A Task-Optimized Neural Network Model of Decision Confidence

Taylor W. Webb (taylor.w.webb@gmail.com)
University of California, Los Angeles

Kiyofumi Miyoshi
University of California, Los Angeles

Tsz Yan So
The University of Hong Kong

Hakwan Lau
University of California, Los Angeles

Abstract

Our decisions are accompanied by a sense of confidence, a metacognitive assessment of how likely those decisions are to be correct, but the mechanisms that underlie this capacity remain poorly understood. A number of recent behavioral and neural data have suggested that decisions are made in accord with an optimal 'balance-of-evidence' rule, whereas confidence is estimated using a heuristic 'response-congruent-evidence' rule. We developed a deep neural network model optimized to classify images and predict its own likelihood of being correct, and found that this model naturally accounts for some of the key behavioral dissociations between decisions and confidence ratings. Further investigation revealed that neither the 'balance-of-evidence' rule nor the 'response-congruent-evidence' rule fully characterized the strategy that the model learned. We argue instead that the model learns to flexibly approximate the distribution of its training data, and, analogously, that apparently suboptimal features of human confidence ratings may arise from optimization for the statistics of naturalistic settings.

Keywords: confidence; metacognition; deep neural networks

Introduction

When faced with a decision between multiple options, we are capable of estimating the likelihood that our decision will be correct. This capacity for metacognition has been studied in domains ranging from perceptual decisions (Zylberberg, Barttfeld, & Sigman, 2012) to economic choice (De Martino, Fleming, Garrett, & Dolan, 2013), and has been shown to have important consequences, such as deciding whether to gather more information before making a decision (Balsdon, Wyart, & Mamassian, 2020), or deciding how much to wager on the outcome of a decision (Persaud, McLeod, & Cowey, 2007). What computations support this sense of confidence?

One proposal is that confidence ratings are generated by optimally predicting the probability that a decision will be correct (Kiani & Shadlen, 2009; Sanders, Hangya, & Kepecs, 2016). An alternative view is that, whereas decisions are made using a 'balance-of-evidence' (BE) rule that incorporates both evidence for and against the decision (as is optimal under certain detection-theoretic assumptions), confidence ratings are generated using a simpler heuristic strategy that only considers the 'response-congruent-evidence' (RCE). That is, after weighing the evidence and making a decision, only the evidence in favor of the decision that was made is taken into account when generating confidence ratings. This alternative view is supported by a range of both behavioral and neural findings (Koizumi, Maniscalco, & Lau, 2015; Maniscalco, Peters, & Lau, 2016; Peters et al., 2017; Zylberberg et al., 2012), and also has parallels to the phenomenon of 'confirmation bias', in which reasoners tend to overweight evidence in favor of the views they already hold (Nickerson, 1998).

These findings raise the question of *why* confidence would be computed according to an apparently suboptimal heuristic. This is especially puzzling given findings suggesting that decisions themselves are made in accord with the BE rule (Peters et al., 2017), because it suggests that the evidence against one's choice is available in the decision-making process, but simply not incorporated into one's sense of confidence. One proposed answer to this question focuses on the detection-theoretic assumptions underlying the apparent suboptimality of the RCE approach. According to these assumptions, decisions are made on the basis of evidence sampled from distributions with equal variance in both the target dimension (evidence for the correct answer) and non-target dimensions (evidence for incorrect answers). However, when there is greater variance in the target dimension than the non-target dimensions, as is thought to be the case in more naturalistic settings (Green & Swets, 1966), it has been shown that the RCE rule can actually outperform the BE rule in terms of metacognitive sensitivity (the ability to discriminate correct from incorrect decisions). In other words, confidence ratings might be made in accord with an RCE rule because doing so is actually useful in naturalistic settings, even if it is suboptimal in the context of certain controlled laboratory experiments (Miyoshi & Lau, 2020).

In this work, we extend this perspective, by asking whether deep neural networks, optimized to perform the tasks of object recognition and generating confidence ratings directly from high-dimensional naturalistic data, will display similar behavioral phenomena as those taken to support the RCE view. Our key contributions are: 1) we show that previously observed dissociations between decisions and confidence are naturally accounted for when networks are trained over a wide range of conditions, but not when networks are trained directly on the narrow conditions that characterize controlled laboratory experiments, and 2) we show that these dissociations, while consistent with an RCE rule for rating confidence, can also be explained as arising from a more flexible and complex strategy. We argue that this latter account is best understood in terms of optimization for the statistics of naturalistic settings, rather than in terms of simple decision rules.

Methods

To train neural networks to perform both object recognition and generate confidence ratings over a range of conditions, we modified common machine learning image datasets (MNIST and CIFAR10) by manipulating image contrast and incorporating noise. Specifically, during training, we scaled each image by a contrast factor μ (sampled from a uniform distribution) and added gaussian noise with variance σ (also sampled from a uniform distribution) to each pixel. The range of conditions generated by this procedure is illustrated for the MNIST dataset in Figure 1. Our standard training regime involved variability in both μ and σ. These parameters were also modified to create alternative training and test regimes (as described in subsequent sections). For simulations with MNIST and CIFAR10, all training and testing was performed using images from the training and test sets respectively.

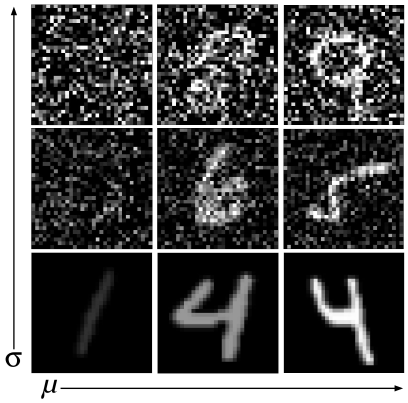


Figure 1: Training regime incorporating a range of contrast (μ) and noise (σ) levels.

We evaluated two separate architectures, but they both conformed to the same general schematic, illustrated in Figure 2. First, each image $\boldsymbol{x}$ was passed through a multilayer encoder[1] f_θ (with parameters θ) to generate a low-dimensional representation $\boldsymbol{z}$[2]. Then, this low-dimensional representation was passed through two separate output layers. One output layer, $g_{\phi_{class}}$ (with parameters ϕ_{class}), was trained to classify the image (i.e. to generate a prediction $\hat{y}$ of the true class y). The other output layer, $g_{\phi_{conf}}$ (with separate parameters ϕ_{conf}), was trained to predict $p\left(\hat{y}=y\right)$, the probability that the classification response was correct (we first evaluated whether the classification response was correct, and then treated this as a binary target for supervised learning). The entire network, including the multilayer encoder function, was trained through backpropagation using the sum of these two loss functions.

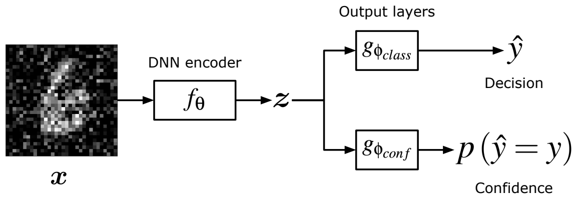


Figure 2: Model diagram. Images ($\boldsymbol{x}$) were passed through a deep neural network (DNN) encoder (f_θ) to generate low-dimensional representations ($\boldsymbol{z}$), which were then passed through two separate output layers ($g_{\phi_{class}}$ and $g_{\phi_{conf}}$) to generate classification responses ($\hat{y}$) and confidence ratings ($p\left(\hat{y}=y\right)$).

To evaluate whether our findings would generalize to a more realistic training signal, we also trained a version of the model with reinforcement learning (RL). Specifically, we used an actor-critic method (Sutton, Barto, et al., 1998) to train networks to discriminate the orientation (left vs. right) of a noisy Gabor patch (with varying contrast and noise levels). The networks had 3 actions available to them: 'left', 'right', and an 'opt-out' action that resulted in a guaranteed, but smaller, reward. The opt-out rate can be used as a proxy for (inverse) confidence, since the model should only select the opt-out action when it is not confident about its response.

We trained 100 separate networks (with different random initializations) for each experiment. We plot the average performance of these networks ± the standard error of the mean, and use the following conventions to indicate statistically significant differences: 'ns', $p > 0.05$; '****', $p < 0.0001$.

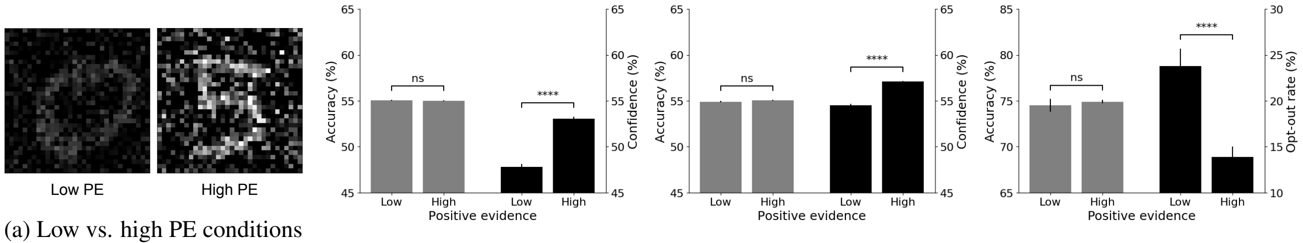
Results

The Positive Evidence Bias

Many studies have shown that human confidence ratings are characterized by a positive evidence (PE) bias (Koizumi et al., 2015; Zylberberg et al., 2012). To elicit this bias, participants are typically presented with two conditions (the 'low PE' and 'high PE' conditions, illustrated in Figure 3a) in which one condition has a greater amount of 'positive evidence', the evidence in favor of making a correct decision (corresponding to the contrast μ in our implementation). Importantly, the signal-to-noise ratio is balanced between these conditions, such that the high PE condition also contains more noise. The PE bias refers to the tendency to be more confident in the high vs. low PE conditions, despite balanced decision accuracy. This bias has been taken as a key piece of evidence in favor of the RCE model.

[1] In our default architecture, f_θ consisted of 3 convolutional layers followed by 3 fully-connected layers. In experiments on CIFAR10 we used an architecture in which f_θ consisted of a 56-layer deep residual network (resnet) based on He et al. (2016).

[2] All experiments used a dimensionality of $d_{\boldsymbol{z}} = 100$.



(a) Low vs. high PE conditions

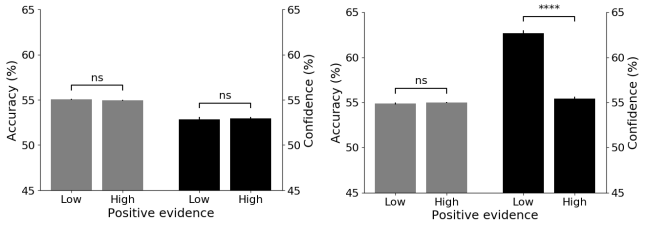
(b) MNIST

(c) CIFAR10

(d) RL

Figure 3: Human confidence ratings display a positive evidence (PE) bias: greater confidence for the high vs. low PE conditions (a) despite balanced signal-to-noise ratio and balanced decision accuracy. This bias naturally emerges in task-optimized neural networks across multiple datasets, architectures, and learning paradigms (b-d).

We first asked whether task-optimized deep neural networks, when trained over a range of contrast and noise levels, also display the PE bias. We did so by identifying two conditions (via grid search) with balanced threshold-level decision accuracy (55% for the 10-class MNIST and CIFAR10 datasets and 75% for the 2-choice tilted gabor RL task[3]), but with differing levels of both contrast (μ) and noise (σ). We found, across multiple architectures, datasets, and learning paradigms (supervised vs. RL), that the PE bias naturally emerged from this framework. Figure 3b shows the presence of the PE bias for models trained on the MNIST handwritten digits dataset. Figure 3c shows the presence of the PE bias for a significantly deeper architecture[4] trained on the CIFAR10 object recognition dataset (involving color images of common objects such as cars, dogs, etc.). Figure 3d shows the presence of the PE bias for networks trained with RL to perform a tilted Gabor discrimination task. The opt-out rate, which is commonly used as a proxy for confidence in experiments with nonhuman animals (Odegaard et al., 2018), is expected to be inversely proportional to confidence, so the observed effect (a lower opt-out rate in the high vs. low PE conditions) is consistent with the PE bias observed for the other tasks.



(a) Trained on fixed μ/σ

(b) Trained on fixed μ

Figure 4: (a) When trained on images with a fixed signal-to-noise ratio, networks do not display a PE bias. (b) When trained on images that vary in noise but not contrast, the PE bias is reversed.

These results confirmed that the PE bias emerged regardless of the dataset, architecture, or learning paradigm used to train networks. This is a noteworthy result, since these networks were optimized only for task performance, rather than being optimized to produce this specific bias. To better understand this phenomenon, we next asked whether the training regime might impact the presence of the PE bias, by training networks on one of two alternative training regimes. First, we trained networks on a regime with a fixed signal-to-noise ratio (similar to training networks directly on the low and high PE conditions). Under these conditions, the optimal approach to rating confidence should not be biased toward positive evidence. Consistent with this, we found that networks trained in this manner did not display a PE bias (Figure 4a). Second, we trained networks on a regime in which the noise level (σ) varied, but contrast (μ) was set to a fixed intermediate value (corresponding to the conditions in the middle column of Figure 1). We found that under these conditions, the opposite effect, higher confidence in the *low* PE condition, emerged (Figure 4b). Under this training regime, accuracy is primarily a function of stimulus noise (since contrast doesn't vary), so it makes sense to adopt a strategy for rating confidence based primarily on the level of sensory noise. Such a strategy results in a reversal of the PE bias, since the high PE condition contains both higher positive evidence and higher noise.

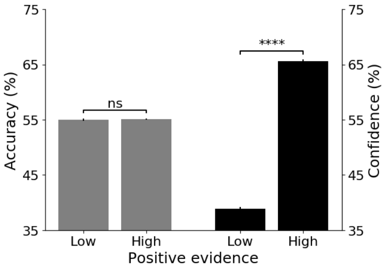
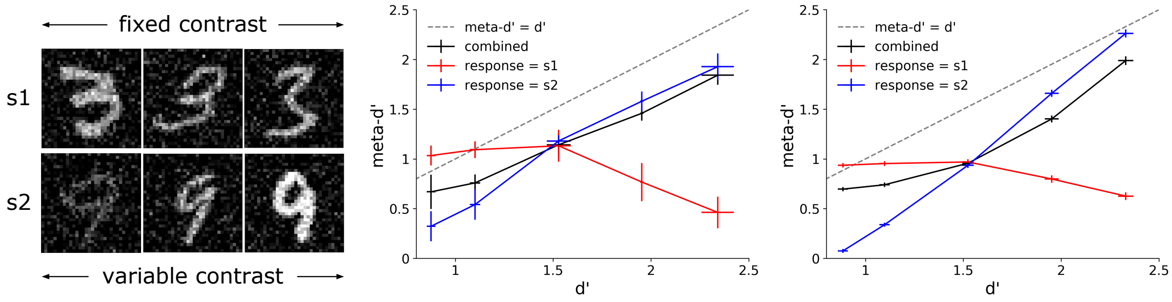


Figure 5: When trained on images that vary in contrast but not noise, the PE bias is significantly expanded.

This suggests that the PE bias emerges in the standard training regime (in which both contrast and noise are varied)

[3] Threshold performance was defined as the midpoint between chance performance (e.g. $1/10$ in the case of 10-class discrimination) and ceiling (100%).

[4] The PE bias was also present on this task for the default architecture, but we evaluated the more sophisticated resnet architecture to test whether the PE bias resulted from an architectural limitation.



(a) Task

(b) Original behavioral results

(c) Model results

Figure 6: The model successfully captures behavioral results from Maniscalco et al. (2016) demonstrating a dissociation between performance and metacognitive sensitivity.

despite the tendency for variable noise to induce a reversed PE bias, suggesting that variable contrast is the key factor driving the emergence of the PE bias. Based on this reasoning, we hypothesized that networks trained on a regime with a fixed noise level, but variable contrast (i.e. the middle row in Figure 1), ought to display an even larger PE bias than in the standard regime. Consistent with this prediction, we observed that networks trained on such a regime displayed a 5-fold increase in the size of the PE bias relative to the standard training regime (Figure 5, note the scale of the Y axis). These results confirmed that the emergence of the PE bias in task-optimized networks is primarily a consequence of whether it is useful within the context of a particular training regime, and in particular whether the training regime contains images with varying contrast. This in turn strongly suggests that the PE bias in human confidence ratings arises as a consequence of optimization for naturalistic settings, in which signal strength will naturally vary.

Dissociation Between Performance and Metacognitive Sensitivity

Maniscalco et al. (2016) tested and confirmed a seemingly paradoxical prediction of the RCE model: human confidence ratings, under certain conditions, are characterized by a pattern of increasing type 1 performance and *decreasing* metacognitive sensitivity (Figure 6b). That is, confidence ratings become less diagnostic of decision accuracy as decision accuracy increases. This finding sharply contradicts the BE decision rule, under which metacognitive sensitivity (as measured by meta-d' (Maniscalco & Lau, 2012)) should always be directly proportional to type 1 performance (as measured by d'). This pattern has therefore been taken as strong evidence in favor of the RCE model.

We tested whether our neural network model would also display this specific signature of human confidence ratings. To do so, we first trained networks on a two-choice variant of the standard training regime, in which each network was trained to discriminate between two digit classes randomly sampled from the ten classes present in the MNIST dataset (e.g. 3 vs. 9). We then tested these networks on stimuli modeled after the task used by Maniscalco et al. (2016) (Figure 6a). In that task, one stimulus class (s1) always appears at an intermediate contrast, whereas the other stimulus class (s2) appears at a range of contrast values, including values below, equal to, and above the contrast of s1. Under these conditions, the RCE model predicts that trials on which participants choose s1 should be characterized by decreasing metacognitive sensitivity as a function of increasing performance, whereas trials on which participants choose s2 should be characterized by the opposite pattern, resulting in the 'crossover' shown in Figure 6b. To test whether our model displayed this same pattern, we performed a grid search to fit five stimulus contrast values (μ) to the observed d' values from Maniscalco et al. (2016), presenting s2 at all five values, and s1 at the intermediate value only. To model the fact that additional noise may accumulate between the time at which participants make a type 1 decision and the time at which they make a confidence rating, we also fit an additional noise parameter and added this to the network's confidence output layer (before applying the final nonlinearity).

Figure 6c shows that the network naturally captured both the crossover effect between meta-d' for trials on which the network chose s1 ($response = s1$) vs. s2 ($response = s2$), and the pattern of decreasing meta-d' as a function of increasing d' for $response = s1$. Thus, the network displayed a characteristic signature of human confidence ratings, thought to be indicative of the use of an RCE heuristic, despite not being optimized to do so[5].

Analysis of Learned Decision Rules

Next we sought to determine whether the confidence rating strategy learned by our model was best characterized by the BE rule, the RCE rule, or by a more complex pattern not completely characterized by either of these rules. To do so, we adapted an approach introduced by Koizumi et al (2015). After training networks on the two-choice variant of the standard

[5] The same qualitative effects (crossover between $response = s1$ and $response = s2$, and negative slope for $response = s1$) were also present when no noise parameter was added to the network's confidence ratings, though the y-intercept for all data was shifted upward.

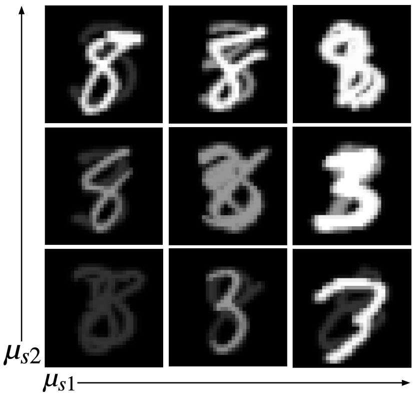
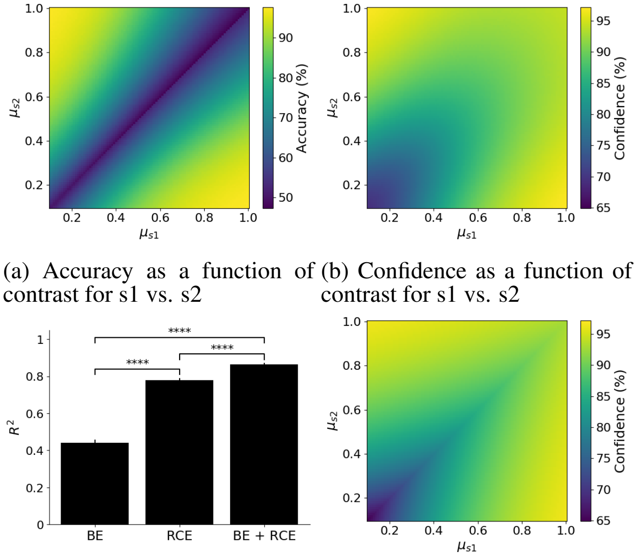


Figure 7: Stimuli used to estimate confidence as a function of stimulus evidence. Each image consists of two superimposed digits (belonging to classes $s1$ and $s2$) with independently varying contrast levels (μ_{s1} and μ_{s2}).

training regime (introduced in the previous section), we presented these networks with images consisting of two superimposed digits, each belonging to one of the two digit classes that the networks were trained on (Figure 7; note that noise was added to the images actually used to evaluate networks, but is not pictured for the sake of illustration). We treated the output of the classification layer as an indication of the class of the digit with a higher contrast. We then measured both accuracy and confidence as a function of the contrast for each digit class (μ_{s1} and μ_{s2}), treating these contrast values as a proxy for the internal evidence for s1 and s2[6].

This revealed that accuracy conformed very closely to an optimal BE rule (Figure 8a), whereas confidence followed a more complex pattern (Figure 8b). Qualitatively, this pattern bears some resemblance to the RCE rule, in that confidence increases as one moves along either the X or Y axes, but there is also some resemblance to the BE rule, in that confidence increases as one moves away from the diagonal.

To better understand this pattern, we fit linear regression models to determine whether (a logit transformation of) confidence ratings were best predicted by the balance of evidence ($|\mu_{s1} - \mu_{s2}|$), the response-congruent evidence (μ_{s1} when $\mu_{s1} > \mu_{s2}$, and μ_{s2} when $\mu_{s2} > \mu_{s1}$), or a multiple regression model incorporating both variables. This revealed that confidence was better predicted by the RCE rule than the BE rule (Figure 8c), explaining the RCE-like behaviors (the PE bias and dissociation between performance and meta-d') displayed by the network. Interestingly though, a regression that combined both the BE and RCE rules predicted confidence better than either of them alone, and, as can be seen from the predictions of this regression (Figure 8d), even the combination of these variables did not fully capture the complex pattern of confidence ratings generated by the neural network.



(a) Accuracy as a function of contrast for s1 vs. s2

(b) Confidence as a function of contrast for s1 vs. s2

(c) Variance captured by regression models

(d) Predictions of multiple regression model

Figure 8: Analysis of decision rules learned by the network.

Discussion

We have presented a deep neural network model of decision confidence, trained directly on naturalistic high-dimensional data, and shown that it can account for some of the key behavioral dissociations observed between decisions and confidence ratings. Moreover, we have shown that it can account for these findings while relying on a strategy that does not conform to either of the simple decision rules previously postulated to underlie decision confidence.

Our work has an important link to the work of Miyoshi and Lau (2020). In that work, it was shown that an RCE rule for rating confidence is advantageous when stimulus distributions are characterized by greater target variance than non-target variance. In our work, we employed a training regime that consisted of images with varying contrast. This likely introduced an additional degree of variance in the target dimension (additional variance in the evidence for the correct stimulus class) not present in the non-target dimension, and therefore may explain why RCE-like behavior emerges in these networks when trained on such a regime. Consistent with this, when we trained networks on images with a fixed contrast level, the PE bias was reversed, whereas, when we trained networks on images that varied only in their contrast, the PE bias was dramatically increased. These results underscore the importance of closely examining the detection-theoretic assumptions on which previous theories have been built, which may be violated under naturalistic settings.

Along these lines, our work has implications for debates over whether human confidence ratings are in some sense optimal (Adler & Ma, 2018; Koizumi et al., 2015; Kiani & Shadlen, 2009; Maniscalco et al., 2016; Peters et al., 2017; Sanders et al., 2016). Previous debates over optimality have

[6] Our reasoning was that, although there will be variation across trials in the internal evidence associated with each contrast level, the contrast level for each digit class should be a reliable predictor for the internal evidence when averaged across trials.

focused on the question of whether confidence ratings are optimal with respect to the specific conditions that characterize laboratory experiments. We suggest that it may be useful to reframe the debate in terms of optimization for more ecologically relevant objectives and task settings, such as the broad range of conditions that characterized our training regime. Our work suggests that confidence ratings may indeed be optimal, or at least optimized, for these settings.

One potential avenue for testing this idea comes from the alternative training regimes that we employed, under which the PE bias was eliminated or even reversed. It may be possible to induce similar effects by training participants extensively on tasks with similar properties (e.g. stimuli that vary only in their noise level, but not contrast). Previous work has found that extensive training and feedback can influence, though not entirely eliminate, such biases (Maniscalco et al., 2016). An important caveat is that it may not be possible to entirely reshape confidence mechanisms within the realistic time frame of a laboratory experiment. The training undergone by our neural network model might reflect optimization over much larger, perhaps even evolutionary, timescales.

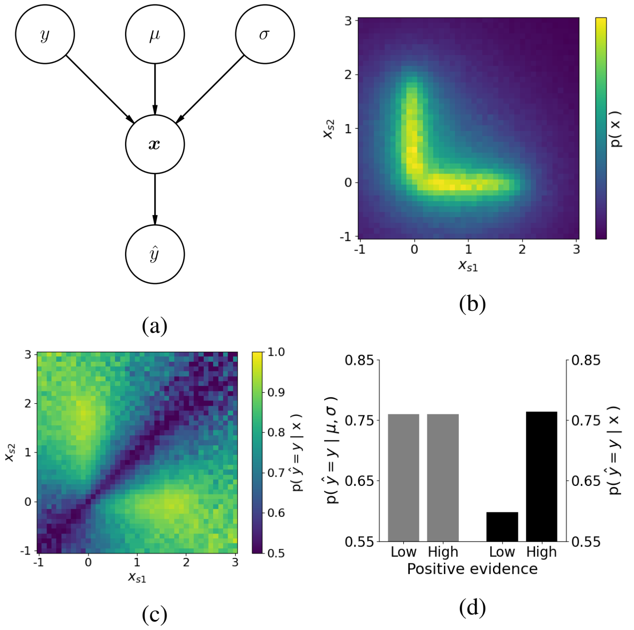


Figure 9: (a) Idealized graphical model. $\boldsymbol{x}$ is sampled from a 2-dimensional Gaussian distribution, with mean in the target dimension (determined by the class y) of μ, mean in the non-target dimension of 0, and variance in both dimensions of σ. μ and σ are themselves sampled from uniform distributions. $\hat{y} = argmax(\boldsymbol{x})$. (b) Distribution of data, collapsing across classes and range of values for μ, σ. (c) Probability of being correct given the data ($\boldsymbol{x}$). (d) Low and high positive evidence conditions were identified by grid search. Probability of being correct given the data is biased toward positive evidence. Probability of being correct given the generative parameters (μ,σ) is unbiased.

We found that our networks adopted a complex strategy for rating confidence. Given that this strategy was not easily explainable by simple decision rules, what is the best way to understand this pattern? In line with recent proposals (Richards et al., 2019), we suggest that this pattern is best understood as an interaction between the objective the network was optimized to perform, the training data used to perform this optimization, and the inductive biases of the network's architecture.

To illustrate this point, consider a simple case in which $\boldsymbol{x}$ is not an image, but is instead a sample from a 2-dimensional Gaussian distribution (Figure 9a). In this example, the mean of the target dimension μ is itself sampled from a uniform distribution, analogous to variable contrast in the image datasets that we used. This results in elongated distributions with greater target variance than non-target variance (Figure 9b). Under these conditions, as shown in Figure 9d (and as shown by Miyoshi and Lau (2020)), the probability of being correct given the data ($p\,(\hat{y} = y|\boldsymbol{x})$) is biased toward positive evidence, while the probability of being correct given the generative parameters that underlie the data ($p\,(\hat{y} = y|\mu,\sigma)$) is not. This simple relationship may partially explain the PE bias in both our neural network model and human confidence ratings, since an observer will generally only have access to the sensory data, not the parameters that generated it.

However, this idealized, low-dimensional model, does not entirely capture the pattern of confidence ratings displayed by our neural network model, or by human participants. In particular, a scenario in which there is strong evidence for both stimulus classes elicits high confidence ratings from the neural network model (upper right quadrant of Figure 8b), even when the relative evidence in favor of the correct answer is weak (close to the diagonal), resulting in low accuracy. This mirrors the pattern observed in human participants (Koizumi et al., 2015), but does not follow from treating confidence as the groundtruth probability of being correct given the data, according to which such a scenario should elicit low confidence (upper right quadrant of Figure 9c).

This dissociation can be explained by the fact that this scenario, in which there is strong evidence for both stimulus classes, is probably never or only rarely encountered during the training regime that we employed. In such a scenario, the behavior of a network is effectively an extrapolation, and will therefore primarily be governed by the network's inductive biases (since there is no training data to constrain the network's behavior in this region of the task space). Thus, the pattern of confidence ratings displayed by the network is most likely an emergent property of interactions between the training objective (to predict $p\,(\hat{y} = y|\boldsymbol{x})$), the distribution of the training data (including variability in signal strength), and the network's inductive biases, which are particularly relevant outside the range of the training data. More work is needed to develop a principled understanding of what drives extrapolative behavior in neural networks (Webb et al., 2020; Xu et al., 2020). Future work should also investigate whether human confidence ratings are best explained by simple decision rules or by a more flexible pattern akin to the behavior of our model.

One remaining question concerns how our model might be implemented in the brain. Though ours is ostensibly a neural network model, it is missing many of the key properties of biological neural networks, including different cell types, temporal dynamics, and biologically realistic learning rules, and therefore may be better understood as occupying a space somewhere between the algorithmic and implementation levels of analysis. Maniscalco et al. (2019) have recently shown how a biologically realistic competing accumulator model that incorporates 'tuned normalization' can capture many of the same dissociations that we've accounted for in the present work. In future work, we hope to explore how our current model might be grounded in biology by incorporating similar mechanisms. At the same time, one strength of our approach is that it allows computations to emerge in a data-driven manner, rather than specifying them by hand, resulting in learned representations and strategies that are best suited to a particular task setting. In future work, we plan to exploit this feature of our model by studying whether the internal representations learned by the network can account for known neural signatures of human confidence ratings (e.g. Peters et al. (2017)).

Acknowledgements

We would like to thank Sivananda Rajananda for his work on an earlier version of this model. We would also like to thank Megan Peters, Matthias Michel, Matan Mazor, and the anonymous reviewers for helpful feedback and discussions. This research was supported by a postdoctoral fellowship from the NIH.

References

Adler, W. T., & Ma, W. J. (2018). Comparing bayesian and non-bayesian accounts of human confidence reports. *PLoS computational biology*, *14*(11), e1006572.

Balsdon, T., Wyart, V., & Mamassian, P. (2020). Confidence controls perceptual evidence accumulation. *Nature communications*, *11*(1), 1–11.

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature neuroscience*, *16*(1), 105–110.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley New York.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *science*, *324*(5928), 759–764.

Koizumi, A., Maniscalco, B., & Lau, H. (2015). Does perceptual confidence facilitate cognitive control? *Attention, Perception, & Psychophysics*, *77*(4), 1295–1306.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and cognition*, *21*(1), 422–430.

Maniscalco, B., Odegaard, B., Grimaldi, P., Cho, S. H., Basso, M. A., Lau, H., & Peters, M. A. (2019). Tuned normalization in perceptual decision-making circuits can explain seemingly suboptimal confidence behavior. *bioRxiv*, 558858.

Maniscalco, B., Peters, M. A., & Lau, H. (2016). Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception, & Psychophysics*, *78*(3), 923–937.

Miyoshi, K., & Lau, H. (2020). A decision-congruent heuristic gives superior metacognitive sensitivity under realistic variance assumptions. *Psychological Review*, *127*(5), 655–671.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*(2), 175–220.

Odegaard, B., Grimaldi, P., Cho, S. H., Peters, M. A., Lau, H., & Basso, M. A. (2018). Superior colliculus neuronal ensemble activity signals optimal rather than subjective confidence. *Proceedings of the National Academy of Sciences*, *115*(7), E1588–E1597.

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature neuroscience*, *10*(2), 257–261.

Peters, M. A., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., ... others (2017). Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature human behaviour*, *1*(7), 0139.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., ... others (2019). A deep learning framework for neuroscience. *Nature neuroscience*, *22*(11), 1761–1770.

Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a statistical computation in the human sense of confidence. *Neuron*, *90*(3), 499–506.

Sutton, R. S., Barto, A. G., et al. (1998). *Introduction to reinforcement learning* (Vol. 135). MIT press Cambridge.

Webb, T., Dulberg, Z., Frankland, S., Petrov, A., O'Reilly, R., & Cohen, J. (2020). Learning representations that support extrapolation. In *International conference on machine learning* (pp. 10136–10146).

Xu, K., Li, J., Zhang, M., Du, S. S., Kawarabayashi, K.-i., & Jegelka, S. (2020). How neural networks extrapolate: From feedforward to graph neural networks. *arXiv preprint arXiv:2009.11848*.

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in integrative neuroscience*, *6*, 79.