Machine learning predicts human prospective decision making

Ning Mei[1], Sean Rankine[3], Einar Olafsson[3], and David Soto[1,2]

[1] *Basque Center on Cognition, Brain and Language, San Sebastian, Spain*
[2] *Ikerbasque, Basque Foundation for Science, Bilbao, Spain*
[3] *Imperial College London, United Kingdom*

April 12, 2019

Correspondence to: d.soto@bcbl.eu, Basque Center on Cognition, Brain and Language, Paseo Mikeletegi 69, 2nd Floor 20009 San Sebastian

Abstract

Metacognition can be deployed retrospectively (i.e. to reflect on the correctness of our recent behaviour) or prospectively (i.e. to make predictions of success in one's future behaviour or make decisions about strategies to solve future problems). We sought to investigate the factors that determine this sort of prospective decision making. Human participants performed a visual discrimination task followed by ratings of stimulus visibility and response confidence. Prior to each discrimination trial participants made prospective judgments concerning the upcoming task. In Experiment 1, they rated their belief of future success. In Experiment 2, they rated their decision to adopt a focussed attentional state. Both types of prospective decisions were related to behavioural performance in different ways. Prospective beliefs of success were associated with no performance changes while prospective decisions to engage attention were followed by better self-evaluation of the correctness of behavioural responses. Using standard machine learning classifiers we found that the current prospective decision could be predicted from information concerning task-correctness, stimulus visibility and response confidence from previous trials. In both Experiments, awareness and confidence were more diagnostic of the prospective decision than task correctness. Notably, classifiers trained with prospective beliefs of success in Experiment 1 predicted decisions to engage in Experiment 2 and vice-versa. These results indicate that the formation of these seemingly different prospective decisions share a common, dynamic representational structure.

Introduction

The capacity to think about one's own thoughts and behaviour is a fundamental constituent of the human mind which is known as metacognition. We rely on it to recognize that we have made a mistake, to realize that we have forgotten something important or to appreciate how confident we are about our own knowledge. An influential model of metacognition[1] highlights the interplay between first-order, task-related processes (i.e., involving our perceptions and responses during task performance) and second-order processes, originating from the prefrontal cortex, that 'monitor' the correctness of the first-order process.[2,3]

Monitoring of one's own behavioural performance is typically assessed by means of retrospective reports, in which participants give ratings of confidence about their perceptual judgments or rate the state of visual awareness associated with the relevant stimulus. However, metacognitive processes are not just about thinking about one's past and ongoing mental states. Metacognition can also be used prospectively to guide our future behaviour. For instance, we can mentally simulate ourselves in future probable scenarios, pre-empt the type of cognitive strategies needed to solve specific problems and adapt behaviour according to learning needs.

Prior research in the memory domain addressed how people make prospective judgments of learning during study,[4,5] and revealed, for instance, how decisions to study further rely on the evaluation of one's own learning[6,7] and how this self-evaluation during study relates to subsequent memory accuracy. However, little is know about the factors that influence prospective metacognition during perceptual decision making. In addition to monitoring one's own behavioral performance and forming prospective beliefs about future success, people also engage in self-regulation. For instance, people may also decide to put more attention when they lack confidence in their knowledge or stop further study when they are confident.

We here sought to investigate whether or not the formation of seemingly different types of prospective decisions (i.e. beliefs of success and decisions to engage with the environment) make use of similar information and recruit similar processes. Specifically, using a paradigm involving visual perceptual decisions we investigated the factors that predict future prospective beliefs of performance success vs. prospective decisions to engage with the environment (i.e. decisions to adopt a focussed attention state). Being successful in a higher proportion of recent trials may influence one's estimation of prospective confidence, leading to predictions of a high probability of success in the next trial.[8] A recent study by Fleming and colleagues (2016) investigated the formation of prospective and retrospective confidence.[9] In this study, participants were asked to perform a motion discrimination task following by confidence ratings, and, every five trials participants also rated the prospective belief of success in the upcoming trial. The results showed that the prospective judgments were not associated with subsequent metacognitive performance (i.e. the association between confidence and accuracy). By contrast, in keeping with prior work, retrospective

confidence judgments were closely aligned with task accuracy. Fleming and colleagues then modeled the influence of the history of previous trials information (e.g. past confidence and past accuracy) on the subjective estimates of prospective and retrospective confidence for a given trial. The results showed that current retrospective confidence can be predicted by the estimate of retrospective confidence in the previous trial. Prospective confidence was however dependent on the previous estimate of prospective confidence, and also on the prior retrospective confidence ratings over a longer timeframe (i.e. involving the previous four trials). The influence of task accuracy on prospective beliefs of success was far weaker by comparison.

Here we wondered whether different types of prospective decisions may rely on specific sources of information. After all, one's certainty of the adequacy of one's behavioural responses may be dependent on a host of different factors, including stimulus visibility, response interference from competing for distracting information and additional biases and heuristics.[10] For instance, given a challenging perceptual task, a state of low visibility of the critical target may encourage the observer to decide to invest more effort in the next trial but may also lead to a reduction in confidence about his prospective accuracy. It is not known whether or not seemingly different types of prospective decision making (i.e. a decision to adopt a more focussed attention state on the next trial vs. prospective beliefs of success) are dependent on the same factors.

Here we modified an existing paradigm[11] to quantify the contribution of the state of visual awareness, task-confidence and task-correctness to prospective decisions. We wondered whether a similar or distinct pattern of experiences influenced different types of prospective metacognitive beliefs, namely, predictions of success and decisions to engage with the environment. Participants were presented with an oriented Gabor patch near the threshold of visual awareness. Prior to the presentation of the Gabor, on each trial, participants indicated their belief of success (i.e. low or high) in the upcoming orientation discrimination task (Experiment 1) or indicated their decision to engage a focussed attention state (low or high; Experiment 2). Following the presentation of the Gabor, participants rated their visual awareness, responded to its orientation and rated their confidence in the orientation response.[11,12] Using standard machine learning algorithms, we sought to predict these seemingly different prospective decisions based on the history of awareness, task-confidence, and task-accuracy from previous trials. We also evaluated the relative importance of these factors for prospective judgements using the coefficients from logistic regression. We also used a random forest classifier in order to estimate the stability of the decoding performance. Similar results were obtained. The results from the random forest classifier are presented in the Supplemental Materials.

Finally, we asked whether these seemingly different prospective decisions play a functional role in shaping our subsequent perception or metacognitive performance. We then tested whether the performance was affected by the type of prospective belief or decision to engage attention. The 'self-fulfilling prophecy' offers a view on the potential effect of prospective beliefs upon behavioural performance. Prospective beliefs may set an expectation that the participant is

motivated to meet.[13,14] One possibility is that estimations of high probability of success may encourage observers to invest more cognitive resources in the upcoming trial and hence facilitate performance in a similar way to decisions to engage focussed attention. This study was devised to test these hypotheses.

Experiment 1: Predictions of success

Methods

Participants

Following informed consent, eighteen participants (19-23 years, mean age: 20.6, 6 males) took part in return of monetary compensation. This sample size was selected based on our prior study in which a similar paradigm was used.[11] Data from three participants were excluded before analyses. One of the participants only reported a total of 3% of aware trials and two participants provided no responses in three conditions of awareness and confidence, impeding further analyses. This was likely due to inadequate pre-calibration of stimulus luminance (see below). The study conformed to the Declaration of Helsinki and was approved by the West London Research Local Research Ethics committee.

Experimental task and procedure

The experiment took place in a dimly lit room with a viewing distance of approximately 90 cm. The task was programmed and controlled by Psychopy.[15] Stimuli were presented in a CRT monitor with a resolution of 1.600 x 1.200 pixels and a refresh rate of 85 Hz.

Figure 1 illustrates the experimental task. On each trial, participants were required to discriminate the orientation of a brief, masked Gabor, presented at the threshold of visual awareness. Prior to the presentation of the Gabor, participants reported their prospective belief of success associated with the upcoming trial (low vs high). Following the offset of the Gabor, participants rated their visual awareness of the Gabor, responded to its orientation and provided confidence ratios on the accuracy of the orientation responses.

During each trial, participants first made the prospective metacognitive decision during an unlimited time window. Then, a Gabor patch was presented in the center of the screen with a grey background (luminance = 10.48 cd/m^2. Mask luminance was 11.34 cd/m^2. The orientation of the Gabor was either 40 degrees to the left or right from vertical, and was randomly varied on each trial, with equal probability for each orientation. Participants responses were recorded using the keyboard.

Participants were instructed to complete a preliminary practice phase in order to get used to the orientation discrimination task. During this phase, the Gabor was presented for 362 ms with a fixed luminance of 11.93 cd/m^2, and followed by immediate feedback regarding the accuracy of their response. Next,

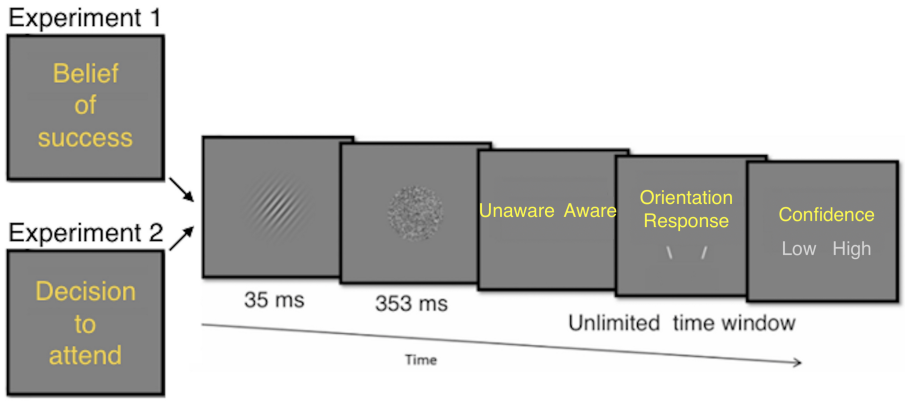

Figure 1: Illustration of the experimental protocol

a calibration phase took place with a 35 ms Gabor stimulus duration and a 353 ms mask, similar to the experimental trials. Here, its luminance was varied using a staircase procedure. This meant that luminance increased when participants reported being unaware of the orientation of the Gabor and vice versa when they reported awareness. Participants were instructed to report 'unawareness' when they had no experience of the Gabor or saw a brief glimpse but otherwise had no awareness whatsoever of the orientation. They were instructed to report awareness of the Gabor when they could see its orientation somewhat or almost clear. The initial luminance was set to 11.97 cd/m^2. The percentage of aware responses was computed on a trial to trial basis and the individual awareness threshold for each participant was determined by the luminance at the point where the probability of aware reports stabilized at around 0.5 for at least 10 trials, which were used to compute the average luminance to be used in the next procedural step.

Following the calibration, participants went through a training phase where they completed 15 trials identical to the experimental trials. The Gabor was presented for 35 ms during these trials. Prior to the presentation of the Gabor target, participants were instructed to report their belief of success. Following the presentation of the target Gabor, participants reported their awareness of the Gabor, its orientation, and then rated their confidence in the orientation response. During the visibility response period, participants were presented with a screen displaying the response options (Unaware Aware). During the confidence response, period participants saw a screen with potential responses (Confidence: low high).

The justification for this particular order of responses is the following. Visual awareness of the stimulus was rated first to make sure that awareness was a genuine estimate of perceptual experience without being contaminated by memory interference from a longer delay between the stimulus and the awareness rating.

The confidence judgment was given last because this referred to the orientation discrimination response, which followed the visual awareness response.

There was no response deadline for any of the three judgments. Participants were asked to provide precise ratings of awareness, and confidence, and accurate orientation discriminations without worrying about the speed of responding. Regarding the orientation response, participants were told that even if they were unaware of the stimulus, they should use their intuition and make their best guess about the orientation of the Gabor. Regarding the confidence report, participants were instructed to report how confident they were about the correctness of the orientation response on a relative scale of 1 (relatively less) to 2 (relatively more) confident. Participants were instructed that confidence ratings should be conceived in a relative fashion and hence they were asked to use all the confidence ratings independently of the awareness rating so that participants would not simply choose low confidence every time they were unaware and vice-versa on aware trials. Previous studies using a similar paradigm indicated that observers can display metacognitive sensitivity in both aware and unaware trials.[11,12]

Prior to the experimental trials, there was a second calibration of the luminance of the Gabor starting with the luminance value from the first calibration. Each participant then completed 12 blocks of 50 trials (600 in total), with breaks between each block.

Machine learning protocols

We used standard machine learning algorithms to investigate whether prospective beliefs of success on the current trial can be predicted based on the participants' past experiences during task performance on previous trials. Hence we aimed to predict whether the belief of success was low or high given a vector of features (i.e. correctness, visual awareness, and confidence) from the previous trials, considering 1-back, 2-back, 3-back, and 4-back trials, separately. Note that all the time series of trials back is not included in the classification. For instance, when we decode the belief of success based on the pattern of confidence, correctness and awareness of 4 trials back, we are only feeding the classifier with the data from that trial and do not include trials 1, 2 and 3 back. This range of trials was included based on a prior study,[9] which showed that prospective confidence estimates relied on the history of retrospective confidence over the previous four trials, with this influence decaying with more trials back.

We employed a logistic regression model (Scikit-learn implementation). Scikit-learn by default implements the logistic regression with an L2 regularization built-in, which reduces the interpretability of the weight coefficients estimated by the regression. Therefore, we set the regularization to be very small in order to emulate no regularization as in the simplest form of logistic regression. During training, the model optimizer (LIBLINEAR), achieved the minimization of the difference between the predictions of the model and the true values we wanted to predict using a coordinate descent algorithm.[16]

Preprocessing: A Pandas dataframe[17] included the trials as rows, and the

features for classification as columns plus the predicted target. Each instance in the target vector was encoded as 0 or 1. The rating of prospective belief of success was set to 0 for the low belief of success and 1 for the high belief of success. Regarding the features for classification, the correctness was coded as 0 or 1 to denote an incorrect or correct response, visual awareness and confidence ratings were coded as 0 or 1 to denote low or high awareness/confidence rating.

Cross-validation: In order to estimate the variance of the decoding performance, we conducted a 100-fold shuffle splitting cross-validation for each subject, each decoding goal (1-back, 2-back, 3-back, and 4-back). Each fold was constructed by shuffling the rows of the data frame. Then 80% of the rows were selected to form a training set while the remaining 20% became the testing set. The predictive performance was estimated in each fold by comparing the target vector with the probabilistic predictions. The comparison was measured by the area under the receiver operating curve (ROC AUC). ROC AUC is a sensitive, nonparametric criterion-free, and less biased measure of predictive performance in binary classification,[18] with 0.5 being the theoretical chance level. The AUC-ROC represents the ratio of the true positive classification rate (TPR, i.e. the classifier predicts 'animal' given the example is an animal) against the false positive rate (FPR, i.e. the classifier predicts 'animal' given the example is a tool).

Post-decoding: In order to estimate the significance level of the decoding performance, we generated a null distribution of decoding scores for each subject using a permutation analysis with 100 iterations. The null distributions were used as the empirical chance level. The null distribution of each subject was estimated by conducting the same cross-validation procedures with the same feature matrices and target vectors, except that the order of rows of the feature matrices and the target vectors were randomized independently. The average performance score over the permutation iterations represented the chance level estimate. This was found to be centered on the theoretical chance level of 0.5.

The statistical significance of the classification scores in each condition (i.e. trial back) was determined by using a non-parametric t-test. In it, the decoding scores were assessed relative to their corresponding chance level estimates. A permutation t-test was conducted to compute the uncorrected p-value for each trial back (1, 2, 3, 4), across all the subjects. However, the level of significance was corrected using Bonferroni multiple comparison correction procedure for each experiment. The same applied for post-hoc tests after an ANOVA.

In plotting the classification results, the error bars represent bootstrapped 95% confidence intervals resampled from the average decoding scores of individual participants by the classifier with 1000 iterations.[19]

Results

Classification analyses

First, we report the pairwise Phi correlations between the features used for classification: (i) awareness and confidence: 0.244 +/- 0.203, p = 0.0001; (ii) correctness and awareness: 0.229 +/- 0.105, p = 0.0001; correct and confidence: 0.169 +/- 0.093, p = 0.0001. Although the correlations are higher than chance, the correlations are rather small. There is room therefore for each variable to provide relevant information to classify the prospective belief of success.

We used a standard logistic regression classifier to predict the prospective belief of success on a given trial by using awareness, confidence, and correctness as training features from the preceding trials. The results show that prospective belief of success could be predicted above chance levels using information from the previous trials. As shown in Figure 2, the prediction of success could be classified with the highest accuracy by using the features from the previous trial, with prediction accuracy dropping close to chance level based on information from 4 trials back. P-values following the permutation tests for the different classification analyses were as follows: 1-back: $p < 0.0004$, 2-back: $p < 0.0004$, 3-back: $p < 0.0056$, 4-back: $p < 0.0084$.

Next, we assessed the relevance of each of the different attributes (awareness, confidence, and correctness) for the classification. As we were interested in understanding the factors that contribute to future beliefs of success, we analyzed the weight coefficients (odd ratios) from the logistic regression (see Methods) by means of an ANOVA with time window (1,2, 3 and 4 trials back) and feature attribute as factors. Note that our main interest here is to understand the contribution of the different attributes for the classification. Since the above classification results already showed that classification accuracy decreases with the number of trials back, additional analyses based on significant main effects of time window are not considered further.

The analysis of the odd ratios of the logistic regression showed a main effect of window, $F(3,42) = 19.15$, $p < 0.00001$, $\eta^2 = 0.176$, and a main effect of attributes, $F(2,28) = 14.43$, $p = 0.00005$, $\eta^2 = 0.179$. Further t-tests showed that the odd ratios of both confidence ($p = 0.000003449$) and awareness ($p = 0.0008536$) was different from correctness. These results show that when observers rated high confidence/high awareness on the previous trial, then a belief of high success on the next trial was over 3 times more likely. Figure 3 illustrates this pattern of results. There was also an interaction between factors, $F(6,84) = 10.66$, $p < 0.00001$, $\eta^2 = 0.083$. In the case of one trial back ($N = 1$), the odd ratio for confidence ($p = 0.00012$) and awareness ($p = 0.0265$) was higher than the odd ratio for correctness, but there was not difference between confidence and awareness ($p = 1$). In $N = 2$, the odd ratios of confidence and awareness were different from correctness ($p = 000876$ and $p = 0.03797$), but there was not difference between confidence and awareness ($p = 1$). This pattern of results was not observed for $N = 3$ and $N = 4$ (all ps > 0.08).

This pattern of results was replicated in further analyses based on a random

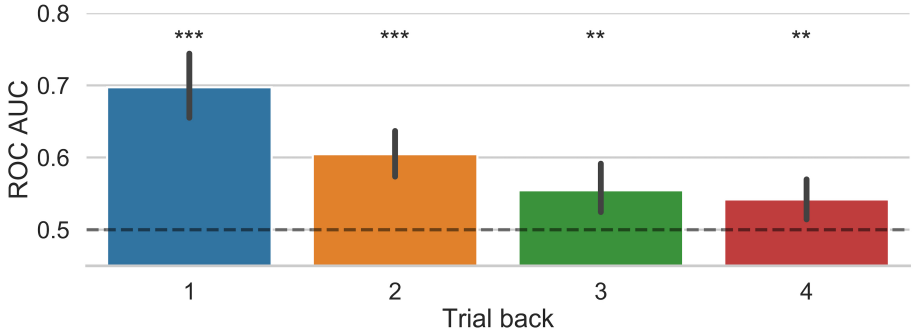

Figure 2: Results from the logistic regression classification model. Error bars represent bootstrapped 95% confidence intervals.

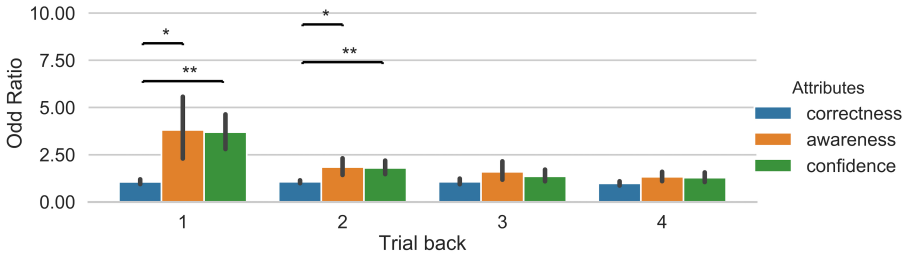

Figure 3: Illustration of odd ratios from the logistic regression for the different features involved in the classification of the belief of success. A value of 1 indicates that the odds of predicting the class is the same regardless of the feature value. Error bars represent bootstrapped 95% confidence intervals.

forest classifier (see Supplemental Figures 2 and 3).

We also performed univariate analyses of the probability of high success as a function of the experimental features of the previous trial. The results showed that the conditional probability of predictions of high success did not differ as a function of the type of feature (i.e. high awareness, high confidence, and correct trials). These results are shown in Supplementary Figure 1 and indicate that the multivariate classifier is able to capture information missed in simple univariate analyses.

Signal detection analyses

We computed type-1 d' to index the observer's sensitivity to discriminate the orientation of the Gabor, and also type-2 d' or meta-d' as a measure of metacognitive sensitivity,[20] which is basically a parametric estimation of the type-2 sensitivity (i.e. the capacity to discriminate correct from incorrect type-1 responses based on the confidence ratings) which is achieved by fitting a type-1 signal de-

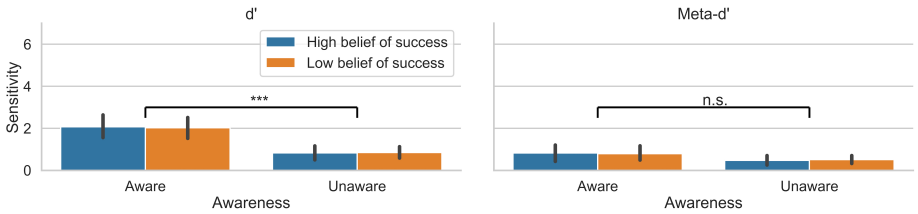

Figure 4: Perceptual sensitivity and metacogntive sensitivity scores as function of the state of awareness and the belief of success. Error bars represent standard error of the mean.

tection model to the observed type-2 performance and estimating the type-2 receiver operating characteristic (ROC) curves. In the type-1 model, the 'signal' and 'noise' were defined as left and right oriented Gabor, respectively. A 'hit' was, therefore, a correct response ('left') to a left-oriented Gabor and a 'correct rejection' was a correct response ('right') to a right-oriented Gabor. A 'false alarm' was an incorrect response (left) to a right-oriented Gabor and a 'miss' was an incorrect response (right) to a left-oriented Gabor.

A 2 x 2 ANOVA with awareness state (unaware, aware) and belief of success (high, low) as factors was carried out on the perceptual sensitivity to the orientation of the Gabor. There was an effect of awareness on type-1 d' ($F(1,14) = 52.44$, $p < 0.000004$, $\eta^2 = 0.353$). Perceptual sensitivity was higher on aware relative to unaware trials. There was no effect of prediction of success on perceptual sensitivity ($F(1,14) = 0.021$, $p = 0.888$, $\eta^2 = 0.00004$) and no interaction effect between factors ($F(1,14) = 0.096$, 0.761, $\eta^2 = 0.0002$). These results are depicted in Figure 4.

A similar ANOVA was carried out on metacognitive sensitivity scores. There was no effect of awareness on type-2 d' ($F(1,14) = 3.273$, $p < 0.0919$, $\eta^2 = 0.063$), no effect of the belief of success ($F(1,14) = 0$, $p = 0.988$, $\eta^2 = 0.00002$) and no interaction effect between belief of success and awareness ($F(1,14) = 0.042$, 0.84, $\eta^2 = 0.0008$). Metacognitive sensitivity was well above chance even in trials that participants reported being unaware of the target ($t(14) = 3.891$, $p < 0.0016$ and $t(14) = 5.218$, $p < 0.00013$, for high and low belief of success cases).

There were no effects of the belief of success on M-ratio (meta-d'/d'), which is an index of metacognitive efficiency that factors out the level of d'[20]($F(1,14) = 2.78$, $p = 0.12$. The interaction between belief of success and awareness was also non-significant ($F(1,14) = 2.44$, $p = 0.14$).

Discussion

We found that visual awareness had a profound effect on perceptual sensitivity, however, the effect of awareness on metacognitive sensitivity was far weaker. However, metacognitive sensitivity was well above chance across (un)awareness

states. This replicates our prior observation[11] and suggests that metacognitive operations are not necessarily carried out on the same type of representation or follow a similar process to that underlying first-order performance. In other words, the precision of retrospective metacognitive confidence judgments can be dissociable from factors that influence task performance (i.e. stimulus visibility). Hence metacognitive confidence and conscious visual awareness can be dissociated (see also.[21] This is also supported by the weak correlation found between confidence and visibility ratings.

Most important for the aims of the present study, we observed that the current prospective belief of performance success could be predicted based on the pattern of visual awareness, task-confidence, and task-correctness, notably from several trials back. The prediction of this belief was strongest considering information from the most recent previous trial, with classification performance decreasing with longer time windows. Furthermore, we found that both confidence and awareness states, relative to task correctness, are more relevant for the classification of the prospective belief. However, this type of prospective beliefs did not appear to play a functional role in behavioural performance. Both perceptual sensitivity and metacognitive sensitivity were similar regardless of whether the prospective belief of success was high or low. This finding seems at odds with the possibility that prospective beliefs are associated with cognitive processing changes deployed to meet the belief (i.e. the 'self-fulfilling prophecy').[13,14] This raises the question of whether a different type of prospective decision (e.g. deciding whether to invest more attention on the current trial) may be predicted by the same information pattern that predicts prospective beliefs of success, and whether this type of prospective decision to engage with the environment may play a functional role in behavioural performance. Experiment 2 was designed with this in mind.

Experiment 2: Decisions to engage attention

Experiment 2 was similar to Experiment 1 except that here participants reported their decision to engage in a more or less focussed attentional state on the upcoming trial, instead of making a prospective belief of their performance success. The goal of Experiment 2 was to test whether or not the findings of Experiment 1 generalize to this new context.

Methods

Participants

Following informed consent, nineteen healthy volunteers (9 males and 8 females) took part in the experiment in return for monetary compensation. They were aged between 18 and 47 years (mean age 21.6). Three participants were excluded prior to data analyses due to the absence of aware trials, likely due to inadequate stimulus calibration. All participants were right-handed and had normal or

corrected-to-normal vision. They were naive to the experimental hypotheses and did not take part in Experiment 1. The study conformed to the Declaration of Helsinki and was approved by the West London Research Local Research Ethics committee

Experimental task and procedure

This was similar to Experiment 1 except that here, instead of reporting their belief of success, participants were instructed to report their decision to engage a focussed attention state in the upcoming trial (low vs high).

Machine learning protocols

These were similar to Experiment 1 except that here we employed the logistic regression to predict the decision to engage focussed attention (i.e. low vs high).

Results

Classification results

First, we report the pairwise correlations between the features used for classification: (i) awareness and confidence: 0.258 +/- 0.255, p = 0.0002; (ii) correctness and awareness: 0.267 +/- 0.117, p = 0.0001; (iii) correctness and confidence is 0.201 +/- 0.157, p = 0.0001. Although the correlations are higher than chance, the correlations are rather small. Tthere is room therefore for each variable to provide distinctive information for the classification of the decision to engage attention.

We then employed a logistic regression classifier to predict the decision to engage attention on a given trial based on awareness, confidence, and correctness features from the preceding trials. We found that the particular decision to engage a focussed state of attention could be significantly predicted above chance levels using information from the previous trials. As shown in Figure 5, classification accuracy was the highest accuracy when information from the previous trial was used. Prediction accuracy dropping close to chance level based on information from 3 trials back. P-values following the permutation analyses for the different time windows were as follow: 1-back: $p < 0.0025$, 2-back: $p < 0.0081$, 3-back: $p = 0.268$, 4-back: $p > 0.99$.

Next, we assessed the relevance of each of the different attributes (awareness, confidence, and correctness) for the classification. Similar to Experiment 1, we performed an ANOVA over the odd ratios of the logistic regression with time window (1,2, 3 and 4 trials back) and attributes as factors. The results showed a main effect of time window, $F(3,45) = 3.224$, $p = 0.0313$, $\eta^2 = 0.043$, but there was no main effect of attributes, $F(2,30) = 2.052$, $p = 0.146$, $\eta^2 = 0.023$, and no interaction, $F(4,60) = 0.97$, $p = 0.45$, $\eta^2 = 0.023$. However, Figure 6 shows that at least when observers rated high confidence on the previous trial, the probability of a decision to engage a highly focussed attention state on the

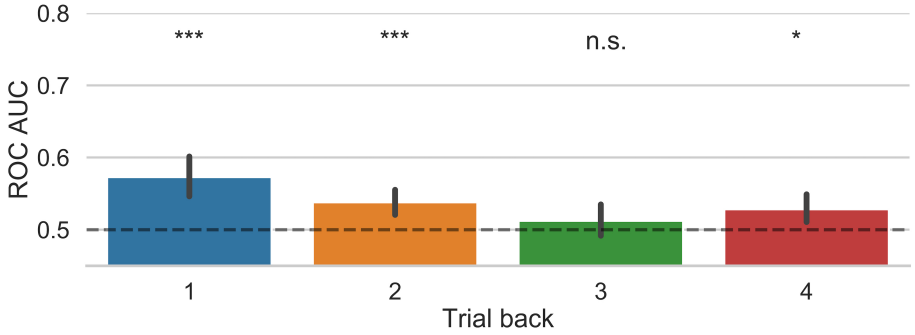

Figure 5: Results from the logistic regression classification model of the decision to engage attention. A value of 1 indicates that the odds of predicting the class is the same regardless of the feature value. Error bars represent bootstrapped 95% confidence intervals.

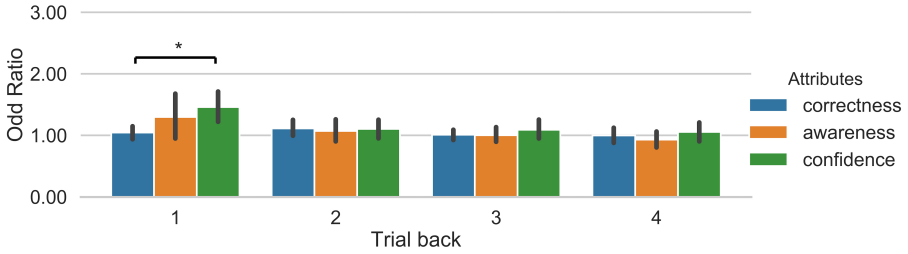

Figure 6: Illustration of odd ratios from the logistic regression for the different features involved in the classification of the decision to engage attention. Error bars represent bootstrapped 95% confidence intervals.

next trial increased by 1.45, which was significantly different from correctness (p = 0.03, corrected).

This pattern of results was replicated using analyses based on a random forest classifier (see Supplemental Figures 4 and 5).

Signal detection results

A 2 by 2 ANOVA with awareness (unaware, aware) and decision to engage a focussed state of attention (low, high) as factors was carried out on the observer's perceptual sensitivity to the orientation of the Gabor. There was an effect of awareness on type-1 d' ($F(1,15) = 69.53$, $p < 0.000000515$, $\eta^2 = 0.462682189$). Perceptual sensitivity was higher on aware relative to unaware trials. There was no effect of the decision to engage attention on type-1 d' ($F(1,15) = 1.388$, $p = 0.257$, $\eta^2 = 0.0054$) and no interaction effect between decision to engage and awareness ($F(1,15) = 1.877$, 0.191, $\eta^2 = 0.00216$).

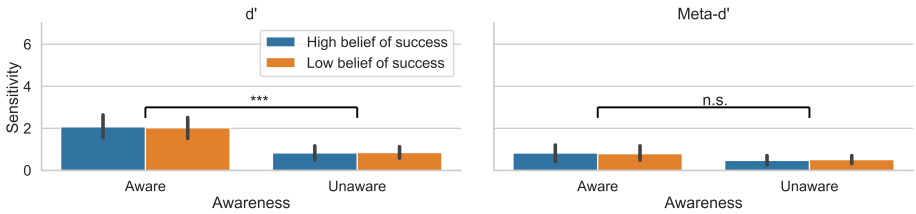

Figure 7: Perceptual sensitivity and metacogntive sensitivity scores as function of the state of awareness and the decision to engage attention. Error bars represent standard error of the mean.

A similar ANOVA was carried out on metacognitive sensitivity scores. There was an effect of awareness on meta-d' (F(1,15) = 7.084, p = 0.0178, $\eta^2 = 0.065$). Metacognitive sensitivity was slightly better on aware relative to unaware trials, though it remained significantly above chance in the unaware trials (i.e. t(15) = 2.88, p $<$ 0.011 and t(15) = 3.47, p $<$ 0.003, for both decisions to attend). There was also an effect of the decision of engage attention on meta-d' (F(1,15) = 8.824, p = 0.00953, $\eta^2 = 0.0631$). Metacognitive sensitivity was better following decisions to engage a more focussed attention state. There was also an interaction between the factors (F(1,15) = 5.279, p = 0.0364, $\eta^2 = 0.027$). Pairwise comparisons showed that on aware trials, meta-d' was better when participants decided to adopt a more focussed attention state (t(15) = 2.79, p $<$ 0.014). This was not the case when the participants reported being unaware of the target (t(15) = 1.772, p = 0.097). These results are depicted in Figure 7. This result indicates that the decision to attend had an effect on metacognitive sensitivity. We note, however, there were no effects of the decision to engage attention M-ratio (F(1,15) = 2.55, p=0.13). There was also no interaction between the decision to engage and the state of visual awareness (F(2,15)=0.502, p=0.48). However, we believe it is unlikely that the influence of the decision to engage attention in meta-d' can be explained by differences in d'. We shall come back to this point in the Discussion.

Discussion

The results showed that decisions to engage a focussed attentional state do not seem to affect perceptual sensitivity, at least in the current experimental conditions with single targets and no competing distracters. However, decisions to engage in a more focussed attention state influenced the observer's metacognitive sensitivity. Although there was no effect of the decision to engage attention on M-ratio, it is unlikely that the effect of the decision to engage in meta-d' can be accounted for in terms of differences in d' given that these were only minimal and statistically insignificant. Most critically, previous research indicates that d' and meta-d' can be sensitive to different factors. For instance, using a similar paradigm to the one used in this study, we have previously shown that

visibility has a strong effect on d' but not on meta-d',[11] indicating that meta-d' and d' are not necessarily based on a similar process, or operate on a similar type of information (see also[21]). Hence the application of M-ratio in the current experimental context seems not optimal. In any case, the data indicate that the decision to engage attention had a functional role in behavioural control, while this was not the case for prospective beliefs of success in Experiment 1.

On the other hand, the present results replicate the findings of Experiment 1 concerning the prediction of the future prospective belief of success, but crucially in a new decision context related to the attention state for the upcoming task. Intriguingly, we observed that high confidence in the previous trial predicted decisions to engage a more focussed attentional state in the next trial. Yet the opposite might be argued, namely, that when observers have held low confidence or have been incorrect, they would then decide to engage more focussed attention on the next trial. This is because errors or low confidence may in principle prompt subjects to be more cautious and try to boost attention to improve performance in the next trial. However, this is not always the case[22] and it has been argued that it may depend on the type of error.[23,24]

We now turn to integrate the findings from the two experiments.

Across-experiment Generalization Results

Finally, we analyzed the data from both experiments together in order to estimate how much information learned from one experiment can be transferred to the other experiment. We fitted the models with the features in one experiment (i.e. prospective of belief of success), and then test the trained model in the other experiment. To estimate the variance of the cross-experiment generalization, we conducted a cross-validation procedure as follows. First, we selected one of the experiment as the source and the other experiment as the target. Second, we preprocessed both the source and target experiment as described in the preprocessing section. Data from all the subjects were concatenated as one whole dataset. Third, the cross-validation method described above was applied to both the source data and the target data, while the training data was sampled from the source data and the testing data was sampled from the target data. Particularly, in each fold, 80% of the source data formed the training set and 20% of the target data formed the testing set. With 100 iterations of the cross-validation procedure, we estimated the variances of the transfer learning given N-back trials (N = 1, 2, 3, 4).

We first trained the classifiers using the data from the prospective belief of success (Experiment 1) and tested the classifiers using the data from the prospective decision to engage attention (Experiment 2). We found that the logistic regression model was able to decode the decision to attend based on the pattern of awareness, confidence, and correctness in the previous trial of the belief of success experiment ($p = 0.0215$), but not with the attributes in 2, 3, or 4 trial back ($p > 0.9$).

We then trained the classifier using the data from the prospective decision to

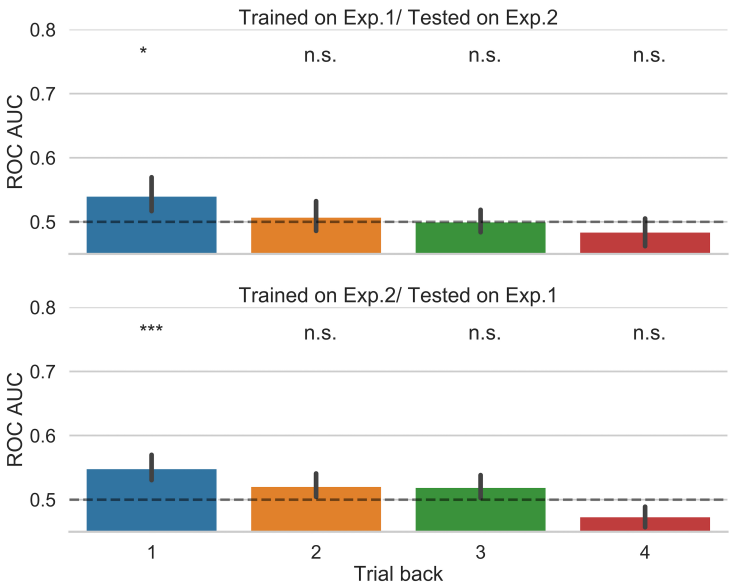

Figure 8: Across-experiment generalization results. Classifiers were trained on data from either experiment and test on the other experiment. Error bars represent bootstrapped 95% confidence intervals.

engage attention (Experiment 2) and tested the classifier using the data from the prospective belief of success (Experiment 1). The classifier was able to decode the prospective belief of success based on the pattern of awareness, confidence, and correctness in the previous trial of the decision to attend experiment ($p = 0.0215$), but this was not the case for $N = 2$ or 3 or 4 trials back (lowest p-value $= 0.0739$).

These results are depicted in Figure 8. This pattern of results was replicated using analyses based on a random forest classifier (see Supplemental Figure 6).

We also performed a cross-experiment comparison on the effect of the prospective decision on metacognitive sensitivity. Recall that in Experiment 1 we did not find that the prospective belief of success had any influence on type-1 perceptual sensitivity or type-2 metacognitive performance. To quantify whether the result of Experiment 2 was significantly different from Experiment 1 we compared the pattern of metacognitive sensitivity across experiments. In particular, we used the aware trials - in which the effect of the decision to engage attention was found- to compute the difference in meta-d' following decisions to adopt a more focussed relative to a less focussed attention state. The same was done in Experiment 1 for the high vs low prospective belief of success. An unpaired t-test showed that the influence of the decision to attend on meta-d' was higher than the influence of the prospective belief of success ($t(28)=2.06$, $p=0.048$).

General Discussion

We sought to understand the sources of information determining prospective decision making in the context of a perceptual task using standard machine learning techniques. Across two experiments, we found that a logistic regression classifier significantly predicted the upcoming prospective belief of success based on the pattern of awareness, confidence, and correctness exhibited in previous trials. Information from the previous trial led to the highest accuracy in the prospective belief of success, with classification accuracy increasingly dropping with up to four trials back. This finding is consistent with prior work.[9] The present study goes beyond the prior work to show that this pattern of results generalises to a different type of prospective decision, namely, individual decisions to engage with the task. Accordingly, we found that prospective decisions to adopt a focussed attention state were dependent on the prior history of task confidence and visual awareness. Across the two experiments, we observed that task correctness was less important than both awareness and confidence for the classification of both prospective decisions. This is also in keeping with the prior study by Fleming and colleagues (2016).

Previous research has shown that components of retrospective confidence estimates (e.g. mean, variance) are highly correlated across testing sessions involving a similar experimental task, although the generalization of confidence components across different task contexts was weaker.[25] Additional studies support the view that the precision of retrospective metacognitive judgments correlates across perceptual and memory tasks[26,27] and across sensory modalities.[28] Other studies have shown that observers use a similar confidence scale for different tasks of the same or different modality[29,30] and that confidence estimates on a given trial of a task carry-over to subsequent trial of a different task.[31] Beyond a theoretical framework based on confidence, the present study indicates that an internal model based on the recent history of experiential states of visual awareness and response confidence determines future prospective decision making. This can happen regardless of the type of prospective judgment (i.e. prospective belief of success or a decision to engage attention). This conclusion is also supported by the finding that a classifier trained on data from one experimental context (e.g. involving prospective beliefs) could be used to predict a different prospective decision (i.e., to engage attention) and vice-versa. However, this generalization only occurred when data from just the most recent trial was considered.

However, despite the common dynamic information pattern underlying the formation of prospective judgments, only the prospective decisions to attend had a functional role in behavioural performance. Retrospective metacognitive sensitivity improved following decisions to engage in a focussed attention state, but this was not the case following a prospective belief of high success. From the perspective of the 'self-fulfilling prophecy', prospective beliefs of success may set an expectation concerning upcoming behavioural performance that the participant is motivated to meet[13,14] and accordingly, a belief of performance success might in principle encourage observers to invest more cognitive resources

in the upcoming trial. Our results suggest that this is not the case. Despite the commonalities in the way that these seemingly different prospective decisions are formed, decisions to engage attention appear to effect a change in the cognitive system with consequences for behavioural control, while prospective beliefs of success do not.

Prospective beliefs of success concerned here a low-level perceptual discrimination task. It is possible that prospective beliefs are more diagnostic of the upcoming behavioural performance in different task domains, namely, memory.[4,32] Another possibility is that decisions to engage attention are more likely to be embodied by comparison to beliefs of success. Accordingly, recent theoretical frameworks borrowing from ecological psychology[33] propose that perceptual biases and decisions are not independent of action. In this framework, perception drives decisions and action, but actions also drive subsequent experiences in a dynamic perception-action loop.[34] We propose that decisions to engage with the environment (i.e. to deploy a focussed attention state) are more likely to be embodied in the action system and hence are very likely to trigger commitment towards that action, while prospective beliefs may not. It is possible that decisions to engage attention trigger preparatory control which in turn can influence subsequent cognitive processing. Decisions to engage attention influenced the precision of self-evaluation of the correctness of perceptual decisions but had no effect on perceptual processing itself, however. The latter is likely due to the absence of visual competition in the displays with single targets at central fixation. It is well known that attention effects on visual processing are stronger when there is space- or object-based competition for selection.[35]

In summary, this study indicates that a common representational structure supports the dynamic formation of seemingly different types of prospective judgements. Additional research is however needed to test whether these observations generalize to different task contexts and cognitive domains beyond perceptual decision making, including those in which the precision of retrospective metacognition is not correlated across tasks (e.g.[36]).

Acknowledgements

D.S. acknowledges support from the Spanish Ministry of Economy and Competitiveness, through the 'Severo Ochoa' Programme for Centres/Units of Excellence in R & D (SEV-2015-490) and project grants PSI2016-76443-P from MINECO and PI-2017-25 from the Basque Government.

Data availability statement

Data and code are available at https://github.com/adowaconan/metacognition.

Author contributions

N.M. performed data analyses and wrote the paper; S.R. and E.O. run the experiments and contributed towards data analyses; D.S. designed the study, wrote the paper, and supervised all aspects of the project.

Competing interests statement

The authors declare no competing interests.

References

[1] T. Nelson, L. Narens, and G. Bower, "The psychology of learning and motivation," *Metamemory: A theoretical framework and new findings*, 1990.

[2] S. M. Fleming, J. Ryu, J. G. Golfinos, and K. E. Blackmon, "Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions," *Brain*, vol. 137, no. 10, pp. 2811–2822, 2014.

[3] K. Miyamoto, T. Osada, R. Setsuie, M. Takeda, K. Tamura, Y. Adachi, and Y. Miyashita, "Causal neural network of metamemory for retrospection in primates," *Science*, vol. 355, no. 6321, pp. 188–193, 2017.

[4] T. O. Nelson and J. Dunlosky, "When people's judgments of learning (jols) are extremely accurate at predicting subsequent recall: The 'delayed-jol effect'," *Psychological Science*, vol. 2, no. 4, pp. 267–271, 1991.

[5] A. Koriat, "Monitoring one s own knowledge during study: A cue-utilization approach to judgments of learning.," *Journal of Experimental Psychology: General*, vol. 126, no. 4, pp. 349–370, 1997.

[6] T. O. Nelson and R. J. Leonesio, "Allocation of self-paced study time and the "labor-in-vain effect."," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 14, no. 4, pp. 676–686, 1988.

[7] J. Metcalfe and B. Finn, "Evidence that judgments of learning are causally related to study choice," *Psychonomic Bulletin & Review*, vol. 15, pp. 174–179, feb 2008.

[8] T. Gilovich, R. Vallone, and A. Tversky, "The hot hand in basketball: On the misperception of random sequences," *Cognitive psychology*, vol. 17, no. 3, pp. 295–314, 1985.

[9] S. M. Fleming, S. Massoni, T. Gajdos, and J.-C. Vergnaud, "Metacognition about the past and future: quantifying common and distinct influences on prospective and retrospective judgments of self-performance," *Neuroscience of Consciousness*, vol. 2016, no. 1, 2016.

[10] A. Koriat, "Metacognition and consciousness," in *The Cambridge Handbook of Consciousness* (P. D. Zelazo, M. Moscovitch, and E. Thompson, eds.), pp. 289–326, Cambridge University Press, 2007.

[11] B. Jachs, M. J. Blanco, S. Grantham-Hill, and D. Soto, "On the independence of visual awareness and metacognition: A signal detection theoretic analysis.," *Journal of experimental psychology: human perception and performance*, vol. 41, no. 2, p. 269, 2015.

[12] L. Charles, F. Van Opstal, S. Marti, and S. Dehaene, "Distinct brain mechanisms for conscious versus subliminal error detection," *Neuroimage*, vol. 73, pp. 80–94, 2013.

[13] R. Weinberg, D. Gould, and A. Jackson, "Expectations and performance: An empirical test of bandura's self-efficacy theory," *Journal of sport psychology*, vol. 1, no. 4, pp. 320–331, 1979.

[14] G. Zacharopoulos, N. Binetti, V. Walsh, and R. Kanai, "The effect of self-efficacy on visual discrimination sensitivity," *PloS one*, vol. 9, no. 10, p. e109392, 2014.

[15] J. W. Peirce, "Psychopy psychophysics software in python," *Journal of neuroscience methods*, vol. 162, no. 1-2, pp. 8–13, 2007.

[16] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.

[17] W. McKinney, "pandas: a python data analysis library," *see http://pandas. pydata. org/. Google Scholar*, 2015.

[18] J. King and S. Dehaene, "Characterizing the dynamics of mental representations: the temporal generalization method," *Trends in cognitive sciences*, vol. 18, no. 4, pp. 203–210, 2014.

[19] T. J. DiCiccio and B. Efron, "Bootstrap confidence intervals," *Statistical science*, pp. 189–212, 1996.

[20] B. Maniscalco and H. Lau, "A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings," *Consciousness and cognition*, vol. 21, no. 1, pp. 422–430, 2012.

[21] M. Zehetleitner and M. Rausch, "Being confident without seeing: What subjective measures of visual consciousness are about," *Attention, Perception, & Psychophysics*, vol. 75, pp. 1406–1426, jul 2013.

[22] W. Notebaert, F. Houtman, F. V. Opstal, W. Gevers, W. Fias, and T. Verguts, "Post-error slowing: An orienting account," *Cognition*, vol. 111, pp. 275–279, may 2009.

[23] J. van Driel, K. R. Ridderinkhof, and M. X. Cohen, "Not all errors are alike: Theta and alpha EEG dynamics relate to differences in error-processing dynamics," *Journal of Neuroscience*, vol. 32, pp. 16795–16806, nov 2012.

[24] A. Navarro-Cebrian, R. T. Knight, and A. S. Kayser, "Error-monitoring and post-error compensations: Dissociation between perceptual failures and motor errors with and without awareness," *Journal of Neuroscience*, vol. 33, pp. 12375–12383, jul 2013.

[25] J. Ais, A. Zylberberg, P. Barttfeld, and M. Sigman, "Individual consistency in the accuracy and distribution of confidence judgments," *Cognition*, vol. 146, pp. 377–386, 2016.

[26] J. Samaha and B. R. Postle, "Correlated individual differences suggest a common mechanism underlying metacognition in visual perception and visual short-term memory," *Proceedings of the Royal Society B: Biological Sciences*, vol. 284, p. 20172035, nov 2017.

[27] C. Song, R. Kanai, S. M. Fleming, R. S. Weil, D. S. Schwarzkopf, and G. Rees, "Relating inter-individual differences in metacognitive performance on different perceptual tasks," *Consciousness and Cognition*, vol. 20, pp. 1787–1792, dec 2011.

[28] N. Faivre, E. Filevich, G. Solovey, S. Kühn, and O. Blanke, "Behavioral, modeling, and electrophysiological evidence for supramodality in human metacognition," *The Journal of Neuroscience*, vol. 38, pp. 263–277, sep 2017.

[29] V. de Gardelle and P. Mamassian, "Does confidence use a common currency across two visual tasks?," *Psychological Science*, vol. 25, pp. 1286–1288, apr 2014.

[30] V. de Gardelle, F. L. Corre, and P. Mamassian, "Confidence as a common currency between vision and audition," *PLOS ONE*, vol. 11, p. e0147901, jan 2016.

[31] D. Rahnev, A. Koizumi, L. Y. McCurdy, M. D'Esposito, and H. Lau, "Confidence leak in perceptual decision making," *Psychological Science*, vol. 26, pp. 1664–1680, sep 2015.

[32] Y.-C. Kao, E. S. Davis, and J. D. Gabrieli, "Neural correlates of actual and predicted memory formation," *Nature neuroscience*, vol. 8, no. 12, p. 1776, 2005.

[33] J. J. Gibson, *The ecological approach to visual perception: classic edition.* Psychology Press, 1979.

[34] N. F. Lepora and G. Pezzulo, "Embodied choice: How action influences perceptual decision making," *PLOS Computational Biology*, vol. 11, p. e1004110, apr 2015.

[35] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annual review of neuroscience*, vol. 18, no. 1, pp. 193–222, 1995.

[36] W. L. Kelemen, P. J. Frost, and C. A. Weaver, "Individual differences in metacognition: Evidence against a general metacognitive ability," *Memory & Cognition*, vol. 28, pp. 92–107, jan 2000.