Are you sure about that? On the origins of confidence in concept learning

Hrvoje Stojic (h.stojic@ucl.ac.uk)
Max Planck UCL Centre for Computational Psychiatry
and Ageing Research, University College London

Eran Eldar (e.eldar@ucl.ac.uk)
Max Planck UCL Centre for Computational Psychiatry
and Ageing Research, University College London

Hassan Bassam (hassan.bassam.17@ucl.ac.uk)
Max Planck UCL Centre for Computational Psychiatry
and Ageing Research, University College London

Peter Dayan (p.dayan@ucl.ac.uk)
Gatsby Computational Neuroscience Unit
University College London

Raymond J. Dolan (r.dolan@ucl.ac.uk)
Max Planck UCL Centre for Computational Psychiatry
and Ageing Research, University College London

Abstract

Humans possess a rich repertoire of abstract concepts about which they can often judge their confidence. These judgements help guide behaviour, but the mechanisms underlying them are still poorly understood. Here, we examine the evolution of people's sense of confidence as they engage in probabilistic concept learning. Participants learned a continuous function of four continuous features, reporting their predictions and confidence about these predictions. Participants indeed had insight into their uncertainties: confidence was correlated with the accuracy of predictions, increasing as learning progressed. There were substantial individual differences. In contrast to many classical models that try to explain only the predictions, we formalized human function learning in Bayesian terms as Gaussian process inference. This model generates posterior distributions, allowing us to link predictions and confidence judgements. Gaussian process inference well matched participants' predictions, and also the confidence judgements of metacognitively competent participants. Our results show that human confidence judgements during learning are tied to uncertainty, suggesting that concept learning is fundamentally probabilistic.

Keywords: function learning; confidence; Bayesian models

Introduction

Consider the problem of predicting the suitability of restaurants given their location and menus, and one's companions. Or consider predicting future stock prices based on historical prices, the state of the economy, and the companies' competitors. Restaurant goers and stock market traders alike leverage their knowledge of what makes a good restaurant or a promising stock to make such predictions. These are just two of innumerable examples of how our knowledge of sophisticated concepts helps us navigate our lives.

Predictions of these sorts are often accompanied by a sense of confidence. Such feelings can serve an important role in guiding learning and choice, for instance helping us to identify better options. Given two restaurants of equal predicted quality, we might explore the one about which we are less confident, if we expect many future opportunities to benefit from the knowledge gained. Confidence in this case grounds an estimate (often called an "uncertainty bonus" Kakade & Dayan, 2002) of the informativeness of an option (Auer, Cesa-Bianchi, & Fischer, 2002), capturing the value of potential improvement. Confidence can also help us avoid coming dangerously close to irreparable harm. Given two stocks of equal predicted mean performance, we can avoid purchasing the one which we cannot be confident does *not* lead to ruinous outcome. In other words, confidence can help us explore the world safely (Sui, Gotovos, Burdick, & Krause, 2015).

Despite their importance, confidence judgements are only rarely asked in concept learning experiments (e.g. Nosofsky, 1984), and the cognitive processes behind them remain mysterious. Influential models of both category learning (concepts with discrete predicted variables) and function learning (respectively continuous; which we study here), only model predictions and thus cannot easily account for confidence judgements (Hoffmann, von Helversen, & Rieskamp, 2016; McDaniel & Busemeyer, 2005). In contrast, confidence has been extensively studied in perceptual decision making and sensorimotor learning (Pleskac & Busemeyer, 2010; Moran, Teodorescu, & Usher, 2015; Körding & Wolpert, 2004). However, the resulting ideas, which center on the evidence available in external stimuli, do not readily transfer to concept learning, for which the evidence is mostly internal, based on a model relating observable variables to a variable of interest.

Here, participants completed function learning tasks (McDaniel & Busemeyer, 2005; Speekenbrink & Shanks, 2010), providing both predictions and confidence about the predictions. We assessed whether confidence judgements correlated positively with prediction accuracy, as in the perceptual domain (Moran et al., 2015). We hypothesized that confidence judgements arise from a probabilistic learning process, which we modelled as Gaussian process inference (Rasmussen & Williams, 2006). In this, participants estimate a posterior distribution of the model of the environment and assess confidence according to a measure of the width of this posterior. We varied the type of environment *within* participants – including linear and quadratic mappings between item

features and values. This allowed us address the qualitatively different learning strategies implied in past literature: rule- or similarity-based strategies, that are induced by these environments (Hoffmann et al., 2016; Ashby & Maddox, 2011). We also varied *between* participants whether they provided predictions ("Pred only" condition) or both predictions and confidence ratings ("Pred CR"), to examine whether making confidence judgements itself changes the learning process[1].

Methods

We recruited 30 participants (19 male, 11 female; $M_{age} = 38.1$, $SD_{age} = 12.6$) through Amazon's Mechanical Turk (`http:\\mturk.com`). We rewarded participants with a fixed payment ($6 or $9, depending on the condition) and a performance bonus ($4.7 on average). The study was approved by the UCL Research Ethics Committee 9929/003.

Participants had to use four observable features (x_1, x_2, x_3, x_4) to predict values y of collectable items called beetles and sonics, and had to express their confidence about the predictions (Figure 1). Values depended on features through a function that was initially unknown to participants. We considered two such functions (counterbalanced as beetles or sonic), one non-linear (quadratic):

$$\begin{aligned} y_{NL} = 51.5 + 4.1(x_1 - 2.2)^2 - 5.1(x_2 - 2.2)^2 \\ + 2.6(x_3 - 2.2)^2 - 0.5(x_4 - 2.2)^2 \end{aligned} \quad (1)$$

and the other linear:

$$y_L = 45 + 7.5x_1 - 5.5x_2 + x_3 - 0.5x_4 \quad (2)$$

Previous studies suggest that in the linear environment, people tend to learn simple rules relating individual features and function values via a form of hypothesis testing. In the non-linear environment, they tend to adopt a similarity based strategy, predicting values based on how similar the item at hand is to previously experienced items stored in memory (e.g. Hoffmann et al., 2016; Ashby & Maddox, 2011).

Participants first underwent a training phase in which they were given the true function values, y, as feedback. We constructed a set of 40 items (with $x \in 1,2,3,4$); participants experienced them four times in the training phase, for 160 trials in total. Participants then proceeded to a no-feedback test phase involving 30 interpolation and 30 extrapolation items designed to evaluate the knowledge of the function and for model selection. The interpolation items were similar to items experienced in the training phase ($x \in 1,2,3,4$), but in combinations not present in the training set. In contrast, the 30 extrapolation items also included the feature values 0 and 5 that were absent in the training set.

In "Pred only" condition, we incentivised learning by relating experiment earnings per trial, R_t^{pred}, to prediction accuracy: $R_t^{\text{pred}} = \max(100 - (\hat{y}_t - y_t)^2, 0)$. Participants earned more

[1] In two additional conditions we tested two more methods for eliciting confidence judgements. Results are qualitatively similar to the ones reported for "Pred CR" condition.

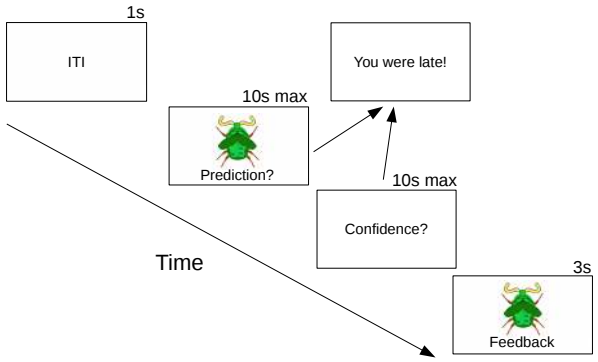

Figure 1: In each trial either a cartoon beetle (illustrated here) or a sonic was displayed, with four features that varied from trial to trial – for example, the size of legs, wings, antennae and the number of spots for beetles. After a 1s inter trial interval, according to the condition, we gave the participant 10s to predict the value of the beetle using a slider ranging from -50 to 150, and then 10s to provide a confidence judgement about that prediction using a slider ranging from 1 ("Completely uncertain") to 10 ("Completely certain"). Only in the training phase, we then presented the true function value for 3s. If participants failed to respond in time, they lost 100 points and had to repeat the trial.

experimental points the smaller the difference between their predictions $\hat{y}$ and the true values y. In "Pred CR" condition, we incentivised confidence ratings in each trial, c_t, by using them to weight the prediction earnings in each trial, R_t^{pred}, in computing the final earnings

$$R = \sum_{t=1}^{T} \frac{c_t}{\sum_{t'=1}^{T} c_{t'}} R_t^{\text{pred}}. \quad (3)$$

Participants were thus encouraged to assign high rating to predictions of whose accuracy they were confident, and a low rating when less confident.

Results

Participants' prediction accuracy improved during the training phase. The root mean square error (RMSE) between predicted and the real values decreased (one-tailed $t_{31.6} = 8.65$, $p < 0.001$) in "Pred only" condition from the first block ($M = 14.1$, $SD = 2.5$) to the fourth ($M = 8.5$, $SD = 1.5$). There was a similar decrease ($t_{11.5} = 4.86$, $p < 0.001$) in "Pred CR" condition from the first block ($M = 18.2$, $SD = 4.9$) to the fourth ($M = 10.7$, $SD = 1.3$) (Figure 2, left panel). Ex-ante we expected similar performance in both conditions, with perhaps slightly better performance in "Pred CR" since confidence judgements required participants to think more extensively about their predictions. Surprisingly, this did not help. In the fourth block RMSE was worse than for the "Pred only" group (two-tailed $t_{23.2} = 4.30$, $p < 0.001$). Identifying the source of this difference is left for future research.

We found no significant difference in performance in the fourth block between the linear and nonlinear environments. As expected, performance was worse in the test phase, where participants encountered items they had not previously seen; the deterioation was much greater (one-tailed $t_{58.8} = 10.60$,

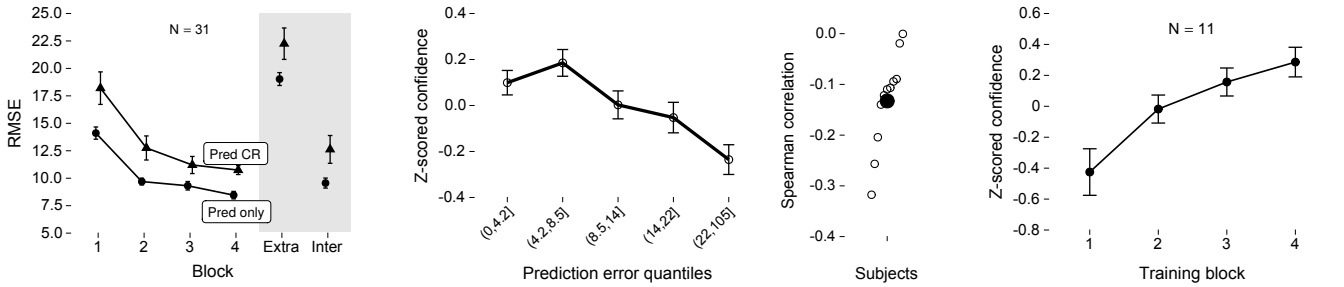

Figure 2: **Left**: People learned over time, with root mean square error (RMSE) between predictions and true values decreasing across the blocks in the training phase. Test phase performance is in the shaded area. **Middle**: Participants reported higher confidence for small prediction errors and lower confidence for larger errors in the test phase. There were also substantial individual differences in metacognitive performance. **Right**: Participants became more confident over time during the training, mirroring the improvement in knowledge of the function.

$p < 0.001$) for extrapolation ($M = 20.2$, $SD = 3.8$) than interpolation trials ($M = 10.6$, $SD = 3.3$).

Confidence is well founded

Confidence judgements are widely considered in perceptual domains because of they correlate positively with decision accuracy (Moran et al., 2015). In the test trials there was an appropriate relationship between participants' confidence and prediction errors (Figure 2, middle panel). A mixed linear model, where we regressed absolute prediction errors on z-transformed confidence ratings with participant random effects on intercepts and slopes, showed a significant effect of -0.012 ($SE = 0.003$, $t = -3.89$, $p = 0.005$). However, individuals differed significantly: from those with rather good metacognitive abilities (correlation of -0.32) to those who were metacognitively challenged (correlation of 0). Individual variability in metacognitive abilities has previously been observed in the perceptual domain (Fleming & Dolan, 2012), albeit not to this degree.

We expected learning effects for confidence judgements to parallel those observed for predictions. Indeed, confidence ratings increased (one-tailed $t_{16.95} = 3.99$, $p < 0.001$) as participants' knowledge of the underlying functions improved, from a mean z-scored confidence rating of -0.42 ($SD = 0.50$) in the first block to a mean rating of 0.29 ($SD = 0.32$) in the fourth block (Figure 2, right panel).

Finally, we predicted that in the test phase, participants' knowledge of the function should be more accurate for interpolation trials, leading to greater confidence about the interpolation than extrapolation items. Indeed, participants were more confident about interpolation items (paired one-tailed $t_{10} = 5.20$, $p < 0.001$), with mean of the differences in z-scored confidence ratings of 0.40 ($SD = 0.32$). This result parallels the decrease in confidence with increase in task difficulty that has been widely observed in perceptual decisions (Pleskac & Busemeyer, 2010; Moran et al., 2015).

Probabilistic learning as a basis of confidence

We modelled function learning using Gaussian process regression (Rasmussen & Williams, 2006; Schulz, Tenenbaum, Duvenaud, Speekenbrink, & Gershman, 2017; Stojic, 2016). This encompasses Bayesian implementations of classical rule- and similarity-based models, allowing us to test the hypothesis that *probabilistic* function learning underlies people's confidence in their knowledge of concepts.

Gaussian process regression assumes outputs y are noisily generated from a function f, $y = f(\mathbf{x}) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon)$. A Gaussian process ($\mathcal{GP}$) defines a distribution $P(f)$ over the functions $f(\mathrm{x})$, parametrized by a mean function $m(\mathrm{x})$ and a kernel (or covariance) function $k(\mathrm{x}, \mathrm{x}')$. The kernel encodes assumptions about the underlying function. By defining a kernel function we can model versions of both similarity- and rule-based learning (Lucas, Griffiths, Williams, & Kalish, 2015). For the former, we use the squared exponential kernel:

$$k(\mathrm{x}, \mathrm{x}') = \sigma_f^2 \exp\left(-\frac{||\mathrm{x} - \mathrm{x}'||^2}{2\lambda^2}\right) \tag{4}$$

For example, for smaller values of λ, the correlation between $f(\mathrm{x})$ and $f(\mathrm{x}')$ is strong when x and x' are very close, but decreases rapidly with distance. A $\mathcal{GP}$ with this kernel (GP-SE) can thus be viewed as a Bayesian implementation of similarity-based learning, producing similar y values for stimuli with similar x values (Nosofsky, 1984).

For rule-based learning, we consider the linear kernel:

$$k(\mathrm{x}, \mathrm{x}') = \sigma_f^2 (\mathrm{x} - c)^\intercal (\mathrm{x}' - c) + \sigma_b^2 \tag{5}$$

A $\mathcal{GP}$ with a linear kernel (GP-LIN) is equivalent to a Bayesian linear regression model (Rasmussen & Williams, 2006); this allows a Bayesian implementation of rule-based learning to be formalized within the same framework (Lucas et al., 2015). Although we could model complex rules with GP-LIN if interaction terms and additional feature transformations were included, here we only capture simple rules, relating features independently to values. This follows arguments from the concept learning literature according to which rule learning is constrained by what can be easily expressed verbally and fits in the working memory (Ashby & Maddox, 2011).

$\mathcal{GP}$ models are probabilistic models, generating Gaussian posterior distributions for their predictions; these can be summarized with mean and variance parameters. The means are the model-based predicted values that can be used to fit participants' predictions in the training phase. We compared the fits of GP-LIN and GP-SE and a guessing model, according

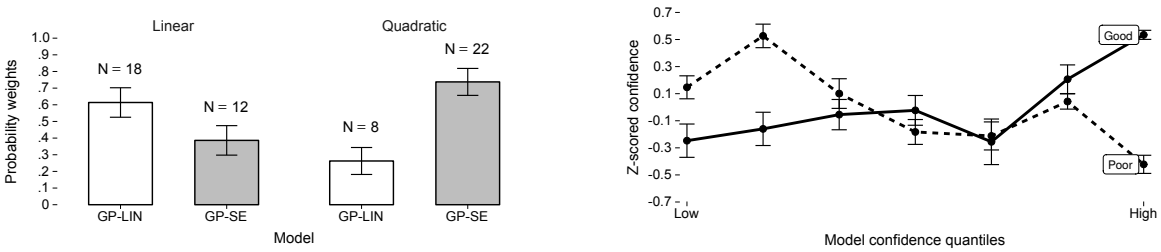

Figure 3: **Left**: Average probabilities that each model is the best; the numbers count the participants best predicted with a model. GP-LIN and GP-SE predict best in the linear and non-linear environments respectively. **Right**: Confidence of participants with good metacognitive abilities roughly followed model-based confidence measure ($-log(\sigma)$); this is not true for the meatacognitively challenged.

to which participants predict using the mean of experienced true values. Using probability weights to select the best model (Wagenmakers & Farrell, 2004) on the test phase predictions, we found that the linear environment behaviour was best predicted by GP-LIN, and the quadratic environment by GP-SE (Figure 3, left panel); no participant was best predicted with the guessing model (so it is not shown). This finding replicates previous results using non-probabilistic function learning models (Hoffmann et al., 2016).

Given the learning model that best predicted participants' responses, we then examined whether participants' confidence correlated with model-based confidence. The variances (inverse precisions) of the $\mathcal{GP}$ posterior are model-based confidence measures that can be compared with participants' confidence. We regressed the negative log precision on confidence ratings, separately for interpolation and extrapolation trials, with random effects on intercepts and slopes. The relationship between the $\mathcal{GP}$-based and human confidence was positive as predicted, but not significantly so (Table 1). This is partly due to the modest number of participants in "Pred CR" condition ($N = 11$), and partly due to individual variability in metacognitive ability (Figure 2, middle panel). Focusing on participants with good metacognition, the relationship was positive and strong, while for the others, it was even negative (Figure 3, right panel). Correlations between model-based and human confidence ranged from -0.82 to 0.89 in interpolation, and from -0.40 to 0.65 in extrapolation trials.

Table 1: Mixed linear regression results. Standard errors of the coefficients are reported in the brackets.

Coefficient	Inter	Extra
Intercept	7.77 (2.52)	6.88 (2.25)
Model confidence	0.66 (0.72)	0.60 (0.61)

Conclusion

In this study we examined the origin of confidence judgements in human function learning. Although on average participants had good metacognition, as evidenced by positive correlations between their confidence and prediction accuracy, there were substantial individual differences. More importantly, leveraging Bayesian modeling of function learning, we showed that the confidence judgements of metacognitively competent participants reflected a probabilistic learning process.

Acknowledgments

We thank Elliott Wimmer and Rani Moran for helpful discussions. HS, EE and RD are funded by the Wellcome. PD is funded by the Gatsby Charitable Foundation. He is on a leave of absence at Uber Technologies; neither organization played any role in this study.

References

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*, 147–161.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, *47*, 235–256.

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *367*, 1338–1349.

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2016). Similar task features shape judgment and categorization processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 1193–1217.

Kakade, S., & Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Networks*, *15*, 549–559.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244–247.

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*, 1193–1215.

McDaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: comparison of rule-based and associative-based models. *Psychonomic Bulletin & Review*, *12*, 24–42.

Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology*, *78*, 99–147.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 104–114.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*, 864–901.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2017). Compositional inductive biases in function learning. *Cognitive psychology*, *99*, 44–79.

Speekenbrink, M., & Shanks, D. R. (2010). Learning in a changing environment. *Journal of Experimental Psychology: General*, *139*, 266–298.

Stojic, H. (2016). *Strategy selection and function learning in decision making*. Unpublished doctoral dissertation, Universitat Pompeu Fabra.

Sui, Y., Gotovos, A., Burdick, J., & Krause, A. (2015). Safe exploration for optimization with gaussian processes. In *International conference on machine learning* (pp. 997–1005).

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*, 192–196.