and human behavior including speed-accuracy trade-off. Recurrent neural networks (RNNs) are shown to be better models of object recognition system in human brain under challenging conditions (Spoerer, Kietzmann, Mehrer, Charest, & Kriegeskorte, 2020; Spoerer et al., 2017), and in explaining temporal dynamics of the brain (Güçlü & van Gerven, 2017; Kietzmann, Spoerer, et al., 2019; Nayebi et al., 2018). Specifically, Spoerer and his colleagues compared RNNs behavior with FNNs and found that RNNs can explain flexible trading of speed and accuracy better than their feedforward counterparts (Spoerer et al., 2020). They observed that recurrent computation was brief for easy images, whereas for harder images, recurrent computation could proceed longer. They also found that the human RTs correlate better with RTs obtained from RNNs than feedforward ones. However, despite their advantages over FNNs, it is not currently feasible to implement and
train an appropriate RNN model to deal with static images (Spoerer et al., 2020, 2017; van Bergen & Kriegeskorte, 2020). Training an RNN for static images, requires unrolling the feedback or lateral connections for a limited number of times which creates a FNN that is only an approximation of the true RNN (Liao & Poggio, 2016; van Bergen & Kriegeskorte, 2020).

Despite structural differences between dynamic neural networks (e.g., the Anytime Prediction model) and unrolled RNNs, they both share a common rule for stopping their processing and making decisions. Both models work based on thresholded confidence system. That is, when the network has enough evidence in favor of one alternative, it stops the processing and makes the decision. The only difference is that Anytime Prediction benefits from checking the evidence after each layer of processing, whereas unrolled RNNs update the available evidence at each time step but after the whole network processes the input information. In fact, dynamic neural networks have multiple classifiers each specialized for a layer of processing, while unrolled RNNs has only one classifier that make a prediction at the end of each time point. Biologically, RNNs are better models of human vision (Spoerer et al., 2017) and it is beneficial to have recurrent connections in network architecture to achieve greater computational depth using a limited compressed architecture (van Bergen & Kriegeskorte, 2020).

Similar to unrolled RNNs, we propose a model here which only uses one classifier in its structure, but it is not restricted to limited computational depth or limited number of time steps. Our model, RTNet (Figure 1B), accumulates information for each alternative separately in a fashion that a race model of decision-making does (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Heathcote & Matzke, 2022; Vickers, 2007). In a race
model, separate evidence measurement systems are considered for each alternative and noisy evidence accumulates for each option over time until a predefined threshold is achieved. Similarly, our model is a neural network in which each choice option has its own evidence accumulation system. The structure of the neural network is relatively different in that instead of point estimates, the network learns distributions over its edges. In fact, the model consists of a Bayesian neural network (Goan & Fookes, 2020; MacKay, 1995; Mullachery, Khera, & Husain, 2018) with accumulation system at its output layer. This enables the neural network to sample evidence in favor of alternatives and stop the decision process only when it has enough evidence in favor of an alternative.

To assess the model’s ability to make decisions similar to humans, we tested it to see whether it produces ubiquitous benchmarks in decision-making literature. Most of the understanding regarding dynamics of decision making come from simple two-choice tasks performed in a laboratory environment. In addition to speed-accuracy trade-off, several other ubiquitous patterns were reported in literature that any model of human decision-making should try to account for. For example, mean RT is shorter for easy stimuli (Forstmann, Ratcliff, & Wagenmakers, 2016; Ratcliff & Rouder, 1998; Wagenmakers & Brown, 2007), increasing speed stress shortens the mean RT but it also increases the proportion of erroneous responses (Forstmann et al., 2016; Heitz, 2014; Heitz & Schall, 2012), error trials correspond to higher mean RTs (Brown & Heathcote, 2008; Luce, 1986; Ratcliff, 2002; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008), RT distributions are right-skewed and the skewness changes as a function of task difficulty (Forstmann et al., 2016), confidence ratings for correct trials are higher than that for error trials (Rahnev, 2021), etc. We first verified these quality metrics in a dataset we collected where human
subjects were engaged in an eight-choice digit discrimination task. We then tested the model under same conditions to explore its capability in producing a behavior similar to human agents.

Further, we explored the model’s capability in predicting RT, accuracy and confidence scores for novel images under a specific condition. For this aim, we embedded unique images in each condition of our 2 x 2 design experiment and found the correlation between human and model produced responses for each image. This enables us to examine the model’s capability in predicting choice property distributions (i.e., RT, accuracy or confidence). Throughout the paper, we compared our model to Anytime Prediction, since it was shown to produce a better account of human data when compared to other existing models which can explain speed-accuracy trade-off including unrolled RNNs (Subramanian et al., 2022). It should be noted that the final structure of an unrolled RNN depends on many factors such as the position of feedback connections, position of lateral connections, number of enrollments, etc. and there is no systematic way of finding an optimal structure which matches an arbitrary architecture (van Bergen & Kriegeskorte, 2020). Therefore, we only considered Anytime Prediction which was shown to be the best available neural network model that resembles human behavior and we matched the basic architecture with RTNet to provide a fair comparison between them.

In the following, we will go through the methods of our study in more details, provide the results, and discuss them from different perspectives. Specifically, in methods section (CHAPTER 2), we first discuss the experimental design 2.1), and provide details regarding data collection, pre-processing, and behavioral data analyses 2.2). Rest of this section is devoted to technical details about model architecture (2.3), and neural network
implementation (2.4) such as model training, and model fitting procedure. In results section (CHAPTER 3), we first visualize the human produced data and examine whether RTNet or Anytime Prediction create the observed patterns in human data (3.1). We then compare the models’ performance in predicting RT, accuracy, and confidence of individual images provide by human subjects, both between experimental conditions (3.2) and within them (3.3). Finally, in discussion section (CHAPTER 4), we summarize our findings (4.1), and extensively discuss the similarities between RTNet and existing models of decision-making (4.2), biological plausibility of RTNet against Anytime Prediction (4.3) and limitations of our study (4.4). The concluding remarks are provided in the last chapter of this dissertation (CHAPTER 5).
CHAPTER 2. METHODS

2.1 Behavioral Experiment

2.1.1 Pre-registration

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

2.1.2 Subjects

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

2.1.3 Stimulus, task, and procedure

The task stimulus was selected from publicly available handwritten digits (MNIST) dataset (L. Deng, 2012). This dataset contains 60,000 training images and 10,000 testing images. Since the training images were used to train the models in this study, we randomly selected images from MNIST test set to include in our experiment. This ensures that the
selected images for the experiment are novel both for the human subjects and for the trained models. We randomly selected 480 images for the experiment (120 for each condition).

We incorporated two different levels of difficulty in experiment (i.e., “difficult” or “easy”) by adding uniform noise to the selected images (Figure 2). The uniform noise with average of 0.25 (range = 0-.5) or 0.4 (range = 0-.8) was added to the images to generate easy or difficult stimuli, respectively. These noise levels were chosen based on the pilot testing to produce two different performance levels. The MNIST dataset images are of size 28 x 28 pixels which was very small for the screens we were using for our experiment. Therefore, before adding noise, the selected images were first resized to 84 x 84 pixels (using MATLAB’s `imresize` function), and they were padded with the background color of MNIST images to size 256 x 256 pixels (Figure 2).

Subjects performed a digit discrimination task where they reported their perceived digit followed by rating their decision confidence. Each trial began with subjects fixating on a small white cross for 500-1000 ms, followed by presentation of the stimulus for 300 ms (Figure 2). The stimulus was a digit between 1-8 (0 and 9 were excluded) superimposed on a noisy background. Subjects’ task was to report their perceived digit using a computer keyboard followed by reporting their decision confidence on a 4-point scale (where 1 corresponds to the lowest confidence rating and 4 corresponds to the highest confidence rating). There was no deadline on response or confidence rating. In each trial, the stimulus was randomly selected from pool of 120 images chosen for each condition.
Figure 2. **Experiment task.** Each trial starts with a fixation cross that takes 500-1000 ms. Followed by that, a digit between 1-8 with a noisy background appears on screen and remains for 300 ms. The stimulus is either easy or difficult, depending on the level of background noise. After 300 ms, the stimulus disappears, and subjects are asked to report their perceived digit and confidence in their decision. Both the decision and the confidence reports are untimed. Each block contains 60 of these trials: 30 easy images and 30 difficult images pseudo-randomly incorporated in a block. At the beginning of each block, subjects were asked to adjust their responding policy based on the instructions given to them. They were asked to either focus on accuracy of their responses or speed of their decisions throughout the block. Each subject underwent four runs each included four of these blocks. In each run, subjects were instructed exactly two times to respond to stimuli by focusing on accuracy and to respond exactly two times by focusing on speed, in a randomized manner. The images for each condition (e.g. accuracy focus with easy images) are unique and no image appears in more than one condition. Each experiment contains 960 trials in total.

Subjects came for one session and underwent a training session before completing the main experiment. Before the start of the session, subjects were given detailed instructions about different conditions. The experiment included two different speed-accuracy tradeoff conditions. Therefore, it was specifically emphasized to focus on accuracy of responses in “accuracy focus” condition and to provide fast responses with less
focus on accuracy in “speed focus” condition. Then subjects were moved to a dark room and were asked to place their fingers on numbers 1-8 on a standard keyboard by placing four fingers of left hand on numbers 1-4 and placing four fingers of right hand on numbers 5-8. This helped in providing faster motor responses without looking at the keyboard before each button press, especially in “speed focus” condition.

The session started with three blocks of training each containing 50 trials. The first block contained images from MNIST dataset without any noise. This was done to familiarize the subjects with the task and button press. Followed by that, two blocks were performed by subjects, where in first they were asked to focus on accuracy of their responses and in next one we demanded them to focus on speed. The difficulty level of stimuli in these two training blocks was same as the difficulty levels incorporated in the main experiment. During the whole training session, the experimenter was standing beside the subject quietly and was answering the questions to clarify potential ambiguities. None of the images used in training session, were used in the main experiment.

Once the subject confirmed that he/she understood the task, the experimenter left the room and subjects went through 960 trials in four runs each containing four blocks of 60 trials. Each block consisted of a single speed-accuracy trade-off condition, and each run included exactly two “accuracy focus” and two “speed focus” conditions in a randomized order. At the beginning of each block, subjects were given the name of the condition for that block (“accuracy focus” or “speed focus”) to adjust their responding policy according to the name of the condition. In each block, we pseudo-randomly interleaved trials with two different difficulty levels such that each was presented exactly 30 times. All 480 images were shown to subjects in first two runs and the procedure was repeated with
completely different order of blocks and trials in second two runs. All noisy images were same for all subjects, and each image was assigned only to one specific condition (i.e. one image does not appear in two different conditions).

2.1.4 Apparatus

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

2.2 Behavioral analyses

We followed the data analyses steps mentioned in pre-registration (https://osf.io/kmraq). We first removed the subjects who did not follow experimental instructions by not providing faster RT, on average, for “speed focus” condition compared to “accuracy focus” condition. This resulted in removing two subjects (out of 64 subjects). We also removed the subjects that showed ceiling or floor effects in average accuracy or average confidence scores throughout the experiment. Specifically, we set our criterion to keep subjects that provided average accuracy of greater than 15% and less than 95% across all trials. None of the subjects were removed due to this cleaning process. However, two subjects were removed from further analyses because they showed ceiling effects for confidence scores where we set the removal criterion to subjects who provided average confidence scores of less than 1.15 or greater than 3.85 across all trials. In pre-registration document, we set the exclusion criterion for confidence ratings as average scores less than 1.3 or more than 3.7. These values were chosen according to the previous studies performed
in our research group where subjects performed a two-choice task with artificial stimuli (e.g. Bang, Shekhar, & Rahnev, 2019; Shekhar & Rahnev, 2018). However, since a huge number of subjects were removed due to providing average confidence ratings greater than 3.7, we slightly changed this criterion and set the new value to 3.85. Therefore, we ended up having sixty subjects after applying these two steps.

To remove the outliers for each subject in each condition, as declared in pre-registration document, we used Tukey’s interquartile criterion. For each subject, we computed 25th and 75th quantile of RT distributions in each condition. We then removed all RTs with values more than 1.5 times the interquartile range from the quartiles. This step resulted in removing an average of 5.46% of total trials (range 1.35–8.22% for each subject).

After preprocessing, for each subject in each condition, we computed average accuracy, average RT and skewness of RT distributions. The skewness was computed 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 then aggregated these measures across subjects to examine the patterns under different conditions. We also computed average RT and average confidence scores for error and correct trials across subjects to examine how the RT or confidence score patterns change as a function of response accuracy.

In addition, we examined how the shape of RT distributions change for individual subjects in each condition. For this reason, we provided kernel density estimation (KDE) of the observations in each condition. Similar to a histogram, KDE approximates the underlying probability density function that generated the data (Y. C. Chen, 2017). The
only difference is that, in histogram, the density function is provided by binning and counting the observations, whereas a KDE plot smooths the observations with a Gaussian kernel, producing a continuous density estimate. All the KDE plots in this study were created using Seaborn’s KDE plot with a smoothing bandwidth of 1.2 (Waskom, 2021).

2.3 Model specifications

2.3.1 RTNet

The model consists of two main modules (Figure 1B). The first module is a Bayesian neural network (BNN) which is capable of making predictions regarding an image. BNNs are a type of artificial neural network built by introducing stochastic components into the network to simulate multiple possible models with their associated probability distribution (Goan & Fookes, 2020; Jospin, Buntine, Boussaid, Laga, & Bennamoun, 2020; MacKay, 1995; Mullachery et al., 2018). The main difference between a BNN and standard feedforward neural network is that in BNN the weights are biases are distributions instead of point estimates. A random sample from these distributions result in a unique deterministic network. This random sampling enables variabilities in network decision which in turn can be fed into an accumulation process that drives a decision. That is actually the second module of our model. Here we consider a separate information accumulation process for each choice option similar to a race model (Brown & Heathcote, 2005, 2008; Heathcote & Love, 2012; Heathcote & Matzke, 2022). The decision process begins from a starting point and continues until the total amount of accumulated information reaches a predefined threshold for an alternative. The alternative for which the threshold was hit, is the response of the model. The response time for a decision is the time
(i.e. number of samples) required to reach a decision threshold. The confidence of the model is set to be the difference between highest accumulated information and the next highest option. This was selected since it was shown that confidence reports are best explained by the difference between the posterior probability of the best and next-best options, rather than by the posterior probability of the selected option alone (H. H. Li & Ma, 2020).

There are free parameters associated to each module of the model. In the first module, the free parameters are the mean and standard deviation of normal distributions over weights and biases. These parameters are learnt in the training process and their fitting process is dependent on factors such as the dataset in use, the initial condition of distributions, hyperparameters used for training (e.g. learning rate), etc (Goan & Fookes, 2020; Jospin et al., 2020). However, these free parameters are learnt end-to-end using modern learning techniques based on gradient descent and backpropagation (Rawat & Wang, 2017). Therefore, we do not adjust them manually anywhere during the analysis. In second module, the threshold parameter is the only free parameter. This is the parameter we adjust manually to numerically fit the human data after training the first module. This parameter is equivalent to the threshold parameter in race model (Heathcote & Matzke, 2022), or drift diffusion model (Forstmann et al., 2016), which accounts for speed-accuracy trade-off. Low threshold value results in faster but less accurate decisions since with only few samples the preset criterion is often met, whereas high threshold value result in longer deliberation and more accurate responses. In addition to these parameters, we adjust image noise to control for model precision. The image noise is not an inherent model parameter, but it is important to tune image noise since it was reported that the noise level for human
and neural networks are different to achieve a similar performance (Geirhos et al., 2017, 2018). Therefore, when fitting the parameters for the RTNet model, we only adjusted the threshold value and image noise level to match the desired values.

2.3.2 Anytime Prediction

The Anytime Prediction model has an architecture similar to a FNN (Figure 1A) but with early-exit classifiers to make anytime predictions (Figure 1C) (Kumbhar et al., 2020). Depending on the number of hidden layers in feedforward architecture, specific number of output layers are incorporated which are specialized to make the decision according to the available evidence using the computations up until that point of processing. At each output layer, the probability of a choice being correct is computed using a softmax function. The softmax function is often used at output layer of a neural network to perform multiclass classification. It converts the scores to normalized probability distribution and hence it takes a value between 0 and 1. For Anytime Prediction, if the softmax value for a choice option exceeds a predefined value at an output layer of processing, the network stops the decision process and produces the response. The layer at which the response was made is indicative of the decision time, and the softmax value at that layer is indicative of decision confidence (Huang et al., 2017; Kumbhar et al., 2020). Hence, the only free parameter in model architecture is the threshold value which is indicative of stopping criterion and it accounts for speed-accuracy trade-off. Therefore, similar to RTNet model, when tuning the parameters to achieve desired level of performance for Anytime Prediction, we only adjusted the threshold value and image noise level, and all other parameters were tuned end-to-end while training the network.
2.4 Deep neural network implementation

2.4.1 Network architecture

All neural networks in this study were implemented using Alexnet architecture (Krizhevsky, Sutskever, & Hinton, 2012). The Alexnet has eight layers with learnable parameters. The model consists of five convolutional layers with a combination of max pooling followed by three fully connected layers (Table 1). All neural networks in these experiments were implemented in PyTorch (Paszke et al., 2019). Bayesian networks were implemented using Pyro (Bingham et al., 2019) which is a probabilistic programming library built on PyTorch.

Table 1. Alexnet structure. The backbone architecture of deep neural network used for all of the trained instantiations. For RTNet instantiations, we trained the network using Bayesian inference and hence the network learnt distributions over weights instead of point estimates. For Anytime Prediction instantiations, additional early exits were applied immediately after each convolutional layer of processing. ReLU: Rectified Linear Unit.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Number of filters</th>
<th>Filter size</th>
<th>Stride</th>
<th>Padding</th>
<th>Size of feature map</th>
<th>Activation function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>227 x 227</td>
<td>-</td>
</tr>
<tr>
<td>Convolutional 1</td>
<td>96</td>
<td>11 x 11</td>
<td>4</td>
<td>-</td>
<td>55 x 55 x 96</td>
<td>ReLU</td>
</tr>
<tr>
<td>Max Pool 1</td>
<td>-</td>
<td>3 x 3</td>
<td>-</td>
<td>2</td>
<td>27 x 27 x 96</td>
<td>-</td>
</tr>
<tr>
<td>Convolutional 2</td>
<td>256</td>
<td>5 x 5</td>
<td>1</td>
<td>2</td>
<td>27 x 27 x 256</td>
<td>ReLU</td>
</tr>
<tr>
<td>Max Pool 2</td>
<td>-</td>
<td>3 x 3</td>
<td>2</td>
<td>-</td>
<td>13 x 13 x 256</td>
<td>-</td>
</tr>
<tr>
<td>Convolutional 3</td>
<td>384</td>
<td>3 x 3</td>
<td>1</td>
<td>1</td>
<td>13 x 13 x 384</td>
<td>ReLU</td>
</tr>
<tr>
<td>Convolutional 4</td>
<td>384</td>
<td>3 x 3</td>
<td>1</td>
<td>1</td>
<td>13 x 13 x 384</td>
<td>ReLU</td>
</tr>
<tr>
<td>Convolutional 5</td>
<td>256</td>
<td>3 x 3</td>
<td>1</td>
<td>1</td>
<td>13 x 13 x 256</td>
<td>ReLU</td>
</tr>
<tr>
<td>Max Pool 3</td>
<td>-</td>
<td>3 x 3</td>
<td>2</td>
<td>-</td>
<td>6 x 6 x 256</td>
<td>-</td>
</tr>
<tr>
<td>Fully Connected 1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4096</td>
<td>ReLU</td>
</tr>
<tr>
<td>Dropout 1</td>
<td>rate = 0.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4096</td>
<td>-</td>
</tr>
<tr>
<td>Fully Connected 2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4096</td>
<td>ReLU</td>
</tr>
<tr>
<td>Dropout 2</td>
<td>rate = 0.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4096</td>
<td>-</td>
</tr>
<tr>
<td>Fully Connected 3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>10</td>
<td>Softmax</td>
</tr>
</tbody>
</table>
In addition to the standard Alexnet structure (Table 1), we incorporated additional readout layers in the Anytime Prediction models. These readout layers were located right after each layer of processing (Figure 1C). Specifically, for first and second layer of processing, they were considered immediately after Max pooling layers, whereas for third and fourth layers, they were located right after convolutional layers. The feature map size of all these readout layers were set to the number of classes.

2.4.2 Network training

Before passing the images to the network, a number of pre-processing steps were taken. The standard input size to Alexnet model architecture is 227 x 227 pixels. Therefore, we first resized MNIST images to this size and normalized them with mean of 0.1307 and standard deviation of 0.3081. This normalization procedure is a standard procedure when using Alexnet for classification of ImageNet dataset (J. Deng et al., 2010).

All the models were trained such that the classification accuracy becomes more than 97% on MNIST test set. To achieve this, we trained Bayesian networks for a total of 15 epochs with a batch size of 500. Evidence lower bound (ELBO) loss function was used for training these networks (Kingma & Welling, 2013). Due to its deterministic nature, for Anytime Prediction, only three epochs were enough to achieve test accuracy of more than 97% with the same batch size and a weighted cumulative loss function (Kumbhar et al., 2020). For all networks, Adam (Kingma & Ba, 2014) was used for optimization with a learning rate of 0.001. To ensure that each network performs greater than 97% on MNIST test set, we followed a specific rule for each model. When testing an image with Bayesian network, we sampled 10 times from the posterior distributions learnt during the training,
and in fact, created 10 different deterministic networks. Each of these networks resulted in a unique response for each image. The response with highest frequency among 10 responses was chosen as the final decision of the Bayesian network. With this procedure, we tested all images in MNIST test set and ensured that the overall accuracy of the trained Bayesian network is greater than 97%. For Anytime Prediction networks, we considered the response of the last output layer as network’s decision. Therefore, the accuracy using this criterion should exceed 97% for MNIST test images to consider a trained Anytime Prediction as acceptable network. When a network did not result in accuracy greater than 97%, we started the training over with same initial values with same hyperparameters. This way, we ensured that all the trained networks are accurate in at least 97% of predictions for MNIST test set, when enough time is granted to them.

We trained sixty instantiations using above procedure but with different initialized values for weights and biases because different initial conditions prior to training lead to substantial differences among models of deep learning (Hermann & Lampinen, 2020; Mehrer, Spoerer, Kriegeskorte, & Kietzmann, 2020). For RTNet, we trained sixty networks by setting the initial values of weights and biases to sixty different values. Since the weights and biases in these networks are normal distributions, we manipulated the initial values by manipulating the mean and standard deviation of those normal distributions. Specifically, the mean of weights and biases was set to a value between 0 and 1.2 with 0.1 increments, and standard deviation of weights and biases was set to a value ranging from 1 to 5 with increments of 1. To train sixty different instantiations of Anytime Prediction, we used the same normal distributions which was used as initial distributions for Bayesian networks. For each network, we sampled from those distributions once, and set those values as the
initial values for training. This makes the initial conditions for RTNet and Anytime Prediction instantiations look similar to each other. All codes and trained models are publicly available at: https://osf.io/akwty

2.4.3 Model fitting

We fit the model parameters such that the average human accuracy in different conditions matches the average model accuracy in corresponding condition. For RTNet, we tested the average accuracy across all instantiations under a wide range of image noise level and threshold values and picked the values that produce performances closest to human performance in each condition. Specifically, we altered the amplitude of uniform noise from 1 to 10 with increments of 1, and we set the threshold to a value from 2 to 12 with increments of 2. The results were close to human produced results when the noise was in range 2-3 for easy images and 4-5 for difficult images. For threshold, the best results achieved when it was set to 2 or 4 for speed focus condition and 6 or 8 for accuracy focus condition. We then searched the parameters space greedily near those values to provide the best parameter tuning for models. We looked at the model performance 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 best performance was achieved when noise level of 2.1 and 4.1 were used for easy and difficult images, respectively. For threshold, the best performance was achieved when it was set to 3 for speed focus condition and 6 for accuracy focus condition.

A Similar procedure was taken to tune the parameters for Anytime Prediction. Note that the threshold value for Anytime Prediction is the probability of a choice option being correct at each early exit. This probability is computed by softmax function at each output
layer. After parameter tuning, we found that the values of 0.58 and 0.82 provide the closest fits to human data for speed focus and accuracy focus conditions, respectively. Within these conditions, the background noise was set to 1.9 and 3.0 for easy and difficult images, respectively.

2.5 Model data analyses

All pre-processing steps for model data were similar to human data pre-processing (see section 2.2). We first checked the model instantiations to ensure that they all provide faster responses in speed focus condition compared to accuracy focus condition. We then assessed whether any of the models show ceiling or floor effects for accuracy or confidence quantities. Since the confidence scales for RTNet or Anytime Prediction were different from human data (i.e., 4-point confidence scale), we first transferred the confidence data to 4-point scale and then examined the floor and ceiling effects with the values used for behavioral data. None of the trained instantiations showed such an effect. Finally, we removed the outliers in each condition for instantiations using Tukey’s interquartile method (refer to section 2.2 for more details). The accuracy of responses was matched to human data in all four conditions after pre-processing steps.

Model data analyses were same as the behavioral data analyses, with only one exception. When skewness of the RT distributions was examined, a random value was added to the individual RTs to reflect the non-decision time in humans. The non-decision time is part of the RT which is not related to the decision process itself and it is the time needed for processes such as encoding and motor response execution. The non-decision time was shown to be an important factor in observed RT patterns within a condition.
(Brown & Heathcote, 2005; Heathcote & Love, 2012; Ratcliff & Rouder, 1998; Weindel, Gajdos, Burle, & Alario, 2021). Therefore, we added a random value that comes from a uniform distribution account for non-decision time in model provided RTs. The uniform distribution was ranged from the minimum RT value for a model to the next minimum value. For example, in case of Anytime Prediction, since the RTs can only take a value between 1-5, we chose a random value in range 1-2 which we believe to be the best representative of non-decision time according to previous empirical studies (Ratcliff, Thapar, & McKoon, 2004; Weindel et al., 2021).
CHAPTER 3.  RESULTS

3.1 Signatures of human decisions

The data from previous decision-making tasks suggests that the choice property distributions have some signatures that any model of decision making should try to account for (Brown & Heathcote, 2008; Forstmann et al., 2016; Luce, 1986; Ratcliff, 2002; Ratcliff & McKoon, 2008; Wagenmakers & Brown, 2007). Some of these signature patterns are very important and are considered as the basic characteristics of any decision-making model. For example, 1) Human responses to a certain stimulus is highly variable and the variability depends on experimental factors (Beck et al., 2012; Renart & Machens, 2014), 2) increasing speed stress shortens mean RT but increase the proportion of erroneous trials (speed-accuracy trade-off) (Forstmann et al., 2016; Heitz, 2014; Heitz & Schall, 2012), 3) mean RT is shorter for easy stimuli than it is for hard stimuli (Forstmann et al., 2016; Ratcliff & Rouder, 1998; Wagenmakers & Brown, 2007), 4) RT distributions are right-skewed, and this skew increases with task difficulty (Forstmann et al., 2016), 5) mean error RT is often slower than mean correct RT (Brown & Heathcote, 2008; Forstmann et al., 2008; Luce, 1986; Ratcliff, 2002; Wagenmakers & Brown, 2007), 6) average confidence score for correct trials is higher than that for error trials (Rahnev, 2021). Most of the studies which reported these signature patterns during a decision-making task employed simple tasks with only two alternatives and artificial stimuli (e.g., Gabor Patches).

In this section, we first examined these effects in our data to see whether they hold for eight-choice task with more naturalistic stimuli. Our dataset consists of data from 60 subjects who were asked to perform a digit discrimination task. The experiment was a 2 x
2 design with factors of task difficulty (easy vs. difficult images) and speed pressure (speed focus vs. accuracy focus). Each condition consisted of 120 unique images, and each subject made decision regarding each image exactly twice. Therefore, each subject went through 960 trials, in total, which was designed to be in four runs of four blocks. All of the images were presented in the first two runs, and the second appearance of them occurred in the second two runs. At the beginning of each block, subjects were asked to adjust their responding policy by either focusing on speed or accuracy of their decisions. Each block contained 30 easy images and 30 difficult images interleaved in a random order.

After examining the patterns in data, we assess RTNet and Anytime Prediction behavior in equivalent situation and compare their behavior with human behavior. To provide a fair comparison between human and model produced data, we matched the average accuracy in each condition for each model to the average performance of the humans. Specifically, we tuned the threshold value and image noise in model population to achieve human like performance. For RTNet, a separate accumulation system was considered for each choice option and hence the height of threshold was manipulated to achieve the desired values. For Anytime Prediction, the threshold value corresponds to the probability by which a choice option is correct according to the trained network. Also, since the level of noise was reported to be different across humans and deep learning models to achieve similar performance (Geirhos et al., 2017, 2018), we manipulated the noise levels to achieve the performance values observed in human data. For easy images, average accuracy across 60 subjects was 83.98% and 80.55% in accuracy focus and speed focus condition, respectively. For equivalent images, after parameter tuning, RTNet performed 83.70% and 82.12%, and Anytime Prediction 83.79% and 80.83% in accuracy
focus and speed focus condition, respectively. For difficult images, humans were 61.31% and 59.83% accurate in accuracy focus and speed focus condition, respectively. For equivalent difficult images, RTNet was 61.14% and 60.18% accurate in accuracy focus condition and speed focus condition, respectively. Finally, for Anytime Prediction, the parameters tuned such that it resulted in 61.38% and 59.81% accuracy level in accuracy focus and speed focus condition, respectively. Statistically, none of the models performed significantly different from each other, within each condition.

3.1.1 Variability in human responses

One important feature for human behavior is that human choices are highly variable even for the same stimulus (Beck et al., 2012; Renart & Machens, 2014; Wyart & Koechlin, 2016). Here we first show this effect in our behavioral data under different conditions, and then explore the similarities between response variability in humans and RTNet or Anytime Prediction. In our experiment, each subject responded to each stimulus exactly twice. The two appearance of the stimulus is within a single condition. Therefore, a specific image (difficult or easy) appears only under its predefined condition (accuracy focus or speed focus) for all the subjects. This ensures that we measure choice properties for a specific image under a certain condition across all subjects and enables us to compute the response variability between the first appearance of an image and its second appearance under a certain experimental condition.

To show the response variability, we computed the consistency of responses to a specific stimulus in its first and second appearance. This results in four different scenarios: 1) the responses are consistent, and they both are correct (consistent-two-correct), 2) the
responses are inconsistent and one of them is correct (inconsistent-one-correct), 3) the responses are consistent but both are incorrect (consistent-zero-correct), and 4) the responses are inconsistent and both are incorrect (inconsistent-zero-correct). We observed that the human responses to the same stimulus is variable. On average, 36% of the trials across subjects yielded in inconsistent responses to the same stimuli, which was significantly greater than zero (one-sample t-test: \( t(59) = 36.78, p < 0.0001 \)) (Figure 3A).

We also examined whether the inconsistency is a function of experimental condition. A repeated measures ANOVA revealed both main effect of speed pressure (F(1,63) = 9.14, \( p = 0.0036 \)) and a main effect of task difficulty (F(1,63) = 871.87, \( p < 0.0001 \)), as well as a significant interaction between the two (F(1,63) = 5.12, \( p = 0.0271 \)).

Figure 3. Variability in human, RTNet and Anytime Prediction responses for same images in different conditions. We measured the consistency of responses in different experimental conditions for (A) humans, (B) RTNet and (C) Anytime Prediction. In each condition, we compared the first and second responses to the same stimulus and computed
the proportion of 1) consistent responses that both were correct (consistent-two-correct), 2) inconsistent responses where only one of them was correct (inconsistent-one-correct), 3) consistent responses where both of them were incorrect (consistent-zero-correct), 4) inconsistent responses where both of them were incorrect. Humans and RTNet provided inconsistent responses for same stimuli whereas Anytime Prediction always produced consistent responses due to its deterministic nature.

To assess whether RTNet produces variable responses to the same stimulus, we tested the model for same stimuli two times under different speed conditions. We then computed the consistency of responses between the two outcomes as it was measured for the human responses. We found that RTNet produces variable responses to the same stimulus (Figure 3B). On average, 20% of trials resulted across model instantiations, resulted in inconsistent response for the same stimuli (one-sample t-test: t(59) = 32.65, p < 0.0001). We also examined the inconsistency proportion under different experimental conditions and found main effect of speed pressure (repeated measures ANOVA: F(1,59) = 87.73, p < 0.0001) and main effect of task difficulty (repeated measures ANOVA: F(1,59) = 120.12, p < 0.0001) but not a significant interaction effect (F(1,59) = 0.10, p = 0.7567). We performed a similar procedure for Anytime Prediction and observed that it only produces consistent results (both responses are either correct or they both are incorrect) when it was shown the same stimuli (Figure 3C). This observation is not surprising, given the deterministic nature of Anytime Prediction. In fact, RTNet relies on noisy accumulation of evidences which can land on inconsistent choices when it is shown the same image. This stochasticity in decision-making enables a behavior similar to human behavior, while deterministic nature of Anytime Prediction is a barrier in providing inconsistent outcomes for the same image similar to what humans show.
3.1.2 Speed-accuracy trade-off

The ability to trade-off speed and accuracy against each other is a hallmark of decision-making across humans and tasks (Heitz, 2014; Heitz & Schall, 2012). Here we show that overall increase in speed of responses decreased the accuracy for humans as well as RTNet and Anytime Prediction. For this reason, we first computed the proportion of correct decisions to the total number of trials in each condition for each subject and computed the average accuracy across subjects (Figure 4A). Results indicated that emphasizing on speed of responses decreases the average accuracy. Two-way repeated measures ANOVA showed statistically significant difference in average accuracy led by speed pressure (F(1, 59) = 4.27, p = 0.0431) or level of noise (F(1, 59) = 1558.50, p < 0.0001). For models, results showed the main effect of speed manipulation both in RTNet (F(1,59) = 11.93, p = 0.0010) (Figure 4B) and Anytime Prediction (F(1,59) = 21.84, p < 0.0001) (Figure 4C). Further, increasing noise level resulted in decreasing accuracy both for RTNet (F(1,59) = 229.46, p < 0.0001) and Anytime Prediction (F(1,59) = 247.52, p < 0.0001). These results indicate that we replicated speed-accuracy trade-off phenomenon with the current design in humans and the models exhibited it robustly by appropriate parameter matching.
3.1.3 **RT is shorter for easy stimuli than difficult stimuli**

One of the ubiquitous findings regarding the human decisions is that RT for difficult stimuli, on average, is greater than RT for easy stimuli (Forstmann et al., 2016; Gold & Shadlen, 2007). To explore the effect here, we computed the average RT for each subject in each condition and found that our data replicates previous findings (Figure 5A). The easy stimuli correspond to shorter mean RT when compared to difficult stimuli (two-way repeated measures ANOVA: $F(1,59) = 411.15, p < 0.0001$). The same trend was observed for RTNet (two-way repeated measures ANOVA: $F(1,59) = 223.97, p < 0.0001$) (Figure 5B) and Anytime Prediction (two-way repeated measures ANOVA: $F(1,59) = 6.17, p = 0.0158$) (Figure 5C). Altogether, the results indicate that both human and the models, on average, take less time to process easy stimuli when compared to difficult stimuli.
**Figure 5. Shorter RT for easy stimuli than difficulty stimuli.** (A) In human data, difficult trials are associated with longer RTs. Both (B) RTNet and (C) Anytime Prediction showed the same effect. *Error bars show standard error of the mean (SEM).*

### 3.1.4 Skewness of RT distributions

For simple two-choice decisions, empirical RT distributions for humans are generally positively skewed and the skewness changes as a function of task conditions (Forstmann et al., 2016; Ratcliff & McKoon, 2008). Here we assess the quality of distributions produced in our eight-choice task by humans and the models. We explored the shape of distributions for a representative subject or model using a kernel density estimate (KDE) **(Figure 6)**. RTNet and human produced distributions are similar in that the skewness increases in their RT distributions when the speed pressure decreases **(Figure 6A, Figure 6B)**. Also, within each speed condition, difficulty trials are more skewed to right than easy conditions. However, the effect of task difficulty on RT distributions for Anytime Prediction changes as a function of speed pressure, and it does not reflect the patterns observed in human data **(Figure 6C)**.
Figure 6. Skewness of RT distributions for a representative subject/model. RT distributions for a representative subject/model change as a function of experimental condition. For (A) humans and (B) RTNet, RT distributions in accuracy focus condition is more skewed to the right than speed focus condition. Within each speed pressure condition, RT distributions associated to difficult condition is more skewed to right than easy condition. (C) Anytime Prediction does not provide patterns observed in human RT data.

In addition, we examined the RT distributions in each condition to assess how the skewness change under different conditions (Figure 7). For this reason, we first computed the skewness of RT distributions for each subject in each condition and averaged them across subjects in human data. We then performed two-way repeated measures ANOVA to assess the effect of speed pressure and task difficulty on skewness of RT distributions. The results indicated that the skewness changes as a function of both difficulty and speed pressure (Figure 7A). RT distributions are more skewed to the right for accuracy focus responses compared to speed focus responses (F(1,59) = 32.84, p < 0.0001). In addition, higher skewness was observed in RT distributions associated with easy stimuli than difficult stimuli (F(1,59) = 5.10, p = 0.0277). We followed the same procedure for model evaluations and found that RTNet shows a similar trend. The skewness of RT distributions decreases by increasing the speed pressure (F(1,59) = 156.71, p < 0.0001) or by increasing the difficulty of stimulus (F(1,59) = 13.19, p = 0.0006) (Figure 7B). However, the trend was different for RTs produced by Anytime Prediction. For this model, we found main
effect of speed pressure (F(1,59) = 52.75, p < 0.0001) and an interaction effect across two different experiment factors (F(1,59) = 13.98, p = 0.0004), but not main effect of task difficulty (F(1,59) = 2.31, p = 0.1336) (Figure 7C). Overall, RTNet produced RT distributions which reflected the observed patterns in human data better than Anytime Prediction. The main reason here is that Anytime Prediction is confined to produce RTs less than or equal to its layer numbers. However, RTNet is capable of going through arbitrary number of samples, similar to sequential sampling models and hence produce RTs with a higher resolution.

Figure 7. Skewness of RT distributions across subjects/models. The skewness of RT distributions change across conditions. For (A) human data, skewness of RT distributions changed by task difficulty and speed pressure. The same trend was observed for (B) RTNet, whereas (C) Anytime Prediction showed main effect of speed pressure and an interaction effect, but not main effect of task difficulty. Error bars show standard error of the mean (SEM).

3.1.5 Error trials RT is slower than correct trials RT

Another ubiquitous observation in fast two-choice tasks is that erroneous decisions are associated with higher RT than correct trials (Brown & Heathcote, 2008; Forstmann et al., 2008; Luce, 1986; Ratcliff, 2002; Wagenmakers & Brown, 2007). We explored this effect in here by separating the data based on accuracy of decisions and computed the mean
RT both for error and correct trials for each subject. We then computed the average RT across subjects and found that error trials are associated to higher RT than correct trials (two-way repeated measures ANOVA: $F(1,59) = 82.08$, $p < 0.0001$) (Figure 8A). We performed the same analysis for RTNet and found the same trend (two-way repeated measures ANOVA: $F(1,59) = 638.78$, $p < 0.0001$) (Figure 8B). However, Anytime Prediction provided slower correct RTs when compared to error RTs (two-way repeated measures ANOVA: $F(1,59) = 65.70$, $p < 0.0001$) (Figure 8C). We replicated the previous findings in our dataset and showed that RTNet shows a similar effect. However, Anytime Prediction failed showed an opposite effect and the correct trials took longer for it when compared to error trials.

Figure 8. Error trials are slower than correct trials. For (A) humans and (B) RTNet error trials are associated with higher RT than correct trials. However, (C) Anytime Prediction showed an opposite pattern and correct trials took a longer processing time. Error bars show standard error of the mean (SEM).

3.1.6 Confidence is higher for correct decisions than error decisions

Finally, it has consistently been shown that confidence in correct choices is stronger than confidence in incorrect choices (Rahnev, 2021; Yeung & Summerfield, 2012). Here we observed the same effect across subjects when explored the differences in average
confidence scores separated by accuracy of trials ($F(1,59) = 472.17, p < 0.0001$) (Figure 9A). Higher average confidence score was also observed for RTNet ($F(1,59) = 1021.53, p < 0.0001$) (Figure 9B) and Anytime Prediction ($F(1,59) = 131.92, p < 0.0001$) (Figure 9C) when a similar procedure was taken to explore the differences in confidence scores for error and correct trials. Therefore, humans, RTNet and Anytime Prediction robustly showed higher confidence for correct trials compared to incorrect trials, a behavior which was vastly reported by previous studies in literature.

Figure 9. Higher confidence for correct trials than error trials. The confidence scores for correct trials in (A) human data was higher, on average, when compared to error trials. The same effect was observed both for (B) RTNet and (C) Anytime Prediction. Error bars show standard error of the mean (SEM).

3.2 Similarities between model and human data

3.2.1 Correlation between the model and average human data

To assess the similarities between human produced data and the models, we looked at the correlations between human and model produced RT, accuracy, or confidence scores across images. Specifically, we first looked at the correlations between each model representative and average human data. The model representative is the average of all 60 instantiations trained for each model. Therefore, each model representative has a value...
associated to each image which simply comes from taking the average of all different instantiations for a choice property (e.g. RT). To find the correlation, we performed a regression analysis by defining the average human data for a measure (e.g. RT) as dependent variable and the model representative as independent variable.

We found that RTNet performs better in predicting the human data than Anytime Prediction. Specifically for RT, the correlation between RTNet model representative and average human data was 0.77 ($F(1,478) = 717.10$, $p < 0.0001$) whereas for Anytime Prediction, the correlation was 0.49 ($F(1,478) = 149.8$, $p < 0.0001$). For accuracy, the correlation between RTNet and average human data was 0.61 ($F(1,478) = 268.1$, $p < 0.0001$) whereas Anytime Prediction and average human data correlation was 0.52 ($F(1,478) = 173.5$, $p < 0.0001$). Finally for confidence, RTNet was significantly correlated with human data and the correlation was equal to 0.61 ($F(1,478) = 280.8$, $p < 0.0001$). Confidence correlation for Anytime Prediction was 0.5 ($F(1,478) = 161.90$, $p < 0.0001$), though. To compare these values, we ran Fisher’s z-test and found that RTNet provided higher correlation for RT ($z(480) = 7.48$, $p < 0.0001$), accuracy ($z(480) = 2.05$, $p = 0.0406$) and confidence ($z(480) = 2.47$, $p = 0.0137$) compared to Anytime Prediction. Overall, the average data from RTNet instantiation population predicted average human data better than average Anytime Prediction instantiations.

3.2.2 Correlation between the model and individual subjects

In addition to correlations between average human data and the models, we examined the similarities between each individual subject data and the models. We computed the correlation between a model representative and each subject data across
images for RT, accuracy, or confidence. We then compared RTNet performance with Anytime Prediction performance and human consistency. Human consistency was computed by correlating the individual subject performance against the average of all other human subjects. This procedure provides a lower bound on the noise ceiling. That means, it provides a lower bound on the performance that a true model could achieve given inter-subject variability (Spoerer et al., 2020). The goal here was to see to what extent the average representative of each model (i.e. RTNet or Anytime Prediction) can predict individual subject data.

We first looked at similarities in produced RTs by human and the models by computing the correlation between representative model RT and individual subject RT across images. While the human consistency for RT was equal to 0.54 (t(59) = 39.74, p < 0.0001), we found average correlation value of 0.43 and 0.27 for RTNet and Anytime Prediction, respectively. These average values was significantly greater than zero both for RTNet (one-sample t-test: t(59) = 19.60, p < 0.0001) and Anytime Prediction (t(59) = 21.72, p < 0.0001). Overall, RTNet performed significantly better than Anytime Prediction in predicting the associated RT values for images (paired t-test: t(59) = 12.80, p < 0.0001) (Figure 10A).
Figure 10. Correlation between average model RT and individual subjects’ RT. (A) There is a RT associated with each image provided by individual subjects or model representatives. Each point shows the correlation between an individual subject RT and model representative RT across images. Human consistency model is provided as a lower bound on noise ceiling. The average correlation for RTNet is significantly higher than that for Anytime Prediction. Error bars show standard error of the mean (SEM). (B) The RT correlation between individual subject and a model representative (i.e. RTNet or Anytime Prediction) is plotted against its corresponding correlation value in human consistency model. The regression line with positive slope indicates that subjects which are more correlated with average of other subjects, they are also more correlated with average of the examined model.

The correlation between an individual subject and rest of the subjects, not only provides a lower bound on noise ceiling, but also it provides the information that to what extent an individual subject data deviates from the best possible model available. The same idea could be applied to look at the deviations of individual subjects from a proposed model. Preferably, for a good model, the direction of deviation for an individual subject should remain same as the best possible model (i.e. average of other subjects). To explore this effect, we performed a regression analysis by defining the correlation between each subject and rest of the subjects as a dependent variable and the correlation between each model and individual subjects as independent variable. Positive regression slope would
indicate that the model performs similar to the average of human data and the amplitude of
the correlation is indicative of how well the model performs.

For RT, we found that the subjects which are more correlated with rest of the subjects, they are also more correlated with RTNet model (adjusted $R^2 = 0.66$, $F(1,58) = 117.50$, $p < 0.0001$) (Figure 10B). Anytime Prediction showed a weaker but similar trend (adjusted $R^2 = 0.63$, $F(1,58) = 103.20$, $p < 0.0001$). Altogether, the results showed that when RT data for a subject is more correlated with rest of the population, it is also more correlated with RT produced by RTNet or Anytime.

Next, we explored the similarities in accuracy of trials between humans and RTNet or Anytime Prediction. We wanted to see to what extent the models can predict the individual subject accuracy data. Similar to the procedure we took for RT analysis, we computed the average model accuracy for each image and found the correlation between this outcome and each subject. The average correlation for RTNet and Anytime Prediction was 0.35 and 0.13 respectively, and both models provided average correlations which were significantly higher than zero (one sample t-test for RTNet: $t(59) = 43.82$, $p < 0.0001$; one sample t-test for Anytime Prediction: $t(59) = 8.52$, $p < 0.0001$) (Figure 11A). The human consistency for accuracy measure was 0.56 (one-sample t-test: $t(59) = 59.21$, $p < 0.0001$). Also, direct comparison of models showed that, on average, RTNet is more correlated with subjects’ data than Anytime Prediction (paired t-test: $t(59) = 18.07$, $p < 0.0001$). Overall, we found that both RTNet and Anytime Prediction are significantly correlated with individual subject data but RTNet performs better than in providing higher correlation values when compared to Anytime Prediction.
Figure 11. Correlation between average model accuracy and individual subjects’ accuracy. (A) There is an accuracy associated to each image provided by individual subjects or model representatives. Each point shows the correlation between an individual subject accuracy and model representative accuracy across images. Human consistency model is provided as a lower bound on noise ceiling. The average correlation for RTNet is significantly higher than that for Anytime Prediction. Error bars show standard error of the mean (SEM). (B) The accuracy correlation between individual subject and a model representative (i.e. RTNet or Anytime Prediction) is plotted against its corresponding correlation value in human consistency model. The regression line with positive slope indicates that subjects which are more correlated with average of other subjects, they are also more correlated with average of the examined model.

In addition to similarities between model and individual subject data, we wanted to see whether the models preserve the deviation direction for a subject similar to what we did for RT measure. Regression analysis showed that when a subject is more correlated with average of other subjects, they are also more correlated with RTNet (adjusted $R^2 = 0.45$, $F(1,58) = 48.90$, $p < 0.0001$) or Anytime Prediction (adjusted $R^2 = 0.50$, $F(1,58) = 58.97$, $p < 0.0001$) accuracies provided for images (Figure 11B). This indicates that both models provide accuracy measures for associated images similar to humans and they both are good representative of the best available model, which is the average human data.
Finally, we assessed which model provides more similar confidence scores for images under different conditions based on individual subject data. A procedure same as the one taken for RT or accuracy was taken for confidence ratings. Human consistency for confidence was equal to 0.37 (one-sample t-test: $t(59) = 18.75$, $p < 0.0001$). The average correlation value for RTNet and Anytime Prediction was equal to 0.25 and 0.20, respectively and they both were significantly greater than zero (one sample t-test for RTNet: $t(59) = 16.74$, $p < 0.0001$; one sample t-test for Anytime Prediction: $t(59) = 15.84$, $p < 0.0001$) (Figure 12A). In addition, the average correlation for RTNet is statistically greater than that for Anytime Prediction (paired t-test: $t(59) = 7.48$, $p < 0.0001$). Despite the fact that confidence judgements are typically noisier than accuracy data (Shekhar & Rahnev, 2021), both of the models showed significant correlations with individual data, with RTNet performing significantly better than Anytime Prediction.

**Figure 12.** Correlation between average model confidence and individual subjects’ confidence. (A) There is a confidence score associated to each image provided by individual subjects or model representatives. Each point shows the correlation between an individual subject confidence and model representative confidence across images. Human consistency model is provided as a lower bound on noise ceiling. The average correlation for RTNet is significantly higher than that for Anytime Prediction. Error bars show standard error of the mean (SEM). (B) The confidence correlation between individual subject and a model representative (i.e. RTNet or Anytime Prediction) is plotted against its
corresponding correlation value in human consistency model. The regression line with positive slope indicates that subjects which are more correlated with average of other subjects, they are also more correlated with average of the examined model.

We also interested in weather RTNet or Anytime prediction preserve the direction of correlation deviation for individual subjects. To answer this question, we performed regression analysis similar to what we did for RT or accuracy and found that when a model is more correlated with average of other human subjects, it is also more correlated with average RTNet confidence scores (adjusted $R^2 = 0.83$, $F(1,58) = 293.20$, $p < 0.0001$) or Anytime Prediction confidence scores (adjusted $R^2 = 0.72$, $F(1,58) = 155.90$, $p < 0.0001$) (Figure 12B). Therefore, both models are behaving very similar to average human data for confidence in that the subjects who are more correlated with average human data, they are also more correlated with RTNet or Anytime Prediction.

3.3 Similarities between model and human in each experimental condition

3.3.1 Correlation between the model and average human data

In this section, we explore at similarities between model and human behavior separately in each condition. This provides us with the opportunity to explore how each model performs for unique images incorporated in each condition regardless of the confounds that can drive the correlation. These confounds are the factors incorporated in the experiment such as different difficulty levels for images, or instructions given to subjects (or models) to respond by focusing on either speed or accuracy. Therefore, exploring the correlations within conditions provides us with information that how human and model behavior are similar for unique images.
To find how well each model performs, here we examine the similarities between average human data and the model representative. As mentioned earlier, the model representative is the average model performance across its all 60 instantiations. Overall, we found significant correlations between average human data and RTNet or Anytime Prediction for all conditions (Table 2). For RT, it means that when an image takes more time for humans to process in a specific condition, it also takes more time for RTNet or Anytime Prediction to process that image. A similar conclusion can be drawn for accuracy and confidence measures. In addition, direct comparison of correlation values between models in each condition, shows that RTNet provides higher correlation values than Anytime Prediction for predicting RT or accuracy. However, the difference was significant, only in accuracy focus condition with difficult images (Fisher’s z-test: $z(120) = 3.17, p = 0.0015$). For confidence, Anytime Prediction outcome resulted in slightly higher correlation values in general, but none of them were significantly different from RTNet produced correlations. Overall, within condition, RTNet and Anytime Prediction performed similarly for predicting average RT, accuracy or confidence across subjects for individual images with only one exception where RTNet showed a higher correlation than Anytime Prediction when predicting average RT in accuracy focus condition with difficult images.
Table 2. Correlation between the average model performance and average human data in each condition. The RTNet model generally performed better than the Anytime Prediction model for RT and decision accuracy measures. Whereas Anytime Prediction was more correlated with average human data for confidence measure in different conditions. All correlation values are significantly greater than zero.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Model</th>
<th>Condition</th>
<th>Accuracy focus easy</th>
<th>Accuracy focus difficult</th>
<th>Speed focus easy</th>
<th>Speed focus difficult</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT</td>
<td>RTNet - Human</td>
<td>Accuracy focus easy</td>
<td>0.37</td>
<td>0.52</td>
<td>0.37</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy focus difficult</td>
<td>0.24</td>
<td>0.16</td>
<td>0.29</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>Anytime - Human</td>
<td>Speed focus easy</td>
<td>0.52</td>
<td>0.5</td>
<td>0.51</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speed focus difficult</td>
<td>0.48</td>
<td>0.38</td>
<td>0.42</td>
<td>0.38</td>
</tr>
<tr>
<td>accuracy</td>
<td>RTNet - Human</td>
<td>Speed focus easy</td>
<td>0.56</td>
<td>0.48</td>
<td>0.51</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speed focus difficult</td>
<td>0.63</td>
<td>0.48</td>
<td>0.56</td>
<td>0.45</td>
</tr>
<tr>
<td>confidence</td>
<td>Anytime - Human</td>
<td>Speed focus easy</td>
<td>0.56</td>
<td>0.48</td>
<td>0.51</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speed focus difficult</td>
<td>0.63</td>
<td>0.48</td>
<td>0.56</td>
<td>0.45</td>
</tr>
</tbody>
</table>

3.3.2 Correlation between the model and individual subjects

Finally, we looked at the similarities between average model performance and individual subjects. This enables us to gain information about the correlation between the model representative and individual subjects which indicates that how a model can predict the choice distributions for unique images in subject-level data. We first assess each model’s ability in predicting individual subject data for RT, accuracy and confidence in each condition and then compare the RTNet performance to Anytime Prediction performance for these measures. We also provide human consistency as the noise ceiling.
to provide information that how well a model can perform given the inter-subject variability.

The results showed that both RTNet or Anytime Prediction produced RT, accuracy or confidence is significantly correlated with individual subject data in all conditions (Figure 13). However, direct comparison of models revealed that for RT and accuracy, RTNet predicts the individual data better than Anytime Prediction. Specifically, for RT, RTNet performs better than Anytime Prediction in all four conditions (paired t-test: \( t_{\text{accuracy focus, easy}}(59) = 5.36, p < 0.0001; t_{\text{accuracy focus, difficult}}(59) = 8.90, p < 0.0001; t_{\text{speed focus, easy}}(59) = 3.16, p = 0.0025; t_{\text{speed focus, difficult}}(59) = 3.08, p = 0.0031 \)). For accuracy, the average correlation for RTNet is significantly higher than Anytime Prediction in speed focus condition (paired t-test: \( t_{\text{speed focus, easy}}(59) = 5.52, p < 0.0001; t_{\text{speed focus, difficult}}(59) = 5.72, p < 0.0001 \)) and accuracy focus condition with difficult images (paired t-test: \( t(59) = 4.84, p = 0.0010 \)) but not in accuracy focus condition for easy images (\( t(59) = 0.07, p = 0.9447 \)).

For confidence, Anytime Prediction performed better than RTNet only in accuracy focus condition with easy images (paired t-test: \( t(59) = 2.62, p = 0.0113 \)), but not in other conditions (paired t-test: \( t_{\text{accuracy focus, difficult}}(59) = 0.14, p = 0.8857; t_{\text{speed focus, easy}}(59) = 0.61, p = 0.5470; t_{\text{speed focus, difficult}}(59) = 1.70, p = 0.0947 \)). Altogether, RTNet model performed better than Anytime Prediction for RT and accuracy measures but not for confidence when the similarities between the model representatives and individual subjects were assessed for each experimental condition separately.
Figure 13. Correlation between average model performance and individual subjects within each condition. (A) shows the correlation between an individual subject and average of other subjects (left), RTNet (middle), Anytime Prediction (right) for RT measure. All average values are significantly greater than zero. RTNet predicts individual RT data significantly better than Anytime Prediction in all experimental conditions. Also, RTNet preserves the patterns observed in human data, but Anytime Prediction only preserves the patterns across task difficulty conditions and fails to preserve the patterns across speed pressure conditions. (B) shows the correlation between an individual subject and average of other subjects (left), RTNet (middle), Anytime Prediction (right) for accuracy measure. All average values are significantly greater than zero. RTNet predicts individual accuracy data significantly better than Anytime Prediction in all experimental conditions except accuracy focus with easy images where both models perform equally well. Also, RTNet showed correlation patterns similar to human data across conditions but Anytime Prediction only succeeded to preserve patterns across speed pressure conditions but not across task difficulty. (C) shows the correlation between an individual subject and average of other subjects (left), RTNet (middle), Anytime Prediction (right) for confidence measure. All average values are significantly greater than zero. Anytime Prediction...
predicts individual confidence data significantly better than RTNet in accuracy focus condition with easy images, but in other conditions the models perform equally well. Both models preserved the correlation patterns across speed pressure conditions but failed to preserve it across task difficulty conditions. Error bars show standard error of mean (SEM).

Further, we examined how the experimental factors (i.e. difficulty and speed pressure) affects the correlation patterns. For this reason, we first looked at the existing patterns in human data by looking at the correlation between individual subjects and average of other subjects. We then explored whether the models preserve the observed patterns across conditions. We found that correlation patterns for RT change as a function of speed pressure and task difficulty using all of the models (Figure 13A). Subjects showed stronger correlation with average of other subjects when they were making decision regarding easy stimuli (repeated measures ANOVA: F(1,59) = 84.71, p < 0.0001) or when they were responding without time pressure (repeated measures ANOVA: F(1,59) = 6.99, p = 0.0105). Similar patterns were observed for RTNet (repeated measures ANOVA: F_sat(1,59) = 12.20, p = 0.0009; F_noise(1,59) = 5.06, p = 0.0282). For Anytime Prediction, noise manipulation had the same effect as human data but stronger correlation was found when the model was focusing on speed rather than accuracy in its decisions (repeated measures ANOVA: F_sat(1,59) = 5.81, p = 0.0190; F_noise(1,59) = 16.63, p = 0.0001). Therefore, RTNet showed correlation patterns across experimental conditions similar to human data, whereas Anytime Prediction only preserved the patterns for speed pressure manipulations and showed opposite effects for difficulty manipulations.
Similar to the procedure that was taken for RT measure, we performed analysis for accuracy and confidence to look at the correlation patterns across conditions. For accuracy, the RTNet performed better in producing patterns similar to human data when compared to Anytime Prediction (Figure 13B). For human data, on average, the accuracy correlation was significantly higher when subjects were responding to difficult stimuli (repeated measures ANOVA: F(1,59) = 95.13, p < 0.0001) or when they were in accuracy focus condition (repeated measures ANOVA: F(1,59) = 18.28, p = 0.0001). Similarly, for RTNet model, we found main effect of task difficulty using repeated measures ANOVA (F(1,59) = 20.19, p < 0.0001) and main effect of speed pressure (F(1,59) = 5.10, p = 0.0276). On contrary, for Anytime Prediction, we did not find main effect of task difficulty (F(1,59) = 0.09, p = 0.7688), but main effect of speed pressure (F(1,59) = 7.31, p = 0.0089). Therefore, for accuracy, RTNet showed patterns similar to human data but Anytime Prediction only succeeded to preserve patterns across speed pressure conditions but not across task difficulty.

Finally, for confidence, we found similar performance for RTNet and Anytime Prediction when the patterns across conditions were examined (Figure 13C). We found main effect of speed pressure (F(1,59) = 107.40, p < 0.0001) and main effect of task difficulty (F(1,59) = 10.04, p = 0.0024) for human data using repeated measures ANOVA. For RTNet, average confidence correlation was significantly higher for accuracy focus condition than speed focus condition (repeated measures ANOVA: F(1,59) = 46.23, p < 0.0001), but we found no main effect of task difficulty (repeated measures ANOVA: F(1,59) = 0.91, p = 0.3446). The same patterns were observed for Anytime Prediction (repeated measure ANOVA: F_{difficulty} (1,59) = 2.33, p = 0.1322; F_{sat} (1,59) = 46.69, p <
0.0001). Overall, both models preserved the correlation patterns across speed pressure conditions, but failed to preserve it across task difficulty conditions.
CHAPTER 4. DISCUSSION

4.1 Summary of aim and results

In this study, we built a model of human decision-making, called RTNet, that reflects the signatures of human choice data, and can predict the properties of a decision associated to a novel image. The model is capable of looking at a novel image and accumulating evidence in favor of available choice options through time to make a decision. Therefore, it can produce a variable RT and even a variable response, associated to a unique image on different trials, which is a basic observation in human decisions. The backbone of the model consists of a Bayesian neural network which enables generalization of predictions to novel images, but with noisy weights instead of deterministic weights. A separate accumulation system is also considered for each choice option which enables sampling and evidence accumulation over the course of time in favor of each alternative. To evaluate the model behavior, we compared the model decisions with human decisions while performing an eight-choice digit discrimination task. Subjects performed a 2 x 2 factorial design task with factors of difficulty (i.e., easy vs. difficult images) and speed pressure (i.e., accuracy vs. speed focus). We first assessed the model’s ability to produce decision behavior similar to human subjects. And second, we investigated the extent to which the RT, accuracy and confidence produced by the model for individual images can predict the same quantities produced by human subjects.

We found that RTNet exhibits several important features of human choice data. Most of the understanding regarding human behavior during decision-making comes from laboratory tasks where subjects are asked to perform two-choice tasks with artificial stimuli.
(e.g., Gabor patches) (Bogacz et al., 2006; Forstmann et al., 2016; Ratcliff & Smith, 2004). However, we validated these findings here for an eight-choice task with more natural images (i.e., handwritten digits), and found that the RTNet model reflects all those findings in human data with appropriate parameter tuning. Specifically, both humans and RTNet showed variability in responses for the same stimulus (Bogacz et al., 2006; Findling & Wyart, 2021; Forstmann et al., 2016; Luce, 1986; Ratcliff & Rouder, 1998; Ratcliff & Smith, 2004), speed-accuracy trade-off effect (Forstmann et al., 2016; Heitz, 2014; Heitz & Schall, 2012), longer RT for difficult trials than easy trials (Forstmann et al., 2016; Wagenmakers & Brown, 2007), right-skewed RT distributions as well as higher skewness for accuracy focus condition than speed focus condition (Forstmann et al., 2016; Luce, 1986), longer RT for error trials than correct trials (Bogacz et al., 2006; Forstmann et al., 2016; Luce, 1986; Wagenmakers & Brown, 2007), and finally, higher confidence for correct trials than error trials (Rahnev, 2021).

In addition to behavioral similarities, we showed that the RT, accuracy, and confidence produced by RTNet is correlated with the same measure produced by 60 human subjects. We first investigated RTNet’s ability to predict the choice properties considering all experimental conditions together. Not only RTNet and human produced RT, accuracy, and confidence were correlated, but also, we found that when a subject is more correlated with the average of the other subjects for a given measure, he/she is also more correlated with RTNet produced data. It further highlights the similarities between RTNet produced data and best available model (i.e., average human data) for individual images. Since these correlations might be driven by the mixture of experimental manipulations, we analyzed the similarities for each condition separately. We found significant correlations between
human- and RTNet-produced RT, accuracy, and confidence in all experimental conditions. Overall, RTNet demonstrated significant similarities with human choice data, both across experimental conditions and even for each condition separately.

Throughout the study, we compared RTNet with Anytime Prediction, which is the best existing neural network model of human decision-making. Behaviorally, RTNet performed better than Anytime Prediction because the latter failed to account for three basic features observed in human data: 1) Anytime Prediction did not produce stochastic responses for the same stimulus, a key observation in all human decision-making tasks (Error! Reference source not found.), 2) Anytime Prediction produced RT distributions where the skewness did not systematically change with task difficulty, whereas for humans, RT distributions were more skewed to right for easy stimuli than difficult stimuli (Figure 6, Figure 7), and 3) correct decisions for Anytime Prediction take longer than erroneous decisions, whereas humans showed the opposite pattern (Figure 8). When their capability in predicting choice properties was investigated, RTNet performed better for RT (Figure 10, Figure 13A) and accuracy measures (Figure 11, Figure 13B). However, for confidence scores none of the models outperformed the other (Figure 12, Figure 13C). To our knowledge, RTNet is the first neural network that exhibits all basic signatures of perceptual human decision-making.

4.2 How RTNet is similar to/different from race models of decision-making?

RTNet exhibited the most important and ubiquitous behavioral findings in decision-making literature. There are already existing models in that literature which proved to be successful in explaining these perceptual decision-making signatures for two-choice tasks
(Forstmann et al., 2016; Heathcote & Love, 2012; Heathcote & Matzke, 2022; Ratcliff & Rouder, 1998; Ratcliff & Smith, 2004). The most successful models in this realm are called sequential sampling models which can be divided into two main class of models: random walk/diffusion models (Forstmann et al., 2016; Ratcliff & Smith, 2004) and accumulator/counter models (Heathcote & Love, 2012; Heathcote & Matzke, 2022). In diffusion models which are typically applied to a binary choice task, the evidence from a stimulus in favor of one response alternative is evidence against the other alternative (Ratcliff, 1978; Ratcliff & Rouder, 1998). Therefore, only one accumulation process is considered, and a decision is made as soon as the process reaches one of the two preset boundaries. On the contrary, in accumulator models, each choice option has its own accumulation system, and evidence for each choice is accumulated in parallel (Brown & Heathcote, 2005, 2008). A decision is made as soon as one counter wins the race to reach its own preset boundary. Similar to accumulator models, RTNet benefits from a separate accumulation system for each choice option in its structure. This was selected because the main part of RTNet structure consists of a neural network which is not necessarily limited to two response alternatives. In fact, neural networks can be generalized to any number of choice options during the training process depending on the type of problem they are being designed to tackle. Therefore, considering a separate accumulation system for each response alternative puts RTNet in the category of accumulator models.

There are two main type of accumulator models: the Linear Ballistic Accumulator (LBA) (Brown & Heathcote, 2005, 2008) and the Racing Diffusion Model (RDM) (Tillman, Van Zandt, & Logan, 2020). Both models benefit from two key parameters. First, the threshold which represents the level of caution or response urgency and accounts for
the speed-accuracy trade-off effect. Second, accumulation/drift rate which represents the amount of evidence accumulated per unit time for each accumulator, and it is an index of task difficulty, or stimulus sensitivity. In the LBA, accumulation occurs at a constant rate within a trial whereas in RDM, the accumulation rate is not constant, and it varies from moment to moment. In both models, the accumulation/drift rate vary from trial to trial, but the rate is higher, on average, for the accumulator that matches the stimulus than the accumulators that mismatch the stimulus. The Bayesian neural network module in RTNet, in fact, represents this evidence encoding system and it provides moment to moment information for the accumulation system. At each moment, a random weight is assigned to each edge in RTNet structure. Random weight for a specific edge at each moment, is a random sample from the distribution learnt during training process, associated to that edge. Since each sampling step, result in an arbitrary network (Jospin et al., 2020), the evidence accumulation rate is not necessarily constant, and hence, this makes the RTNet to be similar to RDM in the way evidence accumulates over time. The difference is that, in RTNet, the neural network component with thousands (or maybe millions) of parameters replaces the accumulation rate parameters (which as mentioned, is an index of stimulus sensitivity), and the encoding can occur directly from stimulus to the accumulation system. However, the direct encoding is not possible in RDM and only after parameter fitting, the RDM can make an estimation regarding a stimulus which was perceived before. Therefore, RTNet can be seen as a RDM which can be generalized to novel scenarios, and it’s not limited to existing data.

One may argue that instead of a Bayesian neural network, a standard FNN with a Gaussian noise over its weights may do the same job, and there is no need to model the
noise over weights with complicated Bayesian inference methods. It should be mentioned that in a neural network, each of the parameters has a specific sensitivity with respect to the other parameters and also to the input (the derivative is different with respect to different parameters) (Saltelli et al., 2009). A Gaussian or uniform noise with a specific amplitude over a specific weight, may not change the performance of a network at all, but the same disturbance over another weight may have destructive effects (Ko, Kim, Na, Kung, & Mukhopadhyay, 2017; Koutník, Gomez, & Schmidhuber, 2010; Kung, Kim, & Mukhopadhyay, 2015). One prominent example of this destructive perturbation is adversarial attacks (Chakraborty, Alam, Dey, Chattopadhyay, & Mukhopadhyay, 2018; Xu et al., 2020). An adversarial image is an image that has pixels purposely and intentionally perturbed to confuse a deep network, whereas to the human eye, the image may look indistinguishable from the original. Like humans, Bayesian neural networks, are robust to adversarial attacks (Uchendu, Campoy, Menart, & Hildenbrandt, 2021; Ye & Zhu, 2018). Hence, using a Bayesian neural network not only enables sampling in favor of response alternatives, but also makes it more robust against perturbations which can infuse destructive effects in decision-making such as adversarial attacks. Adding noise with specific amplitude over the weights of a feedforward network is not sufficient and the uncertainty amplitude should be learnt in a systematic way to enable appropriate sampling, with a higher accumulation rate for stimulus-matched accumulator, and less accumulation rate for others.

We confirmed this fact in our pilot studies, that adding a noise with specific amplitude over the weights of a pre-trained network does not result in appropriate behavior. Specifically, we trained multiple AlexNet-based neural networks to perform digit
discrimination task using MNIST dataset. We then added Gaussian noise to weights with specific amplitude. We did this by considering the learnt weights as the mean of Gaussian distributions and an arbitrary value for variance. We tried a wide range of variance values for weights (0.1-5 with increments of 0.2), and observed a consistent behavior from network. That means, for different sampling of the weights, network produced a consistent response (either correct or wrong) regarding an image, regardless of the level of noise. Therefore, it did not produce variability in moment-by-moment samples to provide appropriate behavior that is essential for decision-making. Therefore, we decided to use a Bayesian network which enables a systematic and more meaningful sampling on favor of alternative. Previous empirical studies showed that there is high variability of spike trains for different neurons when experiencing even the same stimulus. One possible explanation provided is that the internal noise for different neurons differ from each other (Habenschuss, Jonke, & Maass, 2013; Yarom & Hounsgaard, 2011). This is consistent with our observation during the pilot studies and prior to creating RTNet. However, a more systematic study is needed to examine the effect of noise over the weights of RTNet and possibly compare the neuronal spike dynamics with an appropriate RTNet architecture that closely resembles the neuronal behavior for a certain task.

Finally, it is worth mentioning that during pilot studies, we extensively examined unrolled RNNs, and observed that their behavior is highly dependent on the original RNN architecture. The unrolled RNNs consistently showed speed-accuracy trade-off as mentioned in a former study (Spoerer et al., 2020). But the degree by which the effect was observed was highly dependent on number of layers used in feedforward path, position of lateral and feedback connections, use of batch normalization layers, etc. More complicated
models resulted in very high performance even after a single sweep of feedforward path and hence did not produce a monotonic increase in accuracy over the course of time. However, as seen in human data provided in this study, humans are less accurate when responded quickly, even for the easy stimuli. A richer account of response urgency on accuracy of responses can be found in previous study (Rafiei & Rahnev, 2021). This does not mean that we oppose in using these types of models, but in fact, we suggest that future studies should focus on finding a robust way of training these networks, even the unrolled versions. Eventually the ideal model would be the one which incorporates both recurrency and noise in its structure because of plethora of evidence pointing to the existence of those two.

4.3 Biological plausibility of RTNet vs. Anytime Prediction

Throughout this study, we compared RTNet against Anytime Prediction both behaviorally and by investigating their capability in predicting human produced RT, accuracy and confidence for individual images. The two models implemented here, shared a common backbone which was an AlexNet structure (Krizhevsky et al., 2012). But the main difference between RTNet and Anytime Prediction is that RTNet has noisy weights which allows the model to sample evidence in favor of choice options over the course of time, whereas Anytime Prediction uses deterministic weights with early exits in intermediate layers to model the notion of time in biological systems. In the following, we further highlight the differences between these two models. We first discuss the biological plausibility of Anytime Prediction and then explore the advantages of using a RTNet-like structure in overcoming some of the limitations introduced by Anytime Prediction.
4.3.1 *How decision-making in brain is different from the Anytime Prediction model?*

Anytime Prediction assumes that the stimulus processing occurs sequentially in different brain regions and there exists a brain region which has access to the information after each layer of processing. Once there is enough evidence in favor of one alternative after a layer of processing, the decision process stops, and a response is produced. This dynamic is different from the dynamic of standard FNNs used in computer vision. Each unit in a FNN instantaneously transforms its inputs into an output which is in contrast to a feedforward network of biological neurons. When given a static input, biological neurons do not immediately produce their final responses (Ratcliff, Voskuilen, & McKoon, 2018). The movement of electric charges and neurotransmitters, and the opening and closing of ion channels takes time, so the network will gradually transition from its initial to its final state (van Bergen & Kriegeskorte, 2020). Therefore, it is natural to assume that the sources of variability in choice distributions are based on these types of dynamics, which is a key assumption behind the Anytime Prediction models. However, this kind of reasoning has some limitations. First, the trajectory from initial to final state continually becomes perturbed by noise (van Bergen & Kriegeskorte, 2020). The sources of noise can be inherent in the natural world, produced by the signal detection limits of sensory organs themselves, or may reflect the variability of neural responses in the brain at different stages of processing (Ratcliff et al., 2018; Scott, Constantinople, Erlich, Tank, & Brody, 2015). The deterministic nature of Anytime Prediction limits its behavior to capture such noisy dynamics. One important evidence for this limitation comes from the fact that human responses are variable even for the same stimulus (Beck et al., 2012; Renart & Machens,
2014; Wyart & Koechlin, 2016), but we showed that Anytime Prediction never produces inconsistent responses for same image (Error! Reference source not found.).

Second, pure feedforward sweep cannot explain the neural and behavioral data reported in previous experiments. Conduction from one area to another in visual cortex takes approximately 10 milliseconds (Mizuseki, Sirotta, Pastalkova, & Buzsáki, 2009), with signal from photoreceptors reaching inferior temporal cortex at the top of the ventral stream by 70-100 milliseconds (Nayebi et al., 2018). Therefore, a single sweep from input to output in a feedforward network should result in decisions with RT less than few hundred milliseconds. However, human decisions can range from a hundred of milliseconds to even a few seconds. In fact neural dynamics indicating potential recurrent connections take place over the course of 100-200 milliseconds (Issa, Cadieu, & Dicarlo, 2018). Besides, the resolution of RT or accuracy distributions produced by Anytime Prediction directly depends on the number of layers used in the architecture of the model. However, with only limited number of feedforward layers in the brain, humans are able to produce a wide range of RT or accuracy distributions, owing to recurrent connections (Nayebi et al., 2018). It was also shown that human decisions agree best with predictions from intermediate layers of a very deep network (e.g. VGG16 or ResNet) (Eberhardt, Cader, & Serre, 2016). Given that deep network’s output is trained to match human predictions at \( t = \infty \), this can be speculated that maybe ultra-deep networks approximate recurrent computations with a shallower network (Liao & Poggio, 2016). Therefore, a form of recurrency is required to provide a greater and more computational depth (van Bergen & Kriegeskorte, 2020).
4.3.2 Advantages of using RTNet-like structure in modelling human decision-making

The primary goal of RTNet was to produce a model of human decision-making which can account for human behavior and can be generalized to novel images. For this reason, we used a Bayesian neural network to model the internal noise and extend the conventional decision-making models such as race model using state-of-the-art deep neural networks. One advantageous of RTNet over Anytime Prediction is that it works based on accumulation of noisy information over time and hence, can produce noisy trajectories similar to humans (Ratcliff et al., 2018; van Bergen & Kriegeskorte, 2020). The second advantage is that RTNet is not restricted to limited number of computations, and it can provide iterative inference with limited resources. An iterative algorithm is one that employs a computation that improves the initial guess. Applying the computation again to the improved guess yields a further improvement. This process can be repeated until a good solution has been achieved, or until we run out of time or energy. Our brain goes over different iterations of processing by using the recurrent connections.

Even though RTNet does not benefit from any recurrent connections in its structure, it can go over iterative computations by using the accumulation system incorporated in the output of Bayesian network. In fact, the recurrency is embedded in the output where the evidence keeps accumulating until an accumulator reaches a preset criterion. Numerous studies suggested that neurons in lateral intraparietal area (LIP) exhibit the signatures of evidence accumulation (Bahl & Engert, 2019; Hanks, Kiani, & Shadlen, 2014; Huk, Katz, & Yates, 2017; Huk & Shadlen, 2005). Therefore, the evidence accumulation system can be modeled using a recurrent network. Recently, several studies demonstrated advantages of combining a standard feedforward and a recurrent network in performing a range of
tasks (Schwarzschild et al., 2021; Zhou et al., 2022). They showed that these structures are capable of solving sequential reasoning tasks and then extrapolating this knowledge to solve problems of greater complexity than they were trained on (Schwarzschild et al., 2021; Zhou et al., 2022). Hence, due to the methodological constraints regarding training an appropriate RNN for static images (Kietzmann, Spoerer, et al., 2019; Nayebi et al., 2018; Spoerer et al., 2020, 2017; van Bergen & Kriegeskorte, 2020), the decision process can be modeled as a combination of a Bayesian neural network for encoding evidence in favor of response options and a recurrent network which can decode the encoded information with an appropriate stopping criterion. Given the notion that information processing is noisy (Ratcliff et al., 2018) and abundance of recurrent connections in human brain (Douglas, Koch, Mahowald, Martin, & Suarez, 1995; Lamme & Roelfsema, 2000; Supèr, Spekreijse, & Lamme, 2001), this model is more biologically plausible than Anytime Prediction, and in general, the class of dynamic neural networks (Han et al., 2021).

4.4 Limitations of using RTNet-like structure

One limitation with RTNet is that separate evidence accumulation system for each choice option renders its decision-making mechanism non-optimal. This is actually one limitation for all models that consider a separate accumulation system for each response option such as race model. We provide an intuition regarding this non-optimal mechanism in an example. Imagine a situation where the observer has almost same amount of evidence for two available choice options close to the time of the decision. Effectively, in such scenario the observer responds randomly. Previous research showed that guessing can be an appropriate behavior, for example, if the observer knows that the task is very difficult (Malhotra, Leslie, Ludwig, & Bogacz, 2017) or if the observer has been deliberating for a
long time (Drugowitsch, Moreno-Bote, Churchland, Shadlen, & Pouget, 2012). However, in a race model, guessing can happen at any time point regardless of task difficulty. The opposite scenario can also happen. Suppose that the observer has a great deal of evidence in favor of one alternative than the others, and hence, can be almost confident to choose that alternative as the decision. In the race model, the observer will continue to deliberate if the accumulator has not reached its threshold even if there is strong evidence in favor of a specific choice option. It should be noted even though optimal decisions are ideal, it is not clear whether human perceptual decisions are actually optimal. There are plenty of evidence pointing to suboptimal performance in perceptual tasks (Evans, Bennett, & Brown, 2019; Rahnev & Denison, 2018), and hence, this might not necessarily be a drawback for race models when modeling the human perceptual decision-making.

Another limitation of using the current format of RTNet is that each sweep of the feedforward path is independent of the previous states. At each time point, a random sampling of the posterior distributions result in a network which is essentially independent of the resulting network in other time stamps. However, the dynamics in a recurrent network (like human brain) is different and at each time point, the evidence is influenced by the internal representation from previous time point (van Bergen & Kriegeskorte, 2020). To address this limitation in RTNet structure, the sampling process can be modified such that the current state of the network depends on the previous state(s). For example, a weight over an edge at a specific moment can be a function of its previous values, which makes the sequential samples dependent on each other. Additional studies are needed to investigate the effect of such state dependence on model performance improvement.
CHAPTER 5. CONCLUSION

Feedforward neural networks are the dominant models of human visual recognition. However, their decision behavior is remarkably different from human decision-making. Feedforward networks go over same computations when an image is introduced to them, and they always produce a unique RT, response and confidence for same image. In contrast, human decisions are stochastic and highly variable, even for the same image. In this study, we developed a new neural network, RTNet, which can closely approximate all basic features of perceptual decision-making. The assumption behind the model is that the decision process occurs sequentially, by accumulation of noisy evidence through the course of time. RTNet has noisy weights to account for internal noise in human visual system and process the same stimulus multiple times until it has enough evidence in favor an alternative to make the decision. This enables variable RT and stochastic decisions even for the same stimulus which is the signature of human decisions. In addition, the uniquely designed RTNet, exhibits several features of human perceptual decision-making such as speed-accuracy trade-off, right-skewed RT distributions, lower accuracy and confidence for harder decisions, etc. Finally, RTNet data showed significant correlation with data from 60 human subjects on a digit discrimination task. Specifically, RTNet produced RT, accuracy, and confidence for individual images in the task was correlated with the same quantities produced by humans. The RTNet has two main advantages over the existing models of human perceptual decision-making: first, it is the first neural network that exhibits all basic signatures of perceptual decision-making. Second, in contrast to dominant models of decision-making such as race models, it can be generalized to new images.
Future studies can assess the RTNet-like architecture performance more extensively and potentially develop better models which can closely approximate the human decision system.
REFERENCES


tradeoff in macaque area LIP. *ELife, 2014*(3). https://doi.org/10.7554/ELIFE.02260


Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. Proceedings of the National Academy of


Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a


80


