Psychonomic Bulletin & Review
https://doi.org/10.3758/s13423-023-02261-x

BRIEF REPORT

Using serial dependence to predict confidence across observers and cognitive domains

Ning Mei[1] · **Dobromir Rahnev**[2] · **David Soto**[1,3]

Accepted: 22 February 2023

Abstract

Our perceptual system appears hardwired to exploit regularities of input features across space and time in seemingly stable environments. This can lead to serial dependence effects whereby recent perceptual representations bias current perception. Serial dependence has also been demonstrated for more abstract representations, such as perceptual confidence. Here, we ask whether temporal patterns in the generation of confidence judgments across trials generalize across observers and different cognitive domains. Data from the Confidence Database across perceptual, memory, and cognitive paradigms was reanalyzed. Machine learning classifiers were used to predict the confidence on the current trial based on the history of confidence judgments on the previous trials. Cross-observer and cross-domain decoding results showed that a model trained to predict confidence in the perceptual domain generalized across observers to predict confidence across the different cognitive domains. The recent history of confidence was the most critical factor. The history of accuracy or Type 1 reaction time alone, or in combination with confidence, did not improve the prediction of the current confidence. We also observed that confidence predictions generalized across correct and incorrect trials, indicating that serial dependence effects in confidence generation are uncoupled to metacognition (i.e., how we evaluate the precision of our own behavior). We discuss the ramifications of these findings for the ongoing debate on domain-generality versus domain-specificity of metacognition.

Keywords Confidence · Perception · Memory · Machine learning · Metacognition · Cognition

Introduction

Perceptual judgments in the face of uncertainty are accompanied by metacognitive evaluations—a sense of confidence that tracks the probability of one's decision being correct. It remains, however, unclear how confidence is generated on a moment-to-moment basis.

Our perceptual systems appear hardwired to exploit the auto-correlation in perceptual input across space and time, leading to serial dependence effects in perception, such that recent prior perceptual representations bias current perception (Fischer & Whitney, 2014). Serial dependence in perception reflects the influence that prior perceptual representations have on ongoing perception and is a widespread phenomenon demonstrated for low-level features such as oriented gratings (Fischer & Whitney, 2014) to more complex objects such as faces (Liberman et al., 2014) or ensemble visual representations (Manassi et al., 2017). It has also been argued that a mechanism based on serial dependence can account for different experimental results reported across perceptual, memory, and attention tasks (Kiyonaga et al., 2017).

Recent research also indicates the existence of serial dependence for perceptual confidence (Rahnev et al., 2015; see also Mueller & Weidemann, 2008). Serial dependence in confidence judgments across two different visual tasks has been demonstrated even when confidence is not always rated (Aguilar-Lleyda et al., 2021). The current study focuses on understanding the role of serial dependence in confidence generation beyond perception, across multiple cognitive domains. Here, we test the hypothesis that serial dependence

✉ Ning Mei
n.mei@bcbl.eu

✉ David Soto
d.soto@bcbl.eu

[1] Basque Center on Cognition, Brain, and Language, San Sebastian, Spain

[2] School of Psychology, Georgia Institute of Technology, Atlanta, GA, USA

[3] Ikerbasque, Basque Foundation for Science, Bilbao, Spain

Published online: 07 March 2023

effects in confidence generation in perceptual tasks generalize across observers and cognitive domains.

Here, we leverage the power of machine learning classifiers and out-of-sample predictions to investigate how predictable confidence is, whether the moment-to-moment predictability of confidence is task/domain specific, or whether the patterns of serial dependence during confidence generation in the perceptual domain generalize to predict confidence judgments made by different observers across different cognitive domains. Machine learning classifiers were trained using confidence measure estimates from previous trials to predict the level of confidence in the current trial. The classifier was trained in one particular domain (i.e., perception) and then tested across different cognitive domains (i.e., memory), and generalization performance was quantified.

We predicted that if the dynamic representation of metacognitive confidence across trials is shared across different cognitive domains, then the classifiers ought to generalize and hence predict the representation of confidence in different observers and domains. Notably, the question addressed here is impossible to address using traditional statistical approaches, in which all data are fit at once in the same statistical model. This is one of the advantages of using machine learning to study generalization and out-of-sample predictions regarding human decisions, including confidence ratings. For instance, a recent study showed evidence of serial dependence across domains in a paradigm in which recognition judgments were interleaved with perceptual judgments so that confidence in the perceptual judgment influenced subsequent memory confidence in the next trial (Kantner et al., 2019). However, this approach, in which all data are fitted at once within the statistical model, does not enable one to quantify and predict confidence in new examples—namely, across different observers and across cognitive domains.

The distribution of confidence judgments in visual tasks is highly stable within a particular observer when performance is measured across different days in the same visual task and also across visual tasks of similar structure (Ais et al., 2016). Such findings suggest that serial dependence may also be stable within individual observers when similar visual perceptual tasks are considered but raise questions on whether it is also similar across observers. Here, we addressed whether the patterns of confidence across trials associated with serial dependence effects were similar across observers, by training and testing the classifiers in different observers. We also tested whether these confidence patterns are generalizable across correct and incorrect response trials (see Methods). A recent meta-analysis by Rouault et al. (2018) found no behavioral evidence for the association of interindividual measures of metacognitive performance (i.e., how confidence tracks accuracy) across perception and memory domains. This observation predicts that the pattern of moment-to-moment estimation of confidence might be different and not generalizable across cognitive domains.

Methods

This study reanalyzed data from the Confidence Database (Rahnev et al., 2020). We selected all datasets that employed a 4-point confidence scale because this scale was most common in the perceptual domain, thereby providing a rich dataset to train the classifier and assess the degree of generalization across other domains (i.e., cognitive, memory, and mixed). This resulted in a total of 32 datasets: 16 from the perceptual domain, four from the cognitive domain, six from the memory domain, and six from a mixed domain (where the data came from tasks from multiple domains). Details about each dataset—including the name of the dataset, number of subjects, and total number of trials—are included in Supplementary Table S1. In total, there were 3,077 subjects and 925,091 trials across all 32 datasets. The data were downloaded in April 2020 and may not include datasets added later to the Confidence Database.

Machine learning models

We used machine learning models: a random forest (RF) classifier and a linear vector machine (SVM) as a baseline model (see Table 1).

We chose the linear SVM due to its simplicity and its ability to perform a large number of inferences during prediction; thus, the linear SVM served as a baseline classification model. We chose the random forest classifier to perform the four-classes classification task because the random forest classifier allowed us to compute the feature importance and make interpretable relationships between the features and the target. This was the same strategy as in our prior study (Mei et al., 2020).

An RF classifier is an ensemble of multiple, simpler classifiers that are trained on a subset of the training data, and then the final decision is made by a voting algorithm of these classifiers (Breiman, 2001). The RF classifier was implemented using the scikit-learn Python library (Version 0.24), which included some modifications of implementations of Breiman (2001) to avoid overfitting. We used an RF classifier that contained 500 decision-tree classifiers and the entropy loss objective function. The rest of the hyperparameters were the default. The output probabilities from the RF model were normalized within each ensemble tree-based model. The predicted class probabilities of an input sample were computed as the mean predicted class probabilities of the tree-based model in the forest. The class probability of a single tree is the fraction of samples of the same class in a leaf. The predicted class probabilities of an input sample

Table 1 Summary of relevant properties of each the RF and SVM classifier

Model name	Random forest classifier	Linear SVM
Advantages	1. An ensemble model 2. One of the best models for multiclass classification 3. Interpretation of the feature importance is straightforward 4. Overfitting can be regularized by increasing the number ensemble tree-based models	1. A simple linear model 2. Powerful for scaling up the inference of a trained model 3. Interpretation of the feature weights is simple and directional 4. Overfitting can be regularized by L2 penalty
Disadvantages	1. It is difficult to control for the complexity of the ensemble treed-based models due to the large number of hyperparameters, such as the depth, resample sizes for training the individual tree-based models 2. Training and inference time is much longer than SVM for big datasets	1. Not suitable for multiclass classification 2. Feature weights are distributed in multiple trained models, making it difficult to integrate

within the individual tree were summed up to one. This was implemented in the scikit-learn generic "predict_proba" function.

We used the SVM classifier as a baseline model. SVM is often used for binary classification tasks. In the current study, we used the Scikit-learn implemented SVM with the liblinear kernel (Fan et al., 2008; Pedregosa et al., 2011). L2 regularization was also added to control for overfitting. L2 regularization restricts the sum of the parameters in a regression model so that the model is forced to use simpler functions to model the data. This leads to the model having lower variance and makes it less likely to overfit (Hoerl & Kennard, 1970). Because the classification task in the current study is multiclass, the strategy to handle the multiclass support was a one-vs.-rest scheme. In short, the data were reformulated into one of the four classes being class "1" and the rest of the classes being class "0." Thus, the SVM classifier learned four one-vs.-rest binarized discriminative patterns. During the testing phase, for a given matrix of one-hot coded confidence, the SVM classifier predicted the probability of being one of the four classes, giving a vector of four values. Before comparing to the one-hot coded true label, the predicted probabilities were passed to a softmax function to make the sum of the four values equal to one.

Cross-validation procedure

For within-perceptual-domain classifications, the classifier was trained on data from all subjects but one in a particular study, and the classifier was tested on the remaining subject. For across-domain generalization, the classifier was trained on one of the studies in the perceptual domain and then tested on each subject of each study of a different domain. The data was split into "correct" and "incorrect" examples; specifically, "correct" and "incorrect" here referred to whether the target trial (in which confidence was predicted) was correct or not, but the training examples (based on the previous seven trials) could contain correct and incorrect trials. Accordingly, we ran the classification analysis four times: (1) training the classifier with the instances in which the current (to be predicted) trial was correct and testing on similar left-out instances; (2) training the classifier with the instances in which the current trial was incorrect and testing on similar left-out instances; (3) training the classifier with the instances in which the current trial was correct and testing on the instances in which the current was incorrect; and (4) training the classifier with the instances in which the current trial was incorrect and testing on the instances in which the current was correct.

To measure the classification performance, the classifiers predicted the probabilities of the levels of confidence for the test set. Each column of the prediction matrix corresponded to each level of confidence. Thus, each column of the prediction matrix was compared against each column of the one-hot coded label matrix using the area under the receiver operating curve (ROC AUC). The average of the four ROC AUC measures represented the cross-validation performance of a given fold of cross-validation. The range of ROC AUC score is between 0 and 1, where 0.5 is referred to as the theoretical chance level and 1 is the perfect accuracy, while values below 0.5 mean that performance is worse than guessing.

Feature importance

In order to measure how the confidence ratings of the previous trials contribute to classifying the confidence rating of the current trial, we looked into the RF classifiers, computing feature importances using permutation tests (Altmann et al., 2010). For each fold of cross-validation, after an RF classifier was trained on the training data, a permutation test algorithm was applied to both the RF classifier and the testing data. During the permutation, for the testing data, the order of the trials of confidence ratings of one of the previous trials was shuffled while the rest remained unchanged. Predictions of the testing data were made using the shuffled data. A new ROC AUC score was computed by comparing the true labels and the predictions. The difference between

the new ROC AUC score and the true ROC AUC score represented the contribution of the confidence ratings of the particular previous trial that were shuffled. A positive difference meant that the feature was good for the classification, while zero meant that the feature was not important. Particularly, compared with the early proposed feature importance extraction algorithm (Breiman, 2001), the permutation feature importance algorithm could return negative feature importance estimates. Negative feature importance meant the inclusion of the feature would make the classification less effective.

Second-level statistics

After cross-validation of the classifiers for within- and cross-domain predictions, we performed second-level statistics on the results. For the classification results within the perceptual domain, we averaged the ROC AUC scores over the cross-validation folds for each study before the second-level statistics. For the cross-domain classification results, the classifier was trained in the perceptual domain, and then we averaged the ROC AUC scores within each testing fold for each domain (i.e., cognitive, memory, and mixed). Therefore, each average of decoding performance contained 16 independent ROC AUC scores, one per study of the perceptual domain of the Confidence Database. Classification performance within this domain, and for each level of correctness, if applicable, was assessed by comparing the ROC AUC scores of each study to 0.5 (theoretical chance level for the ROC AUC metric) using a one-sample permutation test as implemented in the EnvStats R library (https://cran.r-project.org/web/packages/EnvStats/index.html) but using custom Python scripts.

Initially, we computed the average of the ROC AUC. Then we subtracted the average from each experimental ROC AUC score and added 0.5 to get a vector that had a mean of 0.5 and the same variance as the original vector. We drew 16 observations with replacement from this vector, hence creating 16 "fake" data points, which were then averaged in order to get an estimate of the mean of the population when the ROC AUC was assumed to be 0.5. We repeated these steps (sampling and average) 10,000 times to estimate the chance-level distribution (i.e., the distribution of the null hypothesis). The probability of the chance level being greater or equal to the experimental score was used as the p value to determine the significance of the experimental score compared with 0.5. If the p value was lower than the critical level, we determined the ROC AUC scores were significantly greater than 0.5. For cross-domain classification and each level of correctness, we performed the same statistical analysis to the ROC AUC scores. For within-perceptual domain decoding, the p values of the permutation tests were corrected by the Bonferroni procedure in total, while for cross-domain decoding, the p values were corrected by the Bonferroni procedure for each of the domains.

For the analysis of the feature importance estimates of the random forest classifier, we performed permutation tests similar to those described above to compare the feature importance estimates against zero. Given a set of feature importance estimates, its mean was subtracted from the estimates, hence creating "fake" estimates centered at zero. We then drew samples with replacement from the "fake" estimates and we saved the average of the samples. This was repeated 10,000 times to get the chance-level distribution. The probability of the chance level being greater or equal to the average of the feature importance estimates was used as the p value to determine the significance of the experimental score compared with zero. For the within-perceptual domain decoding analyses, the p values of the permutation tests were corrected by the Bonferroni procedure. For the cross-domain decoding analyses, the p values were Bonferroni corrected on each of the domains.

In order to analyze the linear trend of the feature importance values as a function of the trial indices, we fitted linear regression models using the trial indices as the independent variable and the feature importance values as the dependent variable. The regression models were cross-validated with a 20-fold cross-validation procedure to estimate the performance of the regression. Then, we performed permutation tests to assess the statistical significance of the regression models. We shuffled the correspondence between the feature importance values and the trial indices and then we fitted a new regression model and cross-validated the model using the same 20-fold cross-validation procedure. We repeated this for 1,000 times to estimate the empirical chance level of the regression model. The probability of the chance level being greater or equal to the average of the original regression performance was used as the p value to determine the significance of the regression coefficients compared with zero.

Past history versus recent history

To further investigate how the recency of confidence ratings in the previous trials influenced the prediction of confidence rating in the current trial, we split the confidence ratings in the previous trial into "past" and "recent," where "past" included confidence ratings in trials of T-5, T-6, and T-7, while "recent" included confidence ratings in trials of T-1, T-2, and T-3. Confidence rating in trial T-4 was not used in this analysis to equate the number of features. We applied the same feature and label preparation, machine learning models, and cross-validation procedures to decode the confidence rating in the current trial using either the confidence on "past" or "recent" trials. After

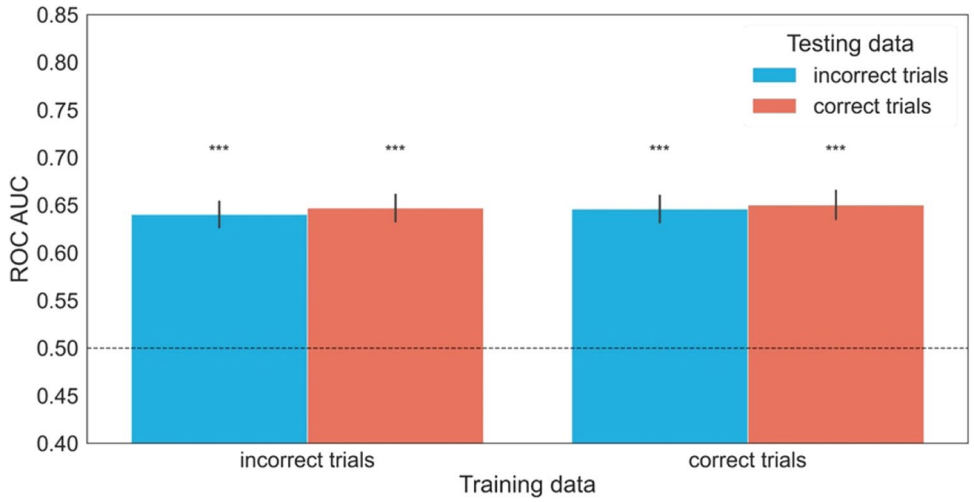

Fig. 1 Decoding confidence levels within the perceptual domain. The random forest classifiers were trained and tested within each study dataset using a leave-one-subject-out cross-validation procedure. The decoding scores were averaged across the cross-validation folds for each study and then the distribution of these averaged scores is plotted. The prediction of confidence level in the current trial based on the previous confidence levels was clearly above chance levels. Statistical significance was estimated by means of resampled one-sample tests against 0.5. Error bars show the standard errors across cross-validation folds. The estimated significant levels were corrected by Bonferroni correction procedure, n.s. = not significant. *$p < .05$. **$p < .01$. ***$p < .001$

cross-validation of the decoding, we applied the same second-level statistics to the ROC AUC results, comparing the ROC AUC against 0.5.

Results

We analyzed studies on metacognition across four different domains from the Confidence Database. We investigated the generation of confidence judgments by training machine learning classifiers using measures of confidence from previous trials to predict the level of confidence in the current trial. We quantified the extent to which the representation of confidence generalizes across different observers across different cognitive domains. Two classifiers were used: a random forest (RF) classifier and a linear vector machine (SVM) as a baseline model. Testing the two classifiers allowed us to test the robustness of the predictions across the different models.

Initially, each classifier was trained in the perceptual domain and then tested within the same domain using a leave-one-subject-out cross-validation procedure for each study (see Methods). Importantly, the cross-validation procedure was performed separately for correct and incorrect trials when decoding the confidence level.

Subsequently, the classifier trained in the perceptual domain was tested across domains (i.e., cognitive, memory, and mixed domains from the Confidence Database; Rahnev et al., 2020).

Classification of confidence within the perceptual domain

We found that the level of confidence in the current trial could be predicted based on the history of previous trials. Figure 1 illustrates that classifier predictions were successful regardless of whether the current trial was correct or incorrect. Averaging across each of the subplots in Fig. 1, the decoding scores were 0.64 ± 0.048 ($M \pm SD$) for SVM and 0.64 ± 0.003 for RF (detailed statistics are shown in Supplementary Table S2). All the cross-validation with the split of data into correct and incorrect examples were significantly greater than 0.5 ($p < .001$, corrected by Bonferroni procedure for multiple comparisons, as indicated in each of the figures). The results suggested that both linear and nonlinear models could learn the patterns of history of confidence ratings to predict the confidence rating of the current trial. Similar results were observed using the SVM classifier (Fig. S1).

We hypothesized that trials that were closer to the current trial may be more important for the prediction of confidence in the current trial than those that were further away from the current trial. To test this hypothesis, we conducted the same decoding analyses based on recent trials (-1, -2, -3 trials back) and based on past trials (-5, -6, -7 trials back). The results showed that confidence in the current trial could be predictive from both recent and also past trials; however, decoding from the most recent trials was better (see Fig. 2). Similar results were observed using the SVM classifier (Fig. S2).

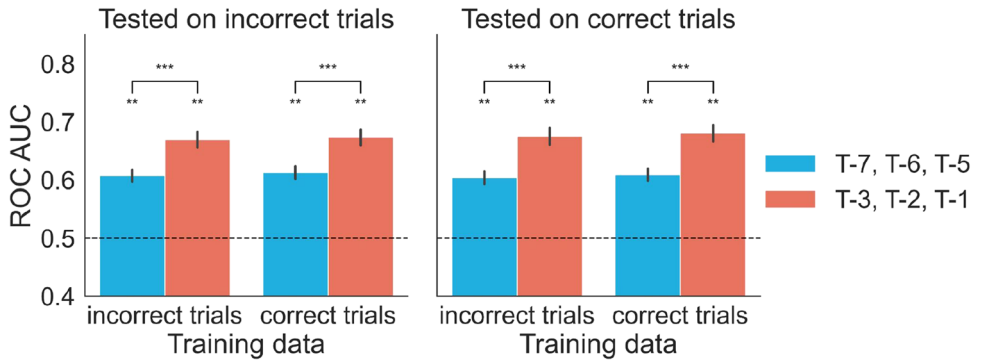

Fig. 2 Decoding of confidence within the perceptual domain based on recent versus past trials back. The random forest classifiers were trained and tested within each study dataset using a leave-one-subject-out cross-validation procedure for trials of T-7, T-6, T-5 and T-3, T-2, T-1 separately. The decoding scores were averaged across the cross-validation folds for each study and then the distribution of these averaged scores is plotted. Statistical significance of each unique condition was estimated by means of resampled one-sample tests against 0.5. Statistical significance of the difference between "past" and "recent" conditions was estimated by comparing the corresponding pair of "past" and "recent" conditions. Error bars show the standard errors across cross-validation folds. The estimated significant levels were corrected by Bonferroni correction procedure. n.s. = not significant. *$p < .05$. **$p < .01$. ***$p < .001$

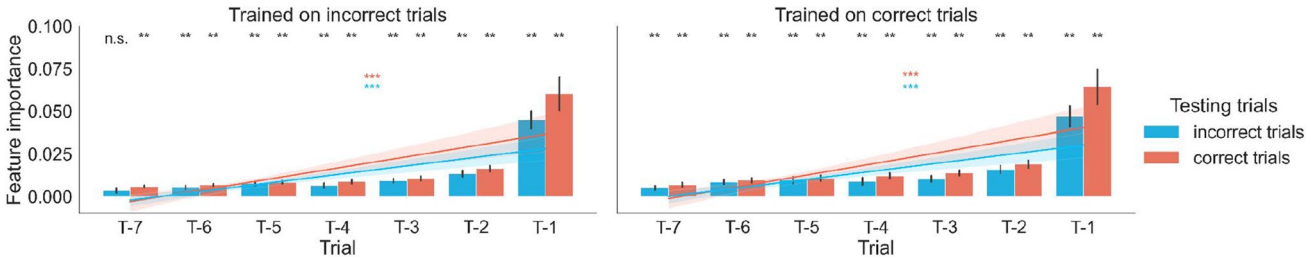

Fig. 3 Illustration of the feature importance estimates of the RF classifiers within the perceptual domain. Permutation tests were performed on the estimates associated with each trial back assessing its significance against zero importance. The vast majority of the trials back had feature importances that were statistically significant after multiple comparison correction. The only feature importance that was not significant was the feature importance of the seventh trial back when training and testing on incorrect trials ($p = .059$). Error bars show the standard errors across cross-validation folds. The rest of the feature importance estimates were all significantly greater than zero ($p < .001$). The shaded areas associated with the regression lines were resampled standard errors of the fitted regression line. The resampling method was random sampling, differing from the cross-validation procedure, which was implemented with the algorithm of the Seaborn library. n.s. = not significant. *$p < .05$. **$p < .01$. ***$p < .001$

We also analyzed the feature importance of the random forest classifier (see Methods) associated with each of the trials back. The results showed that the vast majority of the trials back had feature importances that were significantly different from zero, though the most recent trial was most important for predicting confidence in the current trial (see Fig. 3). We then fit a linear regression model, using the trials as the independent variable and the feature importance as the dependent variable. The model fitting was conducted for (1) results using correct trials as the training set and tested with correct trials, (2) results using correct trials as the training set and tested with incorrect trials, (3) results using incorrect trials as the training set and tested with correct trials, and (4) results using incorrect trials as the training set and tested with incorrect trials. After cross-validating the linear regression model using leave-one-study-out procedure as described above, we derived the statistical significance of the slope of the regression model using a permutation test. During permutation, we shuffled the correspondence between the independent variable and the dependent variable, and fitted a linear regression model using the shuffled data. This procedure was repeated 1,000 times to estimate the empirical chance level of the slope of the regression model we fitted with the unshuffled data. The significant level of the slope was the probability of the fitted slope being greater or equal than the empirical chance levels. The results of permutation tests for four conditions were corrected using Bonferroni correction. The slopes were all positive over the four conditions, and they were all statistically significant (see Fig. 3). Therefore, the results suggested that the more recent trials contributed more to the prediction of the current confidence.

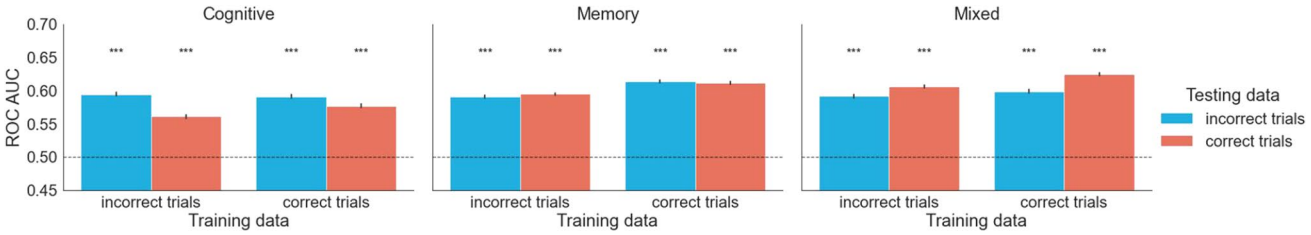

Fig. 4 Cross-domain decoding scores of confidence predictions. The random forest classifiers were trained on the perceptual domain and then tested on the other three domains. The statistical significance was estimated by resampled one-sample tests against 0.5. Each subplot shows the performance of cognitive, memory and mixed domains respectively. Error bars show the standard error across cross-validation folds. The estimated significant levels were corrected by Bonferroni correction procedure. n.s. = not significant. *$p < .05$. **$p < .01$. ***$p < .001$

Testing the generalization of confidence across domains

Having demonstrated that current estimates of perceptual confidence are affected by prior metacognitive decisions, we then examined whether the construction of metacognitive confidence in the perceptual domain generalizes across different other domains. Here, the classifier was trained on one of the studies in the perceptual domain (see Methods) and then tested on all the data of each experiment in a different domain (i.e., cognitive, memory, and mixed domains; see Fig. 4 and Supplementary Table S3). We observed that regardless of the training data, testing data, type of models, or domains of generalizing to, the confidence ratings of the previous seven trials were able to predict the confidence rating in the current trial (all scores were compared against the 0.5 level; all ps $< .001$, corrected for multiple comparisons). Similar results were observed using the SVM classifier (Fig. S3).

The above results demonstrate that a classifier trained in the perceptual domain generalizes to predict confidence across different cognitive domains.

We then tested whether the trials that were closer to the current trial were more important for predicting confidence. The same decoding analyses were conducted but now using just the most recent trials (-1, -2, -3 trials back) or the past trials (-5, -6, -7 trials back). Again, we observed that confidence in the current trial could be predictive from both recent and also past trials; however, decoding from the most recent trials was better (see Fig. 5). Similar results were observed using the SVM classifier (Fig. S4).

As before, we determined the relative importance of each of the trials back at predicting metacognitive confidence in the current trial. The results showed that the vast majority of the previous trials had feature importances that were significantly different from zero—though, again, the most recent trial was most important for generalizing confidence in the current trial across a different domain. Additionally, we fitted linear regression models to measure the feature importance as a function of trial indices, separately for different domains, training data, and testing data. There was a positive linear trend among all of these regression models. The results suggested that the confidence from the more recent trials were the most informative to predict the confidence ratings at the current trial. Figure 6 illustrates these results.

We note here that we elected to train the classifiers in the perceptual domain because the amount of training examples was much bigger in this domain. However, it may be asked whether the pattern of results generalize when the classifiers are trained in the memory or cognitive domains.

To quantify the full generalizability of confidence decoding across all domains, we tested how the decoders performed when they were trained and tested within the same domain (i.e., memory or memory) and when they were tested the decoders in a different domain (i.e., from memory to perception or cognitive). In doing so, we also tested how different attributes beyond the history of confidence ratings, namely, the history of Type 1 reaction time, and the history of accuracy, contributed to predict the confidence in the current trial.

We run the classification for each combination of the different attributes: confidence only, reaction time only, accuracy only, confidence and reaction time, confidence and accuracy, and confidence, reaction time, and accuracy. In each combination, the features of each attribute were concatenated horizontally. Note that for this analysis, the data were not split into "correct" and "incorrect" examples. When the classifier was cross-validated within the same domain, we followed the leave-one-subject-out cross-validation procedure, so that the classifier was trained in all the data except one subject, and then tested in the left-out subject. When the classifier was cross-validated across different domains, we first trained the classifier in a given domain (i.e., Memory) and then tested the classifier in another domain (i.e., Cognitive) for each subject.

Figure 7 shows the decoding results demonstrating that history of confidence alone was the most critical attribute

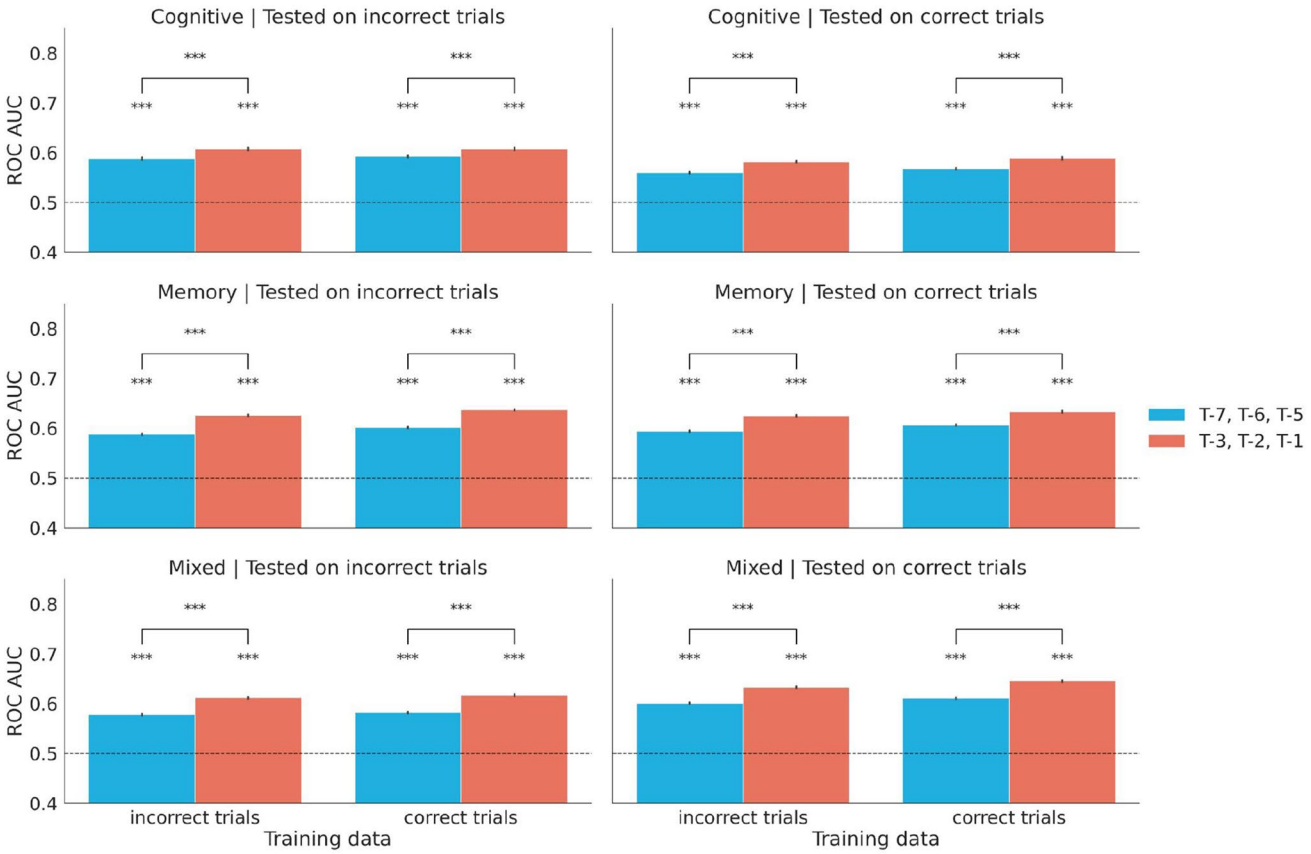

Fig. 5 Decoding confidence across domains based on recent vs. past trials. The random forest classifiers were trained on the perceptual domain and then tested on the other three domains for trials of T-7, T-6, T-5 and T-3, T-2, T1 separately. The statistical significance of each unique condition was estimated by means of resampled one-sample tests against 0.5. Statistical significance of the difference between "past" and "recent" conditions was estimated by comparing the corresponding pair of "past" and "recent" conditions. Error bars show the standard errors across cross-validation folds. The estimated significant levels were corrected by Bonferroni correction procedure. n.s. = not significant. *$p < .05$. **$p < .01$. ***$p < .001$

for predicting confidence in the current trial, and that the recent history of accuracy or Type 1 reaction times did not contribute to predicting the current confidence. Figure 8 displays the feature importance of the confidence ratings in the previous trials. The pattern of results was similar to that reported in the above analyses. Across the different domains, the confidence rating from the previous one trial was generally the most important feature compared with the confidence ratings between T-2 and T-7 trials. This pattern did not change when including or adding information related to the history of accuracy and/or reaction time (Fig. 8).[1]

Similar results were observed using the SVM classifier (Fig. S5). We then analyzed the slopes of the regression lines of the feature importance as a function of previous trials individually for different attributes (confidence, accuracy, and RT). We concatenated the features of all attributes from the previous trials to predict the confidence in the current trial. The results are presented in Fig. 8 (see also Table S1 for the statistical significance measures of the models). The results show that confidences from the previous trials

[1] We conducted a nonparametric comparison between the generalization performance in the within- and cross-domain. We first computed the difference of the average generalization performance between the within- and cross-domain. We then created a vector by concatenating the generalization performance of the within- and cross-domain. This vector was shuffled into two new groups, and the difference between within- and cross-domain decoding was recomputed 10,000 times to estimate the distribution of the empirical chance level. Finally, we computed the significant level by calculating the probability of the absolute value of the experimental difference being greater or equal to the absolute distribution of the empirical chance level (two-tailed). The results showed that the ROC-AUC decoding scores were significantly 0.04 higher in the cross- relative to the within-domain decoding ($p < .001$). However it is difficult to make inferences based on this result because of several factors such as different amounts of examples for training the random forest classifier in the within- versus cross-domain generalization.

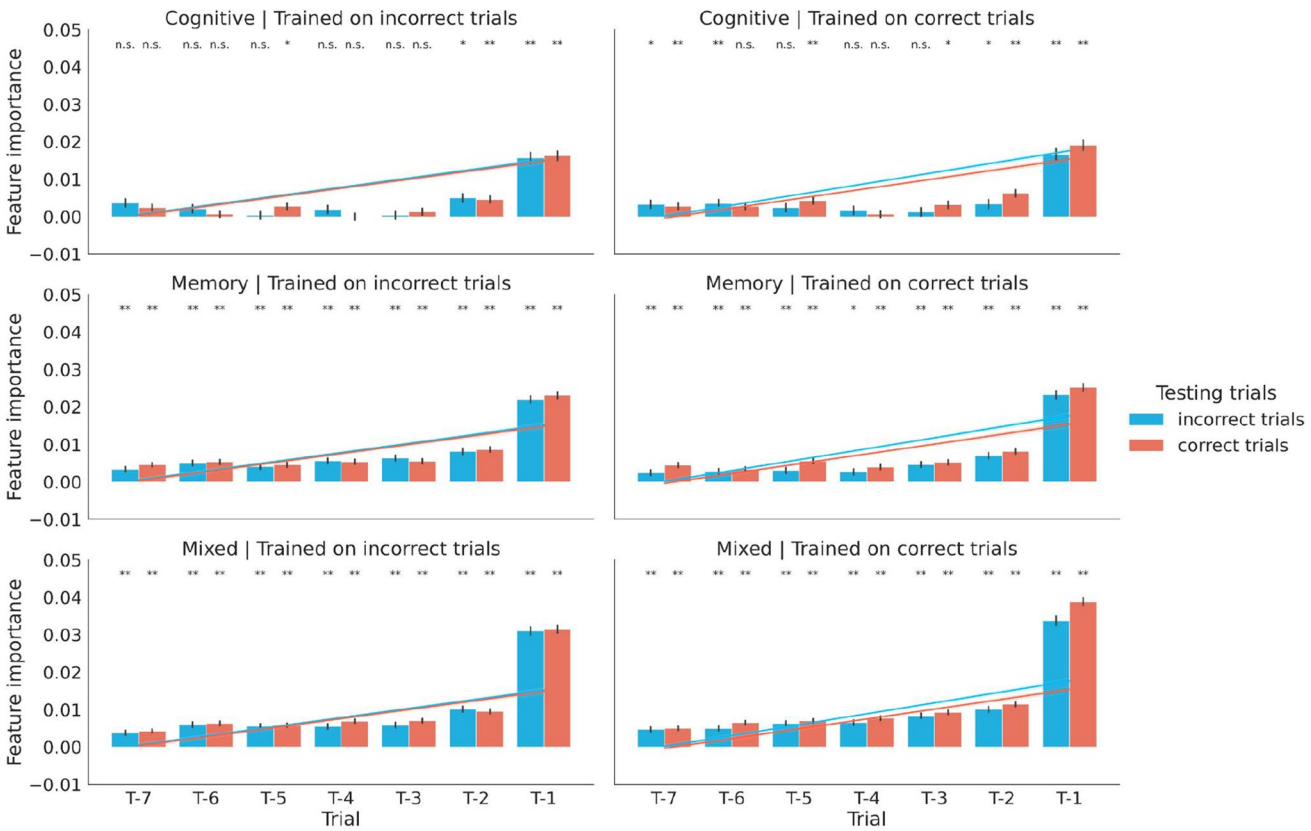

Fig. 6 Illustration of the feature importance estimates of the RF classifiers trained in the perceptual domain for cross-domain decoding. Error bars show the standard errors across cross-validation folds. Permutation tests were conducted on the feature importance estimates, comparing against zero for each of the trials back. The shaded areas associated with the regression lines were resampled standard errors of the fitted regression line. The resampling methods used the random sampling implemented with the algorithm of the Seaborn library. n.s. = not significant. $*p < .05$, $**p < .01$, $***p < .001$, corrected for multiple comparisons within each domain

contributed the most to the prediction of confidence in the current trial, and the most recent trials contributed more to the prediction than past trials. This linear trend was statistically significant within and across domains. On the other hand, relative to confidence, the accuracy and RT from the previous trials contributed little to the prediction of the current confidence.

Finally, we note that the one-versus-rest classification procedure that we used does not take into account the ordinal nature of confidence. We elected to use the random forest classifier to perform classification because it allowed us to compute the importance of each feature and make interpretable relationships between the features and the target (see Mei et al., 2020). Nevertheless, we re-ran the analyses using a random forest regression procedure, which should take into account the ordinal nature of confidence. The results showed that the random forest regression models were not as sensitive as the classification models, but the regression models performed better than chance level when the feature attributes included confidence from the previous trials (i.e., confidence only, confidence + accuracy, confidence + RT, and all three attributes; Supplementary Figs. 7 and 8, and Supplementary Table S2 illustrate the results). One potential reason that the regression models did not perform as good as the classification models may relate to the measure of the model performance—namely, variance explained (i.e., how much variance of the true confidence ratings were explained by the predicted confidence ratings). This measure may not be sensitive enough for the goal of prediction. Additionally, variance explained is more sensitive to extreme values in the predicted confidence ratings, which is not the case for ROC AUC.

Discussion

The goal of the present study was to determine whether serial dependence effects in confidence generation generalize across different observers and cognitive domains. In the perceptual domain, classification results across observers showed that the current level of confidence can be predicted by the prior confidence estimates. This result is in keeping

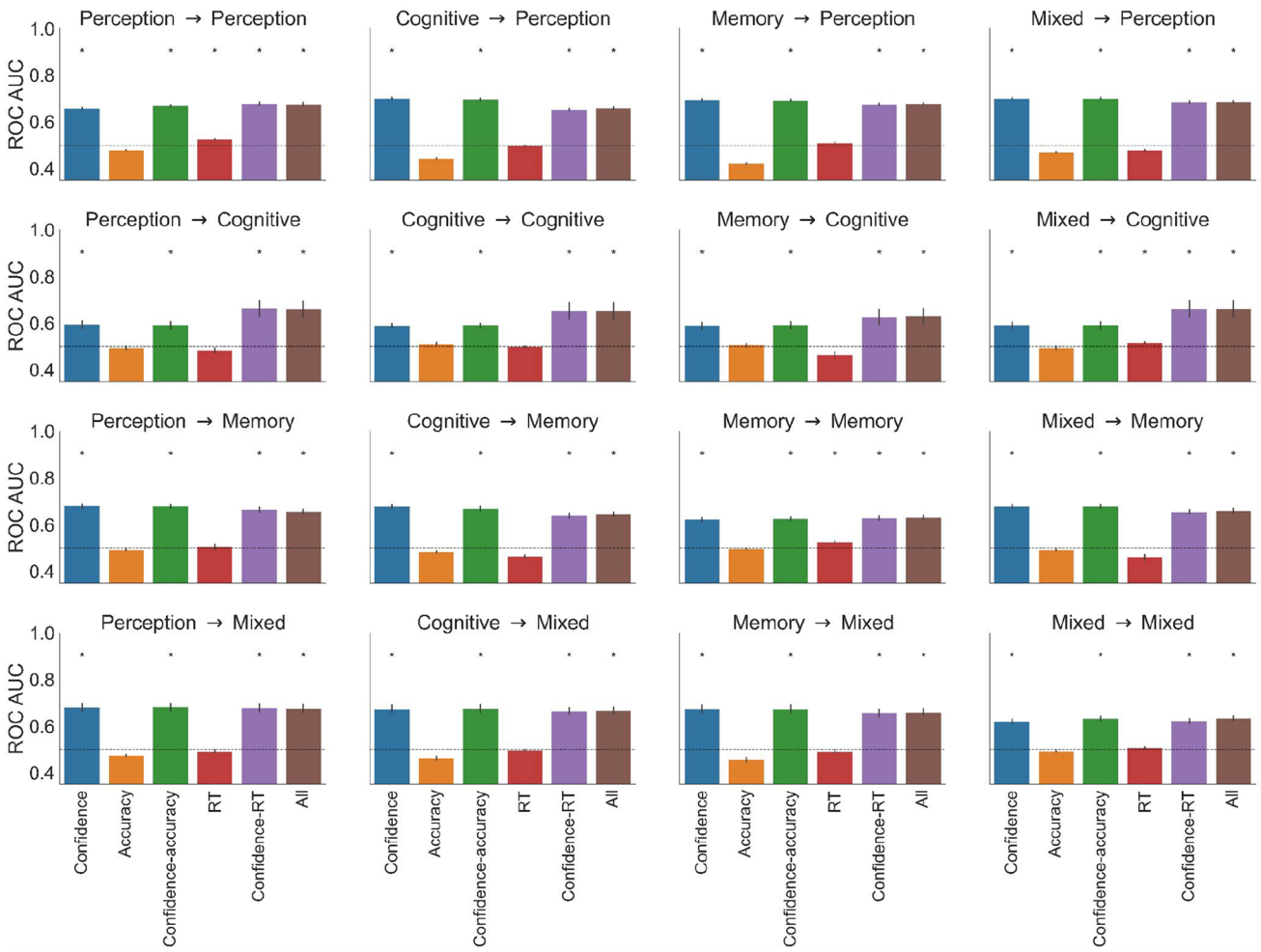

Fig. 7 Within- and cross-domain decoding ROC scores of confidence predictions. The random forest classifiers were trained on a given domain and then tested on the other three domains. The statistical significance was estimated by resampled one-sample tests against 0.5. Error bars show the standard error across cross-validation folds. The estimated significant levels were corrected by Bonferroni correction procedure. n.s. = not significant. $*p < .05$. $**p < .01$

with previous studies on confidence leak (Aguilar-Lleyda et al., 2021; Mueller & Weidemann, 2008; Rahnev et al., 2015). We also observed that classifiers trained to predict confidence in the perceptual domain generalized to other cognitive domains. The extent to which confidence can be predicted is best addressed by using cross-validation with out-of-sample generalization. For instance, a recent study showed evidence of serial dependence across domains in a paradigm in which recognition judgments were interleaved with perceptual judgments so that confidence in the perceptual judgment influenced subsequent memory confidence in the next trial (Kantner et al., 2019). However, this study used standard statistical approaches in which all data are fitted at once within the same statistical model, thereby not allowing quantification and prediction of confidence in new examples and experimental contexts—namely, across different observers and across cognitive domains, as observed in the current study. Intriguingly, confidence predictions generalized across correct and incorrect trials suggesting that serial dependence effects in confidence generation are uncoupled to metacognition (i.e., how confidence tracks accuracy). In line with this, the results showed that the recent history of confidence was the most critical factor and that accuracy or Type 1 reaction time alone, or in combination with confidence, did not improve the prediction of the current confidence. Relatedly, previous observation indicated that interindividual differences in serial dependence are negatively associated with metacognitive sensitivity (Rahnev et al., 2015). This observation highlights that serial dependence may not always confer adaptive value to perception (Cicchini et al., 2018) and cognitive performance, but can under certain circumstances provide suboptimal or maladaptive outcomes (Kiyonaga et al., 2017). This pervasiveness of serial dependence may reflect the habit of human

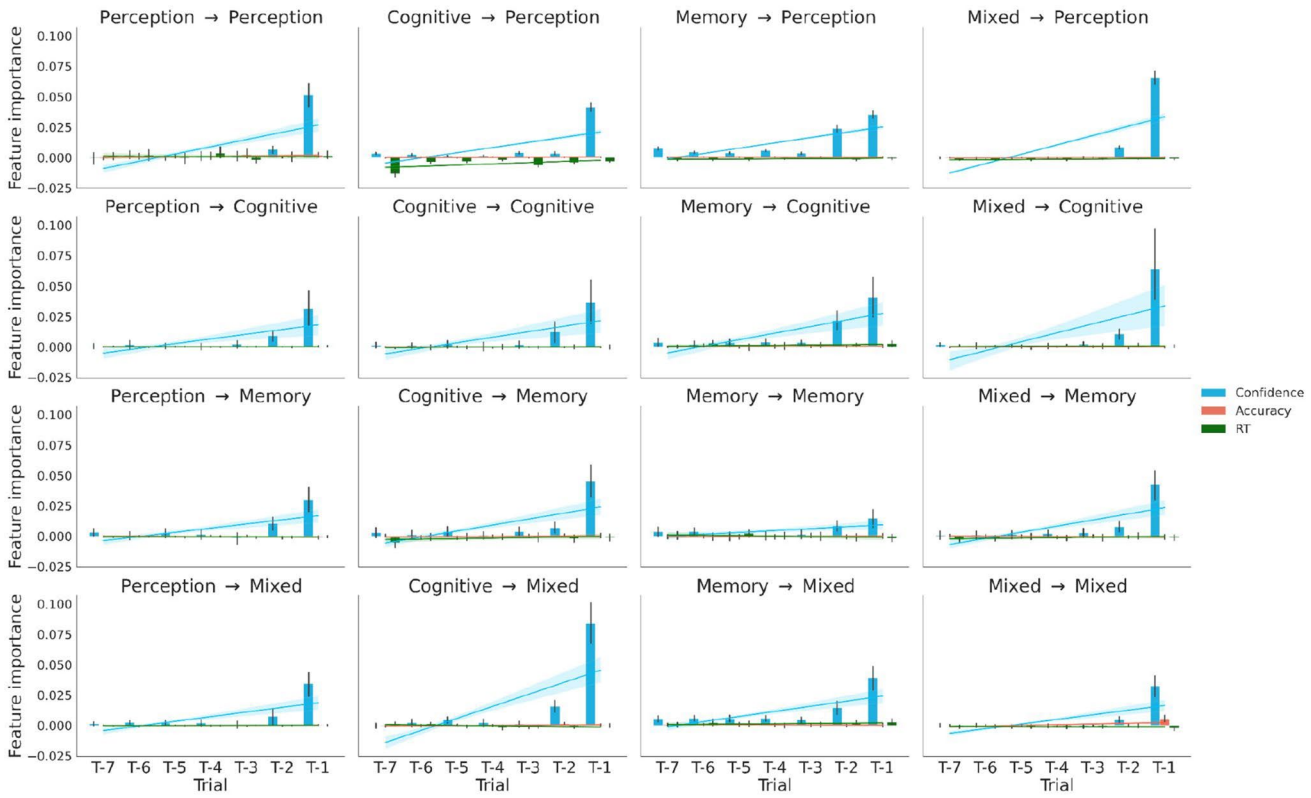

Fig. 8 Feature importance of the attributes of the previous trials, confidence, accuracy, and reaction time, in predicting the confidence in the current trial. The random forest classifiers were cross-validated with confidence, accuracy, and reaction time from the previous seven trials as concatenated features and with confidence in the current trial as the target (see Methods). (Color figure online)

observers to experience the world in a stable, auto-correlated manner (Fischer & Whitney, 2014). The generalization of serial dependence in confidence generation across correct and incorrect trials is difficult to explain according to strict normative models of decision confidence as reflecting the read-out of the perceptual signal (Hebart et al., 2016; Kiani & Shadlen, 2009; Macmillan & Douglas Creelman, 2004), and it is in line with prior studies that observed dissociations between perceptual performance and metacognitive confidence (Desender et al., 2018; Koizumi et al., 2015; Samaha et al., 2016, 2019).

The current results have important implications for the ongoing debate on the domain-generality/specificity of metacognition. Evidence in favor of domain-general metacognitive mechanisms comes from studies showing that observers can successfully assign confidence to two discriminations in different modalities (i.e., visual and auditory), notably, as efficiently as when the two discriminations involve the same sensory domain (de Gardelle et al., 2016). This question is normally addressed in behavioral studies by computing the correlation of individual measures of metacognition (i.e., meta-d' or M-ratio) across different domains (i.e., perception and memory). Studies using this approach have found mixed results with some reporting significant correlations (Fitzgerald et al., 2017; McCurdy et al., 2013; Ruby et al., n.d.) but others failing to find such correlations (Baird et al., 2015; Baird et al., 2013; Morales et al., 2018). A recent meta-analysis (Rouault et al., 2018) of behavioral studies observed mixed evidence for the association of interindividual measures of metacognitive performance across perception and memory domains (but see Mazancieux et al., 2020). Different factors can contribute to these mixed results: lack of statistical power, differences in metacognition measures and differences in task requirement across studies, uncertainty in the estimation of the model parameters across different indices of metacognitive performance such as meta-d' (Rouault et al., 2018), and also different sources of metacognitive inefficiency dominating different tasks (Shekhar & Rahnev, 2021). The current results contribute to this debate by showing that classifier predictions of confidence judgments generalize across cognitive domains. However, we note that the domain-generality issue in metacognitive research typically relates to metacognitive sensitivity or efficiency, rather than confidence per se. The present results indicate that serial dependence in confidence ratings is a robust phenomenon and likely not domain specific.

Psychology findings have been subject to recent scrutiny due to failures to replicate in new samples or replications

that do not hold up to the size of the published effects (Open Science Collaboration, 2015). One of the factors contributing to replication failures may be related to misuse and appropriateness of statistical tests, beyond the so-called *p* hacking or the selective subsampling of the data. Many studies fit a statistical model with all experimental data at once, which can lead to overfitting and poor generalization of the model with new observations that are similar but come from a different sample. Because of this, it has been argued that psychological science can greatly benefit from the field of machine learning in which pattern classifiers are tested in their ability to predict new data coming from the same or different participants (Yarkoni & Westfall, 2017). These authors have made the strong argument that psychological theories contribute little to predict future human behavior with a respectable level of precision. In the same vein, there has been a recent emphasis on changing current statistical practices and encourage the adoption of "estimation thinking," namely, to provide a quantitative model of the effect under investigation, rather than the standard "dichotomous thinking" based on the traditional null hypothesis testing framework to reject the null (Cumming, 2014). The present approach using pattern classification with out-of-sample generalization thereby provides evidence that confidence generation can be reliably quantified and predicted across new samples and experimental contexts. The use of machine learning can prove very useful towards developing predictive models of confidence across different populations and experimental contexts (see Fleming et al., 2016; Mei et al., 2020 for a similar approach to predict prospective beliefs of self-performance). It is difficult to make conclusions regarding the level of ROC prediction scores obtained here without a prior context of similar studies using different measures of predictive performance. However, ROC is arguable the best measure of predictive performance compared with other measures of effect size (Rice & Harris, 2005), and the current ROC confidence prediction scores of ~0.65 for within-domain and slightly lower for cross-domain generalization, indicate that confidence prediction for unobserved data is a robust phenomenon.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13423-023-02261-x.

Acknowledgments D.S. acknowledges support from the Basque Government through the BERC 2022-2025 program, from the Spanish State Research Agency, through the 'Severo Ochoa' Programme for Centres/Units of Excellence in R&D (CEX2020-001010-S). This project was funded by project grant PID2019-105494GB-I00 from the Spanish State Research Agency (DS). We thank Megan Peters and Alan Lee for the helpful feedback on a previous version of the manuscript.

Data Availability The data for all experiments and analysis scripts are available online (https://osf.io/s46pr/) and (https://github.com/nmningmei/decoding_confidence_dataset).

References

Aguilar-Lleyda, D., Konishi, M., Sackur, J., & de Gardelle, V. (2021). Confidence can be automatically integrated across two visual decisions. *Journal of Experimental Psychology: Human Perception and Performance, 47*(2), 161–171.

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics, 26*(10), 1340–1347.

Ais, J., Zylberberg, A., Barttfeld, P., & Sigman, M. (2016). Individual consistency in the accuracy and distribution of confidence judgments. *Cognition, 146*, 377–386.

Baird, B., Smallwood, J., Gorgolewski, K. J., & Margulies, D. S. (2013). Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 33*(42), 16657–16665.

Baird, B., Cieslak, M., Smallwood, J., Grafton, S. T., & Schooler, J. W. (2015). Regional white matter variation associated with domain-specific metacognitive accuracy. *Journal of Cognitive Neuroscience, 27*(3), 440–452.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Cicchini, G. M., Mikellidou, K., & Burr, D. C. (2018). The functional role of serial dependence. *Proceedings. Biological Sciences / The Royal Society*, *285*(1890). https://doi.org/10.1098/rspb.2018.1722

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29.

de Gardelle, V., Le Corre, F., & Mamassian, P. (2016). Confidence as a common currency between vision and audition. *PLOS ONE*, *11*(1), Article e0147901.

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological Science, 29*(5), 761–778.

Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research, 9*, 1871–1874.

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neuroscience, 17*(5), 738–743.

Fitzgerald, L. M., Arvaneh, M., & Dockree, P. M. (2017). Domain-specific and domain-general processes underlying metacognitive judgments. *Consciousness and Cognition, 49*, 264–277.

Fleming, S. M., Massoni, S., Gajdos, T., & Vergnaud, J.-C. (2016). Metacognition about the past and future: Quantifying common and distinct influences on prospective and retrospective judgments of self-performance. *Neuroscience of Consciousness*, *2016*(1), niw018.

Hebart, M. N., Schriever, Y., Donner, T. H., & Haynes, J.-D. (2016). The relationship between perceptual decision variables and confidence in the human brain. *Cerebral Cortex, 26*(1), 118–130.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67.

Kantner, J., Solinger, L. A., Grybinas, D., & Dobbins, I. G. (2019). Confidence carryover during interleaved memory and perception judgments. *Memory & Cognition, 47*(2), 195–211.

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

Kiyonaga, A., Scimeca, J. M., Bliss, D. P., & Whitney, D. (2017). Serial dependence across perception, attention, and memory. *Trends in Cognitive Sciences, 21*(7), 493–497.

Koizumi, A., Maniscalco, B., & Lau, H. (2015). Does perceptual confidence facilitate cognitive control? *Attention, Perception, & Psychophysics, 77*(4), 1295–1306. https://doi.org/10.3758/s13414-015-0843-3

Liberman, A., Fischer, J., & Whitney, D. (2014). Serial dependence in the perception of faces. *Current Biology: CB, 24*(21), 2569–2574.

Macmillan, N. A., & Douglas Creelman, C. (2004). *Detection theory: A user's guide*. Psychology Press.

Manassi, M., Liberman, A., Chaney, W., & Whitney, D. (2017). The perceived stability of scenes: Serial dependence in ensemble representations. *Scientific Reports, 7*(1). https://doi.org/10.1038/s41598-017-02201-5

Mazancieux, A., Fleming, S. M., Souchay, C., & Moulin, C. (2020). Retrospective confidence judgments across tasks: Domain-general processes underlying metacognitive accuracy. *PsyArXiv Preprints.* https://doi.org/10.31234/osf.io/dr7ba

McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., de Lange, F. P., & Lau, H. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 33*(5), 1897–1906.

Mei, N., Rankine, S., Olafsson, E., & Soto, D. (2020). Similar history biases for distinct prospective decisions of self-performance. *Scientific Reports, 10*(1), 5854.

Morales, J., Lau, H., & Fleming, S. M. (2018). Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 38*(14), 3534–3546.

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review, 15*(3), 465–494.

Open Science Collaboration. (2015). PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Rsearch, 12*, 2825–2830.

Rahnev, D., Koizumi, A., McCurdy, L. Y., D'Esposito, M., & Lau, H. (2015). Confidence leak in perceptual decision making. *Psychological Science, 26*(11), 1664–1680.

Rahnev, D., Desender, K., Lee, A. L. F., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., Arbuzova, P., Atlas, L. Y., Balcı, F., Bang, J. W., Bègue, I., Birney, D. P., Brady, T. F., Calder-Travis, J., Chetverikov, A., Clark, T. K., Davranche, K., Denison, R. N., Dildine, T. C., … Zylberberg, A. (2020). The confidence database. *Nature Human Behaviour*, *4*(3), 317–325.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's *d*, and *r*. *Law and Human Behavior, 29*(5), 615–620. https://doi.org/10.1007/s10979-005-6832-7

Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018). Human metacognition across domains: Insights from individual differences and neuroimaging. *Personality Neuroscience*, *1*. https://doi.org/10.1017/pen.2018.16

Ruby, E., Giles, N., & Lau, H. (n.d.). *Finding domain-general metacognitive mechanisms requires using appropriate tasks*. https://doi.org/10.1101/211805

Samaha, J., Barrett, J. J., Sheldon, A. D., LaRocque, J. J., & Postle, B. R. (2016). Dissociating perceptual confidence from discrimination accuracy reveals no influence of metacognitive awareness on working memory. *Frontiers in Psychology, 7*, 851.

Samaha, J., Switzky, M., & Postle, B. R. (2019). Confidence boosts serial dependence in orientation estimation. *Journal of Vision, 19*(4), 25.

Shekhar, M., & Rahnev, D. (2021). Sources of Metacognitive inefficiency. *Trends in Cognitive Sciences, 25*(1), 12–23.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 12*(6), 1100–1122.