Title

Confidence Leak in Perceptual Decision Making

Journal

Psychological Science, 26(11)

ISSN

0956-7976

Authors

Rahnev, Dobromir
Koizumi, Ai
McCurdy, Li Yan
et al.

Publication Date

2015-11-01

DOI

10.1177/0956797615595037

Peer reviewed

HHS Public Access

Author manuscript

Psychol Sci. Author manuscript; available in PMC 2016 November 01.

Published in final edited form as:

Psychol Sci. 2015 November ; 26(11): 1664–1680. doi:10.1177/0956797615595037.

Confidence leak in perceptual decision-making

Dobromir Rahnev[1], **Ai Koizumi**[2], **Li Yan McCurdy**[3], **Mark D'Esposito**[1], and **Hakwan Lau**[4]

[1]Helen Wills Neuroscience Institute, University of California, Berkeley

[2]Department of Psychology, Columbia University, New York

[3]Interdepartmental Neuroscience Program, Yale University School of Medicine, New Haven

[4]Department of Psychology, University of California, Los Angeles

Abstract

We live in a continuous environment in which the visual scene changes on a slow timescale. It has been shown that, to exploit such environmental stability, the brain creates a "continuity field" such that objects seen seconds ago influence the perception of current objects. What is unknown is whether a similar mechanism exists at the level of our metacognitive representations. In three experiments we demonstrate a robust inter-task "confidence leak" that cannot be explained by response priming or attentional fluctuations. Observers' ability to modulate this confidence leak predicted higher capacity for metacognition as well as greater gray matter volume in the prefrontal cortex. A model based on normative principles from Bayesian inference explained the results by postulating that observers subjectively estimate the perceptual signal strength in a stable environment. These results point to the existence of a novel metacognitive mechanism mediated by regions in prefrontal cortex.

Introduction

For various kinds of decisions, we have the ability not only to use the available information to make a choice between several alternatives, but also to introspect on the likelihood that our decision is correct. This metacognitive ability is critical in deciding whether to act immediately on the decision or to continue gathering information, as well as whether to update our model of the world with the newly acquired information (Fleming, Dolan, & Frith, 2012; Koriat, 2007; Metcalfe & Shimamura, 1994; Nelson & Narens, 1990; Shimamura, 2000; Yeung & Summerfield, 2012). Research on metacognitive judgments in perception has typically focused on how confidence ratings are influenced by different properties of the stimulus at hand (Fleming & Lau, 2014). However, other factors beyond

Address correspondence to: Dobromir Rahnev, University of California, Berkeley - Helen Wills Neuroscience Institute, 10 Giannini Hall, Berkeley, California, 94720, Phone: 510-642-2839; Fax: 510-642-5293; drahnev@gmail.com.

Author Contributions

D. Rahnev and H. Lau developed the study concept and design. Testing and data collection were performed by D. Rahnev, A. Koizumi, and L.Y. McCurdy. D. Rahnev performed the data analysis and interpretation under the supervision of H. Lau and M. D'Esposito. D. Rahnev drafted the manuscript, and H. Lau, A. Koizumi, L.Y. McCurdy, and M. D'Esposito provided critical revisions. All authors approved the final version of the manuscript for submission.

Declaration of Conflicting Interests

Authors declare no conflict of interest.

the immediate stimulus also influence confidence ratings (Koriat, 2011). One such factor that has received little attention is inter-trial and inter-task influences on metacognitive judgments.

We live in a continuous environment in which the viewing conditions change relatively slowly and objects tend to persist within our visual field for long periods. It has been demonstrated that the brain exploits such environmental stability (Fischer & Whitney, 2014; Frund, Wichmann, & Macke, 2014; Liberman, Fischer, & Whitney, 2014; Zhang, Wang, & Goldberg, 2014). In particular, it has recently been proposed that the brain creates a "continuity field" such that recently seen objects bias the perception of current objects (Fischer & Whitney, 2014; Liberman et al., 2014).

Here we report evidence that such continuity field exists at the metacognitive level, and that there may be dedicated mechanisms specifically influencing our confidence ratings rather than basic perceptual processing. We demonstrate robust inter-task dependence in metacognitive judgments of confidence even when the two different tasks involve distinct visual features: letter identity versus color. In other experiments, serial dependence of confidence ratings is shown to appear when the same task is presented many times. Finally, we relate this effect to gray matter volume in the anterior prefrontal cortex (aPFC), an area that has previously been linked to metacognition. This phenomenon – which we call "confidence leak" – is explained by a computational model based on the principles of Bayesian inference. The model quantifies the normative principle that observers use their subjective certainty in previous judgments to predict the quality of the perceptual signal for future judgments.

Methods

Observers

Sixty-nine observers (39 women; mean age = 23.6; SD = 5.7) participated in three psychophysical experiments (Experiments 1–3). Experiment 1 consisted of 27 observers; 1 observer was excluded for chance performance. Experiment 2 consisted of 22 observers; 4 observers were excluded for having an extreme bias in the "opt out" task: two of them chose the opt out option on at least 396 trials (out of 400) while the other two chose to opt out at most 5 times. Finally, Experiment 3 was a re-analysis of a previous 20-subject dataset reported in Rahnev et al. (2011). For all 3 studies we sought to collect data from between 20 and 30 observers as in previous work from our laboratory (Maniscalco, Bang, Iravani, Camps-Febrer, & Lau, 2012; Rahnev, Lau, & De Lange, 2011) stopping data collection when the sample size reached the target interval; logistical factors led to slightly different sample sizes. No statistical analyses reported in the paper were performed on partial data. In addition, in Experiment 4 we re-analyzed the data from McCurdy et al. (2013), which included 34 observers. All participants were naive regarding the purposes of the experiments, had normal or corrected-to-normal vision, and signed an informed-consent statement approved by the local ethics committee. Each experiment took approximately one hour to complete and participants were compensated at the rate of $10/hour.

Stimuli and Task / Materials and Procedure

Stimuli in Experiments 1–3 were presented on a gray background (6.0 cd/m^2). Observers were seated in a dimmed room about 60 cm away from the computer monitor. Stimuli were generated using Psychophysics Toolbox in MATLAB (MathWorks, Natick, MA) and were shown on an iMac monitor (19 inch monitor size, 1680 × 1050 pixel resolution, 60 Hz refresh rate).

Experiment 1—On each trial 40 characters were displayed. The characters were X's and O's that were colored in red and blue. The letter identity and color were independent of each other. Observers' task was to judge the dominant letter identity and color in the display (Figure 1). More specifically, with four different button presses observers indicated (1) whether there were more X's or more O's, (2) the confidence in their decision on 1–4 scale, (3) whether there were more red or blue characters, and (4) the confidence in their decision on 1–4 scale. The order of the four questions was the same for all observers. To give their responses, observers used the 1–4 keys on a computer keyboard.

The dominant letter always accounted for 23 of the 40 characters on the screen. In different runs, the dominant color accounted for either 23 or 29 characters. The dominant letter and color were pseudo-randomized such that over the course of the experiment the X's and O's, as well as the red and blue color, were dominant equally often. We used the same objective difficulty level for all observers since achieving very similar level of performance across observers was not critical. Conversely, the use of a staircasing procedure could have led observers to believe that a correlation structure was actually present in the task (even though such structure did not exist in the main experiment), thus potentially compromising our analyses on confidence leak.

All letters were presented in Arial font (size = 0.5°) and placed randomly within an imaginary square centered on fixation with a side of 10°. The letters remained on the screen for one second, after which each of the four questions was presented on the screen in succession. Observers were allowed to take as long as they needed to give their responses.

Participants completed a total of 400 trials separated in 4 runs, each consisting of 4 blocks of 25 trials. At the end of each block observers were given 15-second breaks, while at the end of each run they were allowed to take self-paced breaks.

Observers were given a total of 46 practice trials. The initial practice trials were easier and the difficulty was gradually increased until it reached the difficulty of the actual experiment. Trial-by-trial feedback was provided on the first 36 trials in order to help observers learn how to perform the task. There was no feedback in the last 10 practice trials or in the main experiment.

Experiment 2—Experiment 2 was similar to Experiment 1. The main difference was the method of collecting confidence ratings. In Experiment 1 confidence ratings were always collected on the same 1–4 scale. In Experiment 2, in order to minimize the chance of motor priming between confidence reporting in the two different tasks, we used two separate methods of determining the confidence level that were designed to be as different as

possible from each other. The first reporting method required observers to use the mouse to make a subjective confidence response on a continuous visual analog scale (VAS), while the second was the "opt out" paradigm (Kiani & Shadlen, 2009) in which observers are given the option of not responding if they are unsure about the response (but do not give confidence explicitly).

In the first method, observers gave the confidence rating on the VAS scale by moving a mouse between the two extremes of a straight line. The left extreme of the line was marked as "Not confident at all" and was coded as confidence of 0, while the right extreme was marked as "Very confident" as was coded as confidence of 100. For each observer, this VAS rating followed either the letter identity or color task but this pairing was randomized between observers.

Second, on the other task observers were given the option of choosing either of the two possible responses (either X/O or red/blue). In addition to these two options, a third possibility presented was to choose to "opt out". The response to this task was provided using a computer keyboard. The decision to opt out was coded as confidence of 1, while giving a response was coded as confidence of 2. In order to make the "opt out" choice meaningful, we introduced a point system. Choosing to opt out resulted in 2 points guaranteed, while choosing to respond led to either 4 points (for correct answers), or −1 points (for incorrect answers). The optimal strategy in this task is thus to choose to "opt out" when the probability of being correct is less than 66%. To decrease inter-observer variability, we explicitly informed observers of the optimal strategy. In order to increase the consistency between the two tasks, observers were awarded 4 points for correct and −1 points for incorrect answers in the other task (i.e., the task which was followed by VAS rating). To motivate participants to use the opt-out option optimally, we rewarded the three observers with highest scores across the whole experiment with additional $10.

To remove any other biases potentially present in Experiment 1, we randomized the order of the questions between the observers such that about half of them would always respond first to the letter identity task, while the other half responded first to the color task. Further, in Experiment 2 both tasks had two difficulty levels (the dominant letter identity or color was present in either 23 or 26 of the 40 characters) that were completely randomized between trials.

Experiment 3—Experiment 3 was performed to investigate whether confidence leak depends on the quality of the perceptual signal. It was originally reported in the Supplementary Material in Rahnev et al. (2011). There we focused on comparing the difference in confidence for high vs. low attention conditions. We returned to this dataset to reanalyze it in terms of confidence leak.

In that experiment we varied the number of stimuli (Gabor patches) on the screen in order to manipulate how observers distribute their attention to different objects. In one condition we used 2 items on the screen (a relatively focused mode of attention), while in the other we used 4 items on the screen (a relatively distributed mode of attention). The stimuli were presented for 33 ms (two computer frames). After a delay of 500 ms, observers saw a

response cue that instructed them which stimulus they should respond to (see Figure 5A). Observers had to indicate the tilt (clockwise/counter-clockwise) of the Gabor patch and rate their confidence (high/low). Participants completed 8 blocks of 125 trials each for a total of 1000 trials. Within each block there were always either 2 or 4 patches and a single contrast level (4, 6, 8, 10, or 12%).

Experiment 4—Experiment 4 was re-analysis of the data from McCurdy et al. (2013). All experimental details are included in the original publication. Very briefly, observers completed a 2AFC task in which they indicated which of two noisy stimuli located on the left and right of fixation contained a grating. They then provided a confidence rating on 1–4 scale. Observers completed 510 trials separated in 5 blocks of 102 trials each.

Analyses

Experiment 1—To determine whether confidence "leaks" from one task to the other, we first performed, for each observer, a simple trial-by-trial correlation of the confidence ratings from the two tasks. Next, in order to control for the influence of other factors we performed a regression in which the confidence on the letter identity task was used to predict the confidence on the color task while at the same time controlling for the influence of accuracy and RT on each task, as well as difficulty of the color task (i.e., whether the dominant color was present in 23 or 29 characters).

One possible reason for a confidence leak is simple motoric priming. To control for that possibility, we analyzed all trials in which an observer gave a confidence of 1 to one of the tasks but not 1 on the other task. This allowed us to determine, for each observer the percent of confidence ratings of 2, 3, and 4 on one task paired with a confidence rating of 1 on the other task. Similarly, we determined the percent of confidence ratings of 1, 2, and 3 on one task paired with a confidence rating of 4 on the other task. We excluded any observers that did not have at least 5 trials that entered in either of these analyses because that would lead to excessively volatile estimates. This led to the exclusion of 6 observers in this analysis. Finally, we performed a repeated-measures ANOVA to test whether there was an interaction of the amount of 2, 3, and 4 confidence ratings paired with a confidence of 1 and the amount of 1, 2, and 3 confidence ratings paired with a confidence of 4.

Experiment 2—We performed correlation and regression analyses as in Experiment 1. In addition, for visualization purposes, we created probability density functions (PDFs) for the confidence on the VAS scale conditional on the confidence on the opt-out task. The individual PDFs were averaged, and, for visualization, smoothed with a 10-point window.

Experiments 3 and 4—Since observers completed only one task per trial, confidence leak in these experiments was determined by analyzing the temporal lag-1 autocorrelation of the confidence ratings. Similarly to Experiments 1 and 2, both correlation and regression analyses were performed.

To determine observers' performance on the task, we computed the signal detection theory (SDT) measure *d'* by calculating the hit rate (HR) and false alarm rate (FAR). Then,

$$d' = \Phi^{-1}(HR) - \Phi^{-1}(FAR) \quad (1)$$

where Φ^{-1} is the inverse of the cumulative standard normal distribution that transforms HR and FAR into *z*-scores. The measure *d'* reflects the signal-to-noise ratio for observers performing the task.

Bayes factors—We computed Bayes factors for confidence leak in each experiment. We employed the Bayes calculator in Dienes (2008) for which we used a wide prior distribution for the alternative hypothesis defined as a Half-Normal distribution with a mean of 0, and SD of 1. Values higher than 3 are usually regarded as providing substantial evidence for the alternative hypothesis (Dienes, 2008, 2014). We obtained very large Bayes factors and the results were insensitive to variations in the distribution used for the alternative hypothesis.

Metacognition—To understand the relationship between confidence leak and metacognition, we computed a standard measure for metacognition, the area under the Type 2 ROC curve (Type 2 AUC; Fleming & Lau, 2014; Fleming et al., 2010). The Type 2 ROC curve is similar to a conventional ROC curve with the difference that hits are defined as high confidence correct trials, while false alarms are defined as high confidence incorrect trials. Type 2 AUC was computed for each of the two tasks in Experiment 1 (the average of the two was taken as the observer-specific metacognition score; similar results were obtained if both scores were considered separately). In Experiment 2 the Type 2 AUC score was computed for the task with confidence rating because the opt-out task did not allow us to determine the accuracy of the low-confidence decisions (since a decision to "opt out" meant that observers did not indicate the perceived stimulus). For this analysis VAS confidence scores were transformed into 1–4 scores using cutoffs defined on the 25^{th}, 50^{th}, and 75^{th} percentile of VAS scores for each individual observer. Finally, in Experiment 3 Type 2 AUC was computed taking all trials from the experiment.

We then correlated, across all observers from Experiments 1–3, the amount of confidence leak, defined as the Fisher-transformed correlation values from each experiment, with the observer-specific metacognition (defined as the Type 2 AUC value). To increase power, the correlation was performed on the combined data from the three experiments.

Voxel-based morphometry (VBM)—We repeated the analyses by McCurdy et al. (2013) except that we were interested in the brain correlates of confidence leak rather than metacognition scores. Briefly, we used the same preprocessing which included segmentation into gray matter, white matter, and CSF in native space, aligning the data from different observers, registering to Montreal Neurological Institute stereotactic space, and smoothing with an 8 mm full-width at half-maximum Gaussian kernel. Multiple regression was used with an initial threshold of $p < 0.001$ uncorrected to determine the brain regions that correlated with our measures of confidence leak. Gender was included as a covariate and proportional scaling was used to account for global brain volume variability across participants. As in McCurdy et al. (2013), small-volume correction was applied to the clusters of interest in the prefrontal cortex by defining a 10 mm sphere at the peak voxel coordinates presented by Fleming et al. (2010) that were found to be associated with their

measure of metacognitive capacity [left aPFC, (−20, 53, 12); right aPFC, (24, 65, 18), (33, 50, 9); dorsolateral PFC, (36, 39, 21)].

Modeling framework

In order to explain the confidence leak phenomenon, we developed a single-parameter model and fitted it to the data from the first three experiments. The model was based on the principles of Bayesian inference and is an extension of our previous work on confidence generation (Rahnev, Bahdo, de Lange, & Lau, 2012; Rahnev et al., 2013; Rahnev, Maniscalco, et al., 2011; Rahnev, Maniscalco, Luber, Lau, & Lisanby, 2012)

Bayesian theory describes the optimal way to choose between different alternatives in the space of possible stimuli. Observers attempt to make a decision based on the posterior probability $P(S_i|E = x)$ where S_i are the possible stimulus categories and E is the evidence on the current trial. Bayes' theorem describes how this posterior probability $P(S_i|E = x)$ should be computed based on the likelihood function $f_{E|S_i}(x)$, the prior probability $P(S_i)$, and the density function of E, $f_E(x)$:

$$P(S_i|E=x)=\frac{f_{E|S_i}(x)*P(S_i)}{f_E(x)} \quad (2)$$

For all perceptual decisions in the current set of experiments, there were two possible stimulus categories S_1 and S_2 (in Experiments 1 and 2 these were either X/O or red/blue, while in Experiment 3 these were clockwise/counter-clockwise tilt). Since we informed observers that the two stimuli categories were equally likely in all experiments, we set the prior probabilities for the two stimulus categories to 0.5. Further, since there are only two stimulus categories, we obtain that $P(S_2|E) = 1 - P(S_1|E)$. Thus, observers' choice was made by simply comparing $P(S_1|E)$ and $P(S_2|E)$ and choosing the stimulus that corresponds to the higher posterior probability. In keeping with previous literature (Macmillan & Creelman, 2005), the likelihood functions $f_{E|S_i}(x)$ are assumed to be Gaussian distributions with equal SD that can be set to 1, such that:

$$f_{E|S_i}(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\mu_i)^2}{2}} \quad (3)$$

where μ_i is the mean of the likelihood function $f_{E|S_i}(x)$. For simplicity, and without loss of generality, we set the two means to be symmetric about the origin, that is $\mu_1 = -\mu$, $\mu_2 = \mu$, where μ is positive. Notice that, with these assumptions, *d'* – the standard measure of capacity from signal detection theory – can be expressed as a function of μ:

$$d'=2*\mu \quad (4)$$

Using equation (3), we obtain the following for the posterior distribution $P(S_2|E = x)$ (for full derivation, see Supplementary Material):

$$P(S_2|E=x)=\frac{1}{e^{-2x\mu}+1} \quad (5)$$

How do confidence ratings behave in this framework? According to the principles of Bayesian inference, confidence ratings should be determined by the probability of being correct. Thus, for example, if an observer is given the option of giving low/high confidence, then she would make this choice based on the value of $max[P(S_1|E)]$ and give a high confidence rating if this posterior probability is higher than a threshold (e.g., 0.75) and a low confidence rating otherwise. In general, if observers are asked to produce N discrete confidence ratings, they would need to define N−1 different thresholds $t_1, t_2, \ldots t_{N-1}$, such that a confidence rating of k is given if $max[P(S_1|E)]$ is in the interval (t_{k-1}, t_k), where $t_0 = .5$ and $t_N = 1$ (Macmillan & Creelman, 2005).

The thresholds t_i are defined in the posterior probability space. However, observers can only infer the posterior probability space by assuming a particular shape of the likelihood functions $f_{E|S_i}(x)$. If this assumption is wrong, then the inferred posterior probability space would be incorrect. In this sense, observers do not have direct access to the posterior probability space. Therefore, it is helpful to translate the thresholds t_i from the posterior space to the likelihood space to which observers do have direct access since that space is independent of any assumptions on the shape of the likelihood functions or the exact priors. Such correspondence is fortunately straightforward: each threshold t_i corresponds to a criterion c_i defined in likelihood space, such that, using equation (5):

$$t_i=\frac{1}{e^{-2c_i\mu}+1} \quad (6)$$

from where it follows that:

$$c_i=\frac{log\left(\frac{t_i}{1-t_i}\right)}{2\mu} \quad (7)$$

The likelihood space confidence criteria $c_1, c_2, \ldots c_{N-1}$ are placed on the right of the origin in likelihood space such that a confidence rating of k is given if evidence mapped on the likelihood scale is in the interval (c_{k-1}, c_k), where $c_0 = 0$ and $c_N = \infty$. These confidence criteria reflect the confidence rating when S_2 is the more likely stimulus. In a symmetrical fashion, $-c_1, -c_2, \ldots -c_N$ are used for the confidence ratings when S_1 is the more likely stimulus.

With no trial-to-trial information about the likely state of the environment, an observer can simply fix $c_1, c_2, \ldots c_{N-1}$ based on constant prior and likelihood functions so that they correspond to the desired $t_1, t_2, \ldots t_{N-1}$. However, the environment has higher order structure that makes it predictable (David, Vinje, & Gallant, 2004) and the brain is able to take advantage of this predictability (Fischer & Whitney, 2014; Fründ et al., 2014; Liberman et al., 2014; Yu & Cohen, 2009). To account for such autocorrelation in the environment, an observer will need to update the likelihood functions $f_{E|S_i}(x)$ from trial to trial. If an

observer expects that the difficulty on the subsequent trial is now described by a likelihood function $f'_{E|S_i}(x)$ with a mean of μ', then she ought to update her criteria c_i to c'_i such that:

$$c'_i = \frac{log\left(\frac{t_i}{1-t_i}\right)}{2\mu'} = c_i * \frac{\mu}{\mu'} \quad (8)$$

In other words, if an observer expects *d'* to change by a factor of *F*, then to remain Bayes-optimal, she needs to multiply all criteria, defined in the likelihood space, by a factor of $\frac{1}{F}$ (using equation 8 and the linear dependency between *d'* and μ from equation 4).

Model fitting

We developed a formal model that is based on the Bayes-optimal strategy of adjusting confidence criteria based on the expected change in *d'*. The model had a single free parameter θ, while all other parameters were fixed.

For all experiments, we first computed *d'* and c_i for each task. The value of *d'* was computed using equation (1), while c_i were computed using the formula:

$$c_i = -\frac{1}{2} * [\Phi^{-1}(HR_i) + \Phi^{-1}(FAR_i)] \quad (9)$$

where HR_i is the proportion of correct responses produced with confidence of *i* or higher, while FAR_i is the proportion of incorrect responses produced with confidence of *i* or higher (Macmillan & Creelman, 2005).

To perform the model fitting, for each task, we generated random samples from the distributions of the stimulus categories S_1 and S_2 defined by equation (3). For each task *k* (*k* = 1 or 2, as there were two tasks), we set the means of $S_{1,k}$ and $S_{2,k}$ such that $\mu_{1,k} = -\frac{d'_k}{2}$ and $\mu_{2,k} = \frac{d'_k}{2}$. Since $\mu_{1,k}$ and $\mu_{2,k}$, as well as the criteria $c_{i,k}$, are symmetrical, without loss of generality we could simply sample from the $S_{2,k}$ distributions. The samples for the first task, $x_{j,1}$, were categorized into confidence ratings using the confidence criteria $c_{i,1}$. However, depending on the confidence rating produced on the first task, the criteria for the second task on the corresponding trials are shifted. The idea is that, on a trial-by-trial basis, based on the confidence in the first task, observers adjust their expectation for the signal-to-noise ratio (*d'*) in the second task such that:

$$d'_{expected} = d' * e^{\Delta C * \theta} \quad (10)$$

where Δ*C* is deviation from the average observer-specific confidence produced on the same trial in the first task and θ is a free parameter that controls the amount of adjustment in the expectation of *d'* in the second task based on the confidence on the first task. For example, if an observer has an average confidence of 2.5 on the first task, and on a particular trial she gave a confidence of 4 on that task, then Δ*C* = 4 − 2.5 = 1.5. If, for the same observer θ =

0.2, then $e^{\Delta C*\theta} = e^{1.5*0.2} = 1.35$. In other words, on this particular trial, $d'_{expected}$ for the second task would be 1.35 times higher than its normal value. Note that, $\theta = 0$ indicates no expectation for any change ($d'_{expected}=d'$), while $\theta > 0$ and $\theta < 0$ indicate expectation for an autocorrelated and anti-autocorrelated environment, respectively.

Using equations (4) and (8), it follows that, given the expectation for a change in d' from equation (10), the observer should update her confidence criteria on the second task such that:

$$c_{i\_\ new} = \frac{c_{i\_\ original}}{e^{\Delta C*\theta}} \quad (11)$$

For each observer in each experiment, the single free parameter θ was fitted by starting with a value of 0, and adjusting it in a stepwise fashion until we could reproduce the confidence correlation between the two tasks. For each step we simulated 100,000 trials using the above procedure, and adjusted the parameter θ based on the obtained confidence correlation: if the simulated correlation was smaller/larger than the observed correlation, θ was initially increased/decreased by a step of 0.2 and this step was decreased in half after each reversal until we obtained a fit of the confidence correlation value with error smaller than 0.001.

In Experiment 1, we computed, for each observer, d'_{letter} and the confidence criteria c_{1_letter}, c_{2_letter}, and c_{3_letter} for the letter identity task, as well as d'_{color} and c_{1_color}, c_{2_color}, and c_{3_color} for the color task. The trial-to-trial confidence responses on the letter identity task (first task) were used to adjust the confidence criteria on the color task (second task).

In Experiment 2, we computed, for each observer, d'_{VAS} and the confidence criteria c_{1_VAS}, c_{2_VAS}, … c_{100_VAS} such that confidence ratings from 0 to 100 could be produced. For the opt-out task we could not estimate d'_{optOut} for each observer due to reasons outlined above, so we set d'_{optOut} to the same value for each observer that was equal to the average d'_{VAS}. However, we did compute the single confidence criterion c_{opt_out} for each observer separately. The trial-to-trial confidence responses on the opt out task (first task) were used to adjust the confidence criteria on the VAS task (second task).

In Experiment 3, we computed, for each observer, $d'_{2-stimulus}$ and $c_{2-stimulus}$ for the 2-stimulus task, as well as $d'_{4-stimulus}$ and $c_{4-stimulus}$ for the 4-stimulus task. On each step we generated 50,000 trials for each of the 2- and 4-stimulus tasks (for a total of 100,000). For both the 2- and 4-stimulus tasks, the first trial was generated using the parameters above, while for trials $i \geq 2$, we used the confidence response on trial $i-1$ (equivalent to "first task" above) to adjust the confidence criteria on trial i (equivalent to "second task" above). We fitted the model to the average confidence autocorrelation value across the two tasks but also examined the obtained confidence autocorrelation within each task.

Data and code

All data and codes for the analyses and model fitting in Experiments 1–3, as well as the behavioral analyses in Experiment 4 are freely available online: at http://dx.doi.org/10.5281/zenodo.18396.

Results

Experiment 1

We sought to determine whether confidence "leaks" from one task to another even when the two tasks are about separate visual features such as shape and color. To this end we created a stimulus consisting of X's and O's that were colored in red or blue and asked observers to determine the dominant letter identity and color, as well as to provide confidence ratings related to both of these decisions (Figure 1).

Average accuracy was 68% (SD = 6.1%) for the letter identity task, 68% (SD = 8.6%) in the difficult condition and 88% (SD = 10.5%) in the easy condition of the color task, respectively. Thus, our task was successful in inducing appropriate behavioral performance. Achieving a pre-specified level of performance was not critical in this study.

To check for confidence leak, we first performed, for each observer, a simple trial-by-trial correlation of the confidence ratings from the two tasks. The inter-task confidence correlation was positive for 25 of the 26 observers (Figure 2A), and was significantly positive ($p < .05$) for 21 of the 26 observers. Across the whole group, the average Fisher-transformed correlation coefficient was .23 with 95% confidence interval [.17, .29], which is significantly positive ($t(25) = 7.40$, $p = 9*10^{-8}$) and corresponds to an effect size (Cohen's $d = 1.45$) that is considerably higher than what is conventionally considered as "large" effect size ($d = .80$). Finally, we also computed the Bayes factor (see Methods) and found substantial evidence for the alternative hypothesis (Bayes factor = $9.4*10^{9}$).

Next, in order to control for the influence of other task factors, we performed a regression where the confidence on the letter identity task was used to predict the confidence on the color task while at the same time controlling for the influence of accuracy and RT on each task, as well as difficulty of the color task (i.e., whether the dominant color was present in 23 or 29 characters). After accounting for these other influences, confidence of the letter identity task was now a positive predictor of the confidence on the color task in all 26 observers. The mean beta value was .22 ($t(25) = 8.02$, $p = 2*10^{-8}$) with 95% confidence interval [.16, .27], signifying that one unit of confidence difference in the letter identity task predicted a .22 unit change in the confidence in the color task. This is also illustrated in Figure 2B where we plotted the average confidence on the color task for each confidence rating of the letter identity task and found a significant difference in their means ($F(3,25) = 32.12$, $p = 1.8*10^{-13}$).

Interestingly, neither accuracy nor RT on the letter identity task predicted confidence on the color task ($p = .74$ and .22, respectively). Therefore, what leaks is not the quality of the signal itself but observers' metacognitive assessment of the signal correlation.

These data suggest that confidence indeed leaks between tasks, even when these tasks depend on different visual features. However, there are several possible alternative explanations to rule out.

First, it could be that observers' attentional states varied over the course of the experiment, which could lead to fluctuations in signal quality that is similar for the two tasks. If signal quality correlates between the two tasks, then confidence would also correlate trivially. To test for this possibility, we computed the performance on the color task as a function of the accuracy on the letter identity task. After a correct response on the letter identity task average *d'* was 1.69, while after an incorrect response it was 1.62 with the difference failing to reach statistical significance ($t(25) = 1.03$, $p = .31$, Cohen's $d = .10$; Figure 2C). Further, this small, non-significant "performance leak" did not significantly correlate across subjects with the amount of confidence leak ($r = .19$, $p = .34$). Thus, confidence leak is not simply the result of fluctuations in signal quality because such fluctuations would have produced large corresponding changes in accuracy that would also correlate with confidence leak.

Second, it is possible that confidence ratings for both tasks drifted on a long timescale (e.g., they could have slowly increased for both tasks over the course of the experiment), reflecting perhaps a change in subjective mood state or degree of perceived expertise with the task. Such low-frequency drift could result in an observed confidence leak even without any within-trial leak. To check for that possibility, we applied scaled correlation (Nikolić, Mureşan, Feng, & Singer, 2012), which is a method to compute the correlation from a small window sizes thus excluding the slow-frequency components from contributing to the correlation. We used window sizes of 5, 10, 20, 50 100, 200, and 400 trials and found no influence of window size on inter-task confidence correlation ($F(6,25) = .64$, $p = .7$), suggesting that confidence leak does not depend on slow changes in the observers' overall cognitive or mood state.

Third, it could be that confidence for both tasks fluctuated on a very short time scale (e.g., on a time scale of less than 5 trials) and that this fluctuation was similar for both tasks but that there was no causal influence from one task to the other. To address this question, we took advantage of a particular feature in our design: while the letter identity task always had the same difficulty over the course of Experiment 1, the difficulty of the color task changed once every run (of 100 trials). If confidence did not leak from one task to the other (but simply fluctuated simultaneously for the two tasks), we would expect confidence on the letter identity task to be the same regardless of the difficulty of the color task. However, even though the accuracy on the letter identity task did not depend on the difficulty of the color task ($p = .48$), observers were significantly more confident on the letter identity task in the easy color blocks compared to the difficult color blocks ($t(25) = 2.65$, $p = .014$). That is, an *experimental manipulation* of task difficulty for the color task *led* to a significant difference in confidence in the letter identity task. We obtained a similar result in an additional control experiment in which observers judged the orientation of grayscale gratings and provided confidence ratings: intermediate-contrast gratings were judged with higher confidence when presented in the same block as high-contrast gratings than when presented in the same block as low-contrast gratings (Figure S1 in the Supplementary Material available online).

Finally, it is possible that observers were subject to motoric priming such that a certain button press used as a confidence response on one task was likely to be repeated for the next task. In order to exclude simple motoric priming as the cause of the observed results, we analyzed the trials in which different buttons were pressed for the two tasks in order to determine whether confidence leak is preserved when trials with the same motoric response are removed. We analyzed the trials in which confidence rating of 1 or 4 was given on one of the tasks, while a different confidence rating (i.e., not 1 or not 4, respectively) was given on the other. Among these trials, confidence rating of 1 was paired with a confidence of 2, 3, and 4 on 38%, 35%, and 27% of the trials, respectively, while confidence rating of 4 was paired with a confidence of 1, 2, and 3 on 24%, 30%, and 46% of the trials, respectively. A repeated-measures ANOVA confirmed that these patterns are significantly different ($F(2,19) = 4.6$, $p = .016$), which suggests a simple motoric priming effect cannot explain the data.

Experiment 2

We conducted Experiment 2 in order to obtain a replication of the confidence leak phenomenon, as well as to provide stronger evidence that the effects of Experiment 1 were not due to motoric or response priming. We kept the stimulus the same but made the two confidence ratings as different as possible: one of them was provided on a continuous scale, while the other one was collected in an implicit fashion through the “opt out” paradigm in which observers are given the option of choosing not to respond if they are not certain enough (Figure 1). Further, responses were provided with a mouse to the first task, and with a keyboard to the second task. This design minimized the effect of any possible motoric priming (see Methods).

We again performed, for each observer, trial-by-trial correlation of the confidence ratings from the two tasks. The inter-task correlation was positive for 17 of the 18 observers (Figure 3A), and was significantly positive ($p < .05$) for 13 of the 18 observers. Across the whole group, the average Fisher-transformed correlation coefficient was .26 with 95% confidence interval [.10, .42], which is significantly positive ($t(17) = 3.17$, $p = .006$) and corresponds to a large effect size (Cohen’s $d = 0.75$) and strong evidence for the alternative hypothesis (Bayes factor = 13.8). Controlling for accuracy, RT, and difficulty on both tasks in a regression analysis (as in Experiment 1) produced virtually identical results. As in Experiment 1, neither accuracy nor RT on one task predicted confidence on the other task ($p = .26$ and .72, respectively), suggesting that it is specifically the metacognitive assessment of the signal’s quality that leaks from one task to another.

To better visualize the magnitude of the effect, in Figure 3B we plotted the probability density functions (PDFs) of the confidence distributions in the continuous rating task as a function of confidence on the opt-out task. The plot demonstrates the clear tendency for confidence to leak from one task to another: the mean of the VAS PDF for trials in which observers did not opt out is higher than for trials in which they did.

Further, similarly to Experiment 1, we checked whether the observed confidence leak could be due to correlations in signal quality. We could not compute accuracy for the low-confidence trials in the opt-out task since subjects did not provide a guess when they chose to “opt out”, and therefore were unable to repeat the analysis from Experiment 1. Instead,

we computed the accuracy correlation between the two tasks restricting the analysis to high-confidence trials on the opt out task. This correlation (r = .04) was much smaller than the correlation for confidence and did not reach statistical significance (p = .08). Further, as in Experiment 1, this "performance leak" did not correlate across subjects with the amount of confidence leak (r = −.11, p = .67), confirming that confidence leak is not simply due to fluctuations in signal quality.

Modeling—To explain the data from the Experiments 1 and 2, we created a model of confidence inspired by Bayesian inference. The mathematical description in the model can be found in the Methods, while a graphical intuition is presented in Figure 4. The crux of the model is that observers attempt to maintain confidence criteria consistently with Bayesian principles, according to which confidence should reflect the likelihood of being correct. However, computing the likelihood of being correct depends on one's prior (which we always fixed to 0.5 for each choice alternative), expected likelihood function, and evidence on the current trial. Our model instantiates the intuition that observers interpret a high confidence rating on a previous task as a sign of a high signal-to-noise environment and is thus predictive of likelihood that are farther apart on the current trial. In other words, to account for the expectation of an easier task, confidence criteria are made more liberal. If the expectation is correct, then this adjustment would lead to confidence remaining tightly coupled to a particular accuracy level, as prescribed by normative Bayesian theory. However, if the expectation is incorrect (i.e., high confidence on the previous task does not predict well the difficulty on the current task), this leads to a higher confidence on the current task, and therefore to confidence leak. Mathematically, if an observer expects the signal-to-noise ratio to change by a factor of F, then to remain Bayes-optimal in her confidence ratings, she needs to shift each of her confidence criteria (defined in the likelihood space) by a factor of $\frac{1}{F}$ (see equations 10 and 11 in the Methods).

The model employs a single free parameter θ that reflects the degree to which each observer adjusts her expectation of the signal-to-noise ratio on one task as a function of confidence rating on the other task. Positive values of θ are indicative of positive between-task confidence dependency and therefore indicate the presence of confidence leak. The model provided excellent fit to the data in Experiments 1 and 2.

In Experiment 1, the average value of θ was .27, which means that for every increase of confidence by one unit (on the 1–4 scale) on the letter identity task, observers expected a signal-to-noise increase of .27 on the color task (i.e., a difference of .81 between confidence of 1 vs. 4). Further, θ was positive for 25 of the 26 observers in Experiment 1 ($t(25) = 6.42$, $p = 10^{-6}$; Cohen's d = 1.26).

Similarly, θ was positive for 17 of the 18 observers in Experiment 2. The highest value for θ in Experiment 2 was a significant outlier (3.9 SD higher than the mean group) and was therefore omitted from the statistical analyses which demonstrated that θ was significantly positive ($t(16) = 3.83$, p = .001; Cohen's d = .93). The average value of θ – with the outlier excluded – was .016, which means that for every increase of confidence by one unit (on the 0–100 scale) on the VAS task, observers expected a signal-to-noise increase of .016 on the

opt out task (i.e., a difference of 1.6 between confidence of 0 vs. 100). Note that the exact value of θ depends on factors such as using the full confidence scale (vs. only a restricted range of the scale).

Model fits are shown in Figure 2B and C with open circles, and in Figure 3B with thin lines. Despite using a single free parameter, the model was able to fit the data in Experiments 1 and 2 very well.

Experiment 3

One critical feature of our model is that confidence leak depends solely on observers' expectation of the signal-to-noise ratio on the current task as a function of the confidence rating on a previous task. This means that the phenomenon of confidence leak should be independent of factors such as the contrast of the stimuli, the individual bias for high or low overall confidence ratings, or the attentional demands of the task. In the Supplementary Material available online, we report an experiment that demonstrates that confidence leak is indeed independent of stimulus contrast. In Experiment 3, we further test whether confidence leak depends on attention.

The data from Experiment 3 were first reported in the Supplementary Material in Rahnev et al. (2011). In that study we compared conditions of high attention (2 stimuli) vs. low attention (4 stimuli) (Figure 5A) in order to understand the relationship between confidence and attention. Here, we re-analyzed the same data to check for confidence leak.

We first confirmed that attention led to increased performance, as measured by the signal detection measure *d'*. Indeed, *d'* was higher in the high attention condition (average *d'* = 1.73) than in the low attention condition (average *d'* = 1.25, $t(19) = 6.41$, $p = 3.8*10^{-6}$; Figure 5B).

We then checked for confidence leak by determining the amount of confidence autocorrelation. The assumption in this analysis was that confidence on the previous trial would leak to the confidence on the current trial. The confidence autocorrelation was positive for all 20 observers and significant for 18 of them (Figure 5C). Across the whole group, the average Fisher-transformed correlation coefficient was .28 with 95% confidence interval [.21, .36], which is significantly positive ($t(19) = 7.66$, $p = 3.2*10^{-7}$) and corresponds to a very large effect size (Cohen's $d = 1.71$) and strong evidence for the alternative hypothesis (Bayes factor = $2.1*10^{10}$). Controlling for accuracy and difficulty in a regression analysis (as in Experiments 1 and 2) produced virtually identical results.

Critically, we compared the amount of confidence leak as defined by the beta value in a regression in which the confidence on the previous trial was used to predict the confidence on the current trial, while controlling for the accuracy and contrast on each trial. We found no significant difference in the beta value in the high attention (average beta = .255) and low attention (average beta = .261) conditions ($t(19) = .22$, $p = .83$; Figure 5D) indicating that confidence leak does not depend on attention.

Our model was able to fit these data too (Figure 5B,D). According to the model fit, the amount of confidence leak, as defined by the parameter θ, was positive for 19 of the 20 observers ($t(19) = 7.65$, $p = 3.2*10^{-7}$; Cohen's $d = 1.71$), while the average value of θ was .59 suggesting that increasing confidence from "low" to "high" led observers to expect a signal-to-noise increase of .61 on the next trial.

Relationship to metacognition—According to our model, the confidence leak phenomenon depends on the extent to which a confidence rating on one task (trial) shifts observers' expectation for the signal-to-noise ratio on a different task (trial). However, since in our experiments the environment was kept stable (the difficulty of the tasks on any trial could not be predicted from the previous trial), an adjustment in expectation actually led to larger noise in observers' confidence ratings. In our model, larger amounts of confidence leak in individual observers lead to larger erroneous shifts in the likelihood functions, which, in turn, should result in lower metacognition scores (Maniscalco & Lau, 2012). We checked for such an effect by computing, for each observer, the area under the Type 2 ROC curve (Type 2 AUC), which is a standard measure of metacognition (Fleming & Lau, 2014; Fleming et al., 2010). We then correlated this measure of metacognition with the amount of confidence leak as estimated by the Fisher-transformed inter-task correlation value. To increase power, we performed the correlation on the combined data from Experiments 1–3. We found that confidence leak correlated negatively with metacognition ($r = -.36$, $p = .004$, 95% confidence interval: [−.56, −.12]). The correlation appeared to be influenced by four outliers (two observers with Fisher-transformed correlation > 1, and two observers with metacognitive score worse than chance = .5) but remained significant when we instead applied Spearman's rank correlation that is less sensitive to outliers ($\rho = -.28$, $p = .02$). We nonetheless confirmed that the correlation remained significantly negative ($r = -.26$, $p = .04$) when we removed the outliers altogether (Figure 6). Importantly, when the same analysis was performed within each experiment, we obtained similar correlation magnitudes ($r = -.31$, −.45, and −.26, respectively) but because of the smaller number of observers in each analysis, the p values did not reach the .05 level ($p = .12$, .06, and .27, respectively). Taken together, these results show that, as suggested by our model, the phenomenon of confidence leak impairs observers' metacognitive ability to introspect on their accuracy.

Experiment 4

The empirical relationship between confidence leak and metacognitive accuracy led us to explore whether the two may share a common neurophysiological representation. We re-analyzed the data from McCurdy et al. (2013) which replicated a finding first reported in Fleming et al. (2010) that higher metacognitive sensitivity is related to higher gray matter volume in the anterior prefrontal cortex (aPFC).

Behaviorally, we found very strong evidence for confidence leak in that dataset too (confidence autocorrelation was positive for all 34 observers, mean = .27, 95% CI = [.22, .31], $t(33) = 10.9$, $p = 1.9*10^{-12}$, Cohen's $d = 1.87$, Bayes factor = $2.4*10^{24}$). In addition, replicating the analysis above, we found that confidence leak correlated negatively ($r = -.38$, $p = .025$) with the measure of metacognitive efficiency (meta*d'/d'*, Maniscalco & Lau, 2012) originally used in the analyses in McCurdy et al. (2013).

We then explored whether confidence leak is related to gray matter volume. Given the negative relation between confidence leak and metacognitive scores, we expected a negative relationship between confidence leak and gray matter volume in prefrontal cortex (PFC) regions that had previous been linked to metacognitive sensitivity (Fleming et al., 2010). We used the same methods as in McCurdy et al. (2013) and found two regions in right PFC for which lower gray matter volume predicted higher confidence leak scores. Both regions survived small-volume correction for multiple comparisons (right dorsolateral PFC: peak voxel coordinate (41, 32, 22), $T = 3.76$, cluster familywise error (FWE)-corrected $p = 0.032$; right anterior PFC: peak voxel coordinate (35, 53, 6), $T = 3.71$, cluster FWE-corrected $p = 0.031$; Figure 7).

As we explained above, confidence leak logically leads to lower metacognitive scores. However, it is also possible that observers with lower metacognitive capacity are more susceptible to confidence leak (*i.e.*, the reverse causal link). Such an effect would raise the possibility that the PFC regions identified above do not contribute directly to confidence leak. Therefore, we regressed out the influence of metacognitive scores from the confidence leak scores and repeated the analyses with the residuals. We found that these new confidence leak scores were still significantly predicted by lower gray matter volume in the region in right dorsolateral PFC ($T = 4.46$, cluster familywise error (FWE)-corrected $p = 0.026$) but not by the region in anterior PFC ($p > .05$).

Discussion

Four experiments provided evidence for the novel phenomenon of confidence leak between different psychophysical tasks. In the first two experiments, confidence “leaked” between the tasks of judging the dominant letter identity and color. In the last two experiments, confidence leaked over time in the same task. Our work extends previous demonstrations of serial dependence in perceptual decisions (Fischer & Whitney, 2014; Frund et al., 2014; Liberman et al., 2014; Zhang et al., 2014) and reveals that such dependence is also present for metacognitive judgments.

What causes serial dependence in perception? Previous work has theorized the existence of a “continuity field” which arises from the brain’s attempt to use previous information to interpret the current visual scene (Fischer & Whitney, 2014; Liberman et al., 2014; Yu & Cohen, 2009). This work has shown that in the case of our percepts, this effect is likely to be perceptual rather than cognitive. The metacognitive continuity field that we have demonstrated appears to arise from a similar process: the brain attempts to interpret the fidelity of its current decision based on the perceived fidelity of past decisions. On the other hand, the phenomenon of confidence leak clearly arises from higher-level, rather than perceptual, processes as demonstrated by the fact that it appears across perceptual tasks and is dependent on the brain anatomy of the prefrontal cortex.

Our computational model formalizes the idea that confidence on a previous task is used to predict the quality of the perceptual signal in a current task. Using the principles of Bayesian inference, the model describes normative behavior in cases in which observers expect that the environment has a predictable higher order structure (David et al., 2004) that can be

exploited (Fischer & Whitney, 2014; Frund et al., 2014; Liberman et al., 2014). Specifically, it has been argued that observers assume the world is autocorrelated because it usually is, and that this assumption is not particularly detrimental to performance when the world is not autocorrelated (Yu & Cohen, 2009). The model specifies precisely how expecting higher or lower signal strength on a given trial ought to translate into an adjustment of the metacognitive criterion for confidence. Despite employing a single free parameter, the model fitted well a large amount of data (see Figures 2, 3, and 5). It also made two novel predictions that were confirmed by additional experiments or analyses. First, the model predicted that the phenomenon of confidence leak does not depend on factors that affect the perceptual signal such as attention and contrast, which was confirmed in Experiment 3 and the control experiment reported in the Supplementary Material, respectively. Second, the model predicted that stronger confidence leak leads to impaired metacognitive sensitivity (Fleming & Lau, 2014), which was confirmed with the combined data from Experiments 1–3, as well as in Experiment 4. Thus our model not only fits well the existing data but also makes testable novel predictions.

Several components of our model may seem counterintuitive. For example, why do observers use one task to predict their performance on a completely different task? We think that this is because in the real world even very different tasks tend to vary in difficulty together. For instance, foggy conditions, headache, or high level of distractibility could decrease the signal-to-noise ratio similarly for both the color and letter identity tasks. Another potential concern is whether the fact that high confidence responses make us more likely to give higher confidence on the next trial or task would lead to confidence increasing indefinitely. We note that confidence ratings are inherently based on both signal strength and our perceptual expectations (as well as other potential factors). Thus, even after a maximal confidence rating observers can still have a relatively low confidence rating on the next task or trial if the signal strength happens to be relatively low. Finally, if observers are constantly making predictions about the likelihood functions of the upcoming tasks or trials, why did they not learn to adjust these predictions over time? Since the expected likelihood functions were only modestly shifted compared to the real likelihood functions, most trials would likely not provide enough evidence to overturn observers' expectation for an autocorrelated environment, especially in the absence of trial-to-trial feedback. In a similar manner, purely perceptual decisions remain autocorrelated despite the perceptual conflict this autocorrelation brings (Fischer & Whitney, 2014; Frund et al., 2014; Liberman et al., 2014; Zhang et al., 2014).

Our results have several important implications. First, the fact that confidence leak predicts metacognitive sensitivity means that we have identified one of the causes of suboptimal metacognitive performance in perceptual decision-making (Fleming et al., 2010; Maniscalco & Lau, 2012; McCurdy et al., 2013). Second, our findings provide evidence that confidence for perceptual decisions is not solely determined by the signal strength in visual cortex, as assumed by some dominant theories. For example, theories of population coding in early visual cortex (Ma, Beck, Latham, & Pouget, 2006) postulate that visual cortex activity forms a distribution from which one can directly read the decision (usually the peak of the distribution) and the uncertainty (the width of the distribution)(Drugowitsch & Pouget, 2012). However, if confidence were determined exclusively based on activity in the visual

cortex (and perhaps additionally corrupted by white noise), then we should not have seen evidence for confidence leak in Experiments 1 and 2. Indeed, the two tasks in these experiments focused on different visual features (color and letter identity), which are often assumed to rely on processing in largely independent neural populations (James, James, Jobard, Wong, & Gauthier, 2005; Zeki, 1990). Importantly, even for neurons that may be sensitive to both shape and color, it is hard to imagine how and why sharper representation for one of these features (corresponding to high confidence) could lead to a sharper representation for the other feature.

The phenomenon of confidence leak extends previous work by de Gardelle & Mamassian (2014) who recently demonstrated the existence of a "common currency" for confidence such that the certainty for different perceptual tasks can be directly and meaningfully compared. This finding suggests that confidence is represented in generic, task-independent format, which is a necessary condition for our proposed mechanism for confidence leak. Our results also build on previous work by Mueller and Weidemann (2008) who found confidence autocorrelation in the same task but did not put forward an explanation of the causes of such confidence autocorrelation or show dependence between different tasks. Finally, our results are consistent with previous studies which suggest that confidence is influenced by a subjective perception of ease (Koriat, 2008, 2011), but extends this previous literature by considering between-task influences.

One alternative explanation for our results is that providing confidence on one task simply changes the emotional state of the observer. This explanation is not mutually exclusive with the theory that observers engaged in Bayesian inference: for example, it is possible that an expectation of higher signal quality puts observers in a better mood. Nevertheless, we consider a purely emotional account of the confidence leak phenomenon to be unlikely. Indeed, in Experiment 2 we gave observers the opportunity to earn points and presented them with the "opt out" paradigm where they could choose not to give a response if they were unsure about their response. Further, we explicitly informed them that the optimal strategy was to choose the "opt out" option when they had less than 66% certainty in their response. Observers were promised a monetary reward for high performance, and it is therefore likely that they tried to minimize purely emotional reactions, and instead gave their response based on the objectively computed probability of being correct. Despite that, in Experiment 2 we observed confidence leak that was just as strong as in the other experiments suggesting that a purely emotional account is unlikely.

An important question for future research relates to the limit of the confidence leak phenomenon. Future studies should examine whether confidence leak occurs in tasks that depend on different senses altogether, such as vision and audition. Beyond perceptual tasks, an important question is whether confidence ratings on any task, such as in memory or high-level cognitive tasks, also show confidence leak. If similar confidence leak is present for cognitive tasks, there will likely be a number of real-world implications related, for example, to how the level of confidence of witnesses in courts, doctors examining mammograms, or drivers deciding whether road conditions are dangerous can be influenced by seemingly irrelevant context.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Sneha Subramanian for helpful comments and Tashina Graves for collecting the data for Experiment 3.

Funding

This work was supported by the National Institute of Health [grant number R01 NS088628-01] and John Templeton Foundation [grant number 21569].

References

David SV, Vinje WE, Gallant JL. Natural stimulus statistics alter the receptive field structure of v1 neurons. The Journal of Neuroscience. 2004; 24(31):6991–7006. [PubMed: 15295035]

De Gardelle V, Mamassian P. Does Confidence Use a Common Currency Across Two Visual Tasks? Psychological Science. 2014

Dienes, Z. Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference. Palgrave Macmillan; 2008. Retrieved from http://www.palgrave.com/page/detail/understanding-psychology-as-a-science-zoltan-dienes/?K=9780230542303

Dienes Z. Using Bayes to get the most out of non-significant results. Frontiers in Psychology. 2014; 5

Drugowitsch J, Pouget A. Probabilistic vs. non-probabilistic approaches to the neurobiology of perceptual decision-making. Current Opinion in Neurobiology. 2012; 22(6):963–969. [PubMed: 22884815]

Fischer J, Whitney D. Serial dependence in visual perception. Nature Neuroscience. 2014; 17(5):738–743. [PubMed: 24686785]

Fleming SM, Dolan RJ, Frith CD. Metacognition: computation, biology and function. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences. 2012; 367(1594): 1280–1286. [PubMed: 22492746]

Fleming SM, Lau HC. How to measure metacognition. Frontiers in Human Neuroscience. 2014; 8

Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G. Relating introspective accuracy to individual differences in brain structure. Science. 2010; 329(5998):1541–1543. [PubMed: 20847276]

Fründ I, Wichmann FA, Macke JH. Quantifying the effect of intertrial dependence on perceptual decisions. Journal of Vision. 2014; 14(7)

James KH, James TW, Jobard G, Wong ACN, Gauthier I. Letter processing in the visual system: different activation patterns for single letters and strings. Cognitive, Affective & Behavioral Neuroscience. 2005; 5(4):452–466.

Kiani R, Shadlen MN. Representation of confidence associated with a decision by neurons in the parietal cortex. Science. 2009; 324(5928):759–764. [PubMed: 19423820]

Koriat, A. Metacognition and Consciousness. In: Zelazo, PD.; Moscovitch, M.; Davies, E., editors. Cambridge handbook of consciousness. New York: Cambridge University Press; 2007. p. 289-326.

Koriat A. Subjective confidence in one's answers: the consensuality principle. Journal of Experimental Psychology. Learning, Memory, and Cognition. 2008; 34(4):945–959.

Koriat A. Subjective confidence in perceptual judgments: a test of the self-consistency model. Journal of Experimental Psychology. General. 2011; 140(1):117–139. [PubMed: 21299320]

Liberman A, Fischer J, Whitney D. Serial Dependence in the Perception of Faces. Current Biology. 2014; 24(21):2569–2574. [PubMed: 25283781]

Ma WJ, Beck JM, Latham PE, Pouget A. Bayesian inference with probabilistic population codes. Nature Neuroscience. 2006; 9(11):1432–1438. [PubMed: 17057707]

Macmillan, NA.; Creelman, CD. Detection Theory: A User's Guide. 2nd ed.. Mahwah, NJ: Erlbaum; 2005.

Maniscalco B, Bang JW, Iravani L, Camps-Febrer F, Lau H. Does response interference depend on the subjective visibility of flanker distractors? Attention, Perception & Psychophysics. 2012; 74(5): 841–851.

Maniscalco B, Lau H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Consciousness and Cognition. 2012; 21(1):422–430. [PubMed: 22071269]

McCurdy LY, Maniscalco B, Metcalfe J, Liu KY, de Lange FP, Lau H. Anatomical Coupling between Distinct Metacognitive Systems for Memory and Visual Perception. The Journal of Neuroscience. 2013; 33(5):1897–1906. [PubMed: 23365229]

Metcalfe, J.; Shimamura, AP. Metacognition: Knowing about Knowing. Cambridge, MA: MIT Press; 1994.

Mueller ST, Weidemann CT. Decision noise: An explanation for observed violations of signal detection theory. Psychonomic Bulletin & Review. 2008; 15(3):465–494. [PubMed: 18567246]

Nelson, TO.; Narens, L. Metamemory: A theoretical framework and some new findings. In: Bower, G., editor. The Psychology of Learning and Motivation. New York: Academic Press; 1990. p. 125-141.

Nikolić D, Mureşan RC, Feng W, Singer W. Scaled correlation analysis: a better way to compute a cross-correlogram. The European Journal of Neuroscience. 2012; 35(5):742–762. [PubMed: 22324876]

Rahnev D, Bahdo L, de Lange FP, Lau H. Prestimulus hemodynamic activity in dorsal attention network is negatively associated with decision confidence in visual perception. Journal of Neurophysiology. 2012; 108(5):1529–1536. [PubMed: 22723670]

Rahnev D, Kok P, Munneke M, Bahdo L, De Lange FP, Lau H. Continuous theta burst transcranial magnetic stimulation reduces resting state connectivity between visual areas. Journal of Neurophysiology. 2013; 110(8):1811–1821. [PubMed: 23883858]

Rahnev D, Lau H, De Lange FP. Prior expectation modulates the interaction between sensory and prefrontal regions in the human brain. Journal of Neuroscience. 2011; 31(29):10741–10748. [PubMed: 21775617]

Rahnev D, Maniscalco B, Graves T, Huang E, De Lange FP, Lau H. Attention induces conservative subjective biases in visual perception. Nature Neuroscience. 2011; 14(12):1513–1515. [PubMed: 22019729]

Rahnev D, Maniscalco B, Luber B, Lau H, Lisanby SH. Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. Journal of Neurophysiology. 2012; 107(6): 1556–1563. [PubMed: 22170965]

Shimamura AP. Toward a cognitive neuroscience of metacognition. Consciousness and Cognition. 2000; 9(2 Pt 1):313–326. [PubMed: 10924251]

Yeung N, Summerfield C. Metacognition in human decision-making: confidence and error monitoring. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences. 2012; 367(1594):1310–1321. [PubMed: 22492749]

Yu AJ, Cohen JD. Sequential effects: Superstition or rational behavior? Advances in Neural Information Processing Systems. 2009; 21:1873–1880. [PubMed: 26412953]

Zeki S. A century of cerebral achromatopsia. Brain. 1990; 113(Pt 6):1721–1777. [PubMed: 2276043]

Zhang M, Wang X, Goldberg ME. A spatially nonselective baseline signal in parietal cortex reflects the probability of a monkey's success on the current trial. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111(24):8967–8972. [PubMed: 24889623]

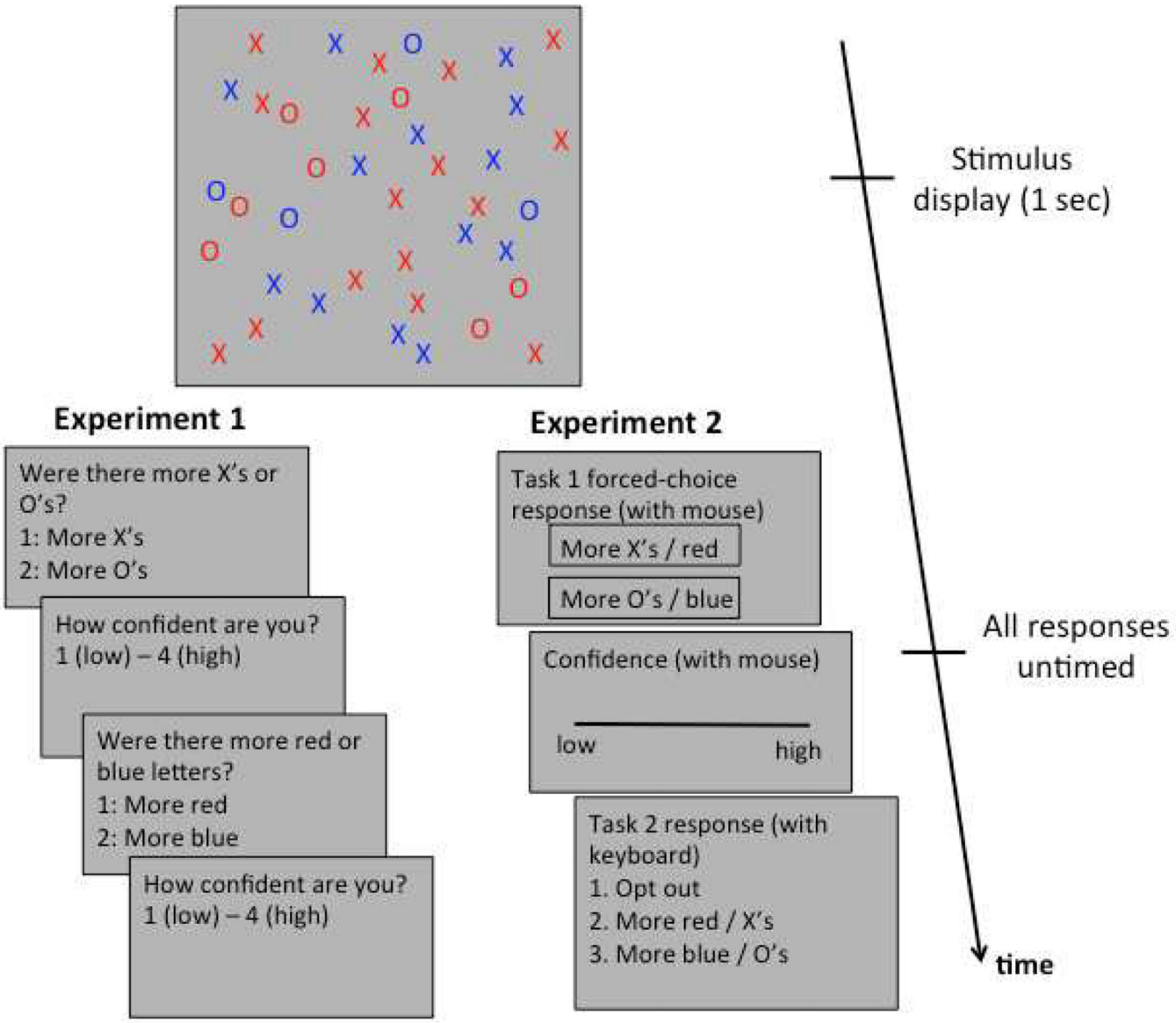

Figure 1.

Tasks for Experiments 1 and 2. The stimulus consisted of 40 X's and O's that were colored in red and blue. The observers' tasks were to decide whether there were more X's or more O's (letter identity task), as well as whether there were more red or blue letters (color task). For each task, observers indicated their level of confidence. In Experiment 1 confidence was indicated using a 1–4 scale. In Experiment 2, for one of the tasks observers used a visual analog scale (VAS) by sliding a marker, while for the other question they decided whether to provide an answer in order to win a larger reward (thereby indicating high level of confidence), or to "opt out" and not give a response thus earning a smaller, guaranteed, reward (thereby indicating low level of confidence). To further minimize response priming, in Experiment 2 the VAS response was provided with a mouse, while the opt out response was provided with a keyboard.

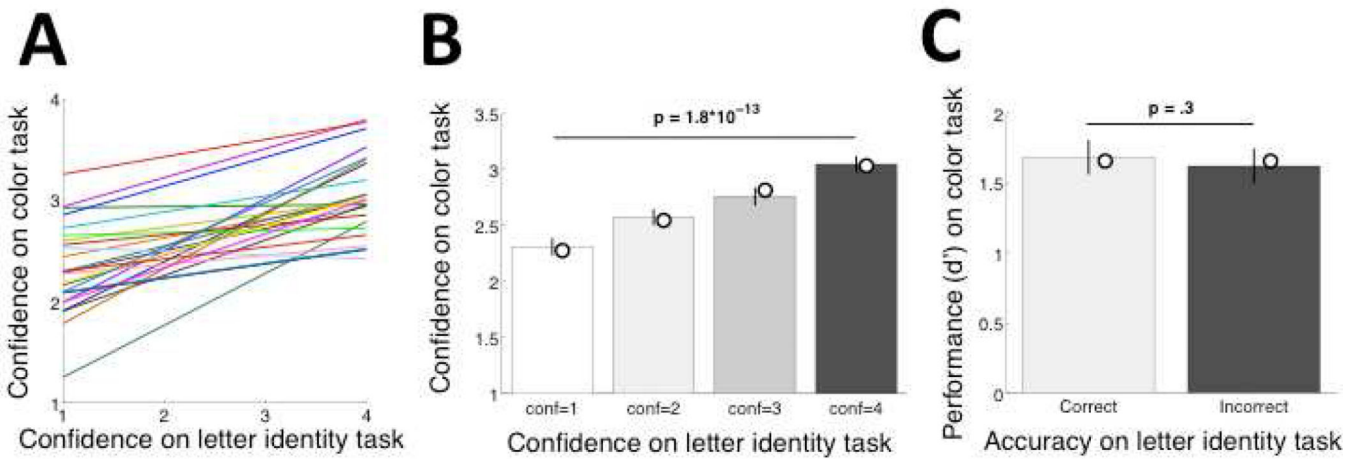

Figure 2.

Results of Experiment 1. A) Individual fits in a regression in which only the confidence on the letter identity task was used to predict the confidence on the color task. B) Mean confidence on the color task increases as a function of the confidence on the letter identity task. C) Mean accuracy on the color task does not vary as a function of the accuracy on the letter identity task. Open circles signify model fits.

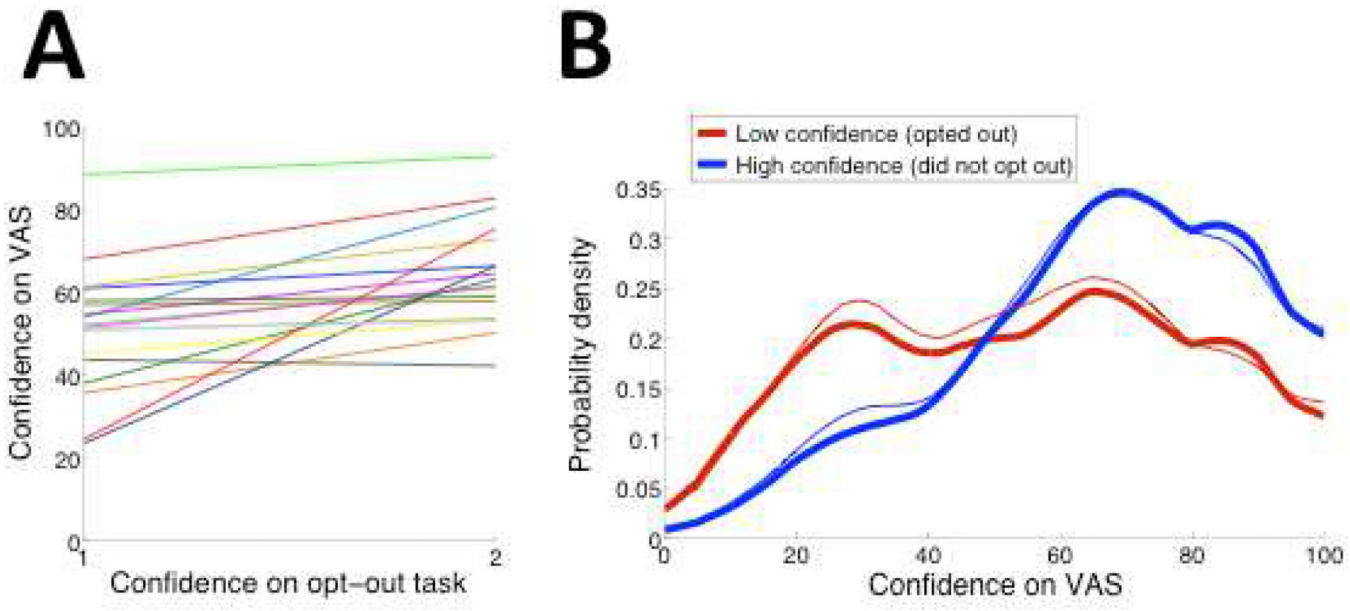

Figure 3.

Results of Experiment 2. A) Individual observers' confidence on one task, measured with the visual analogue scale (VAS), as a function of the confidence on the other task, measured with opt-out procedure. Positive relationship was found for 17 of the 18 observers. B) Probability density functions for the VAS confidence on one task for trials in which the observers decided to opt out (i.e., low confidence) vs. trials in which observers provided a response (i.e., high confidence) on the other task. For display purposes the curves were smoothed with a 10-point moving average. The thick lines indicate data, while the thin lines indicate model fits.

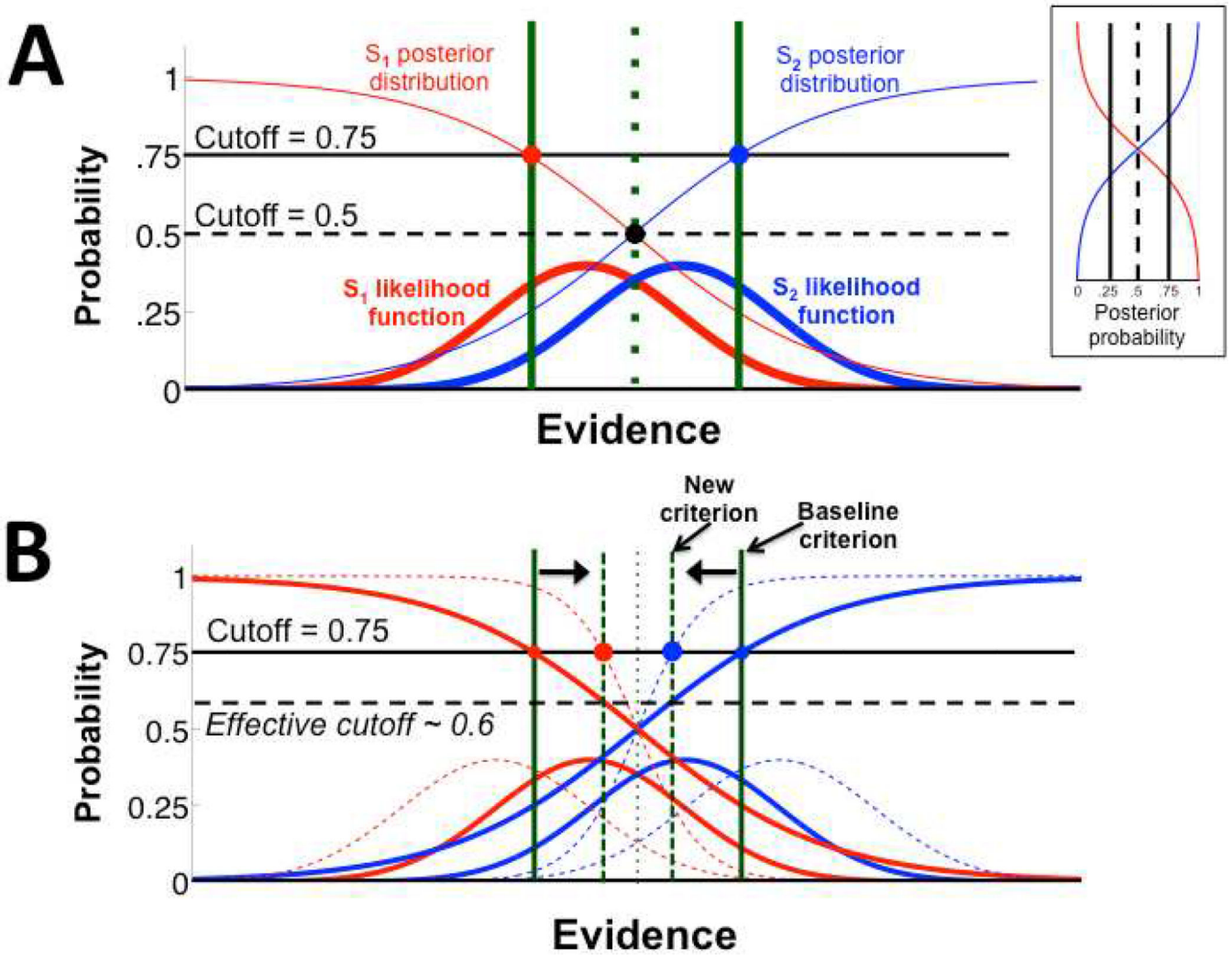

Figure 4.

A graphical depiction of our model. A) Discrimination between stimuli S_1 and S_2. Each stimulus follows a Gaussian likelihood function (thick lines) on an axis that denotes the total evidence available on a given trial. The posterior distributions (thin lines) are drawn for the case when the two stimuli have equal prior probability. Bayes-optimal thresholds are placed horizontally (black lines) based on predetermined cutoff values (in this case, 0.5 for the decision, and 0.75 for the confidence). These cutoffs correspond to vertical criteria defined in the likelihood space (green lines), which intersect the horizontal thresholds on the posterior distributions. Observers' goal is to set stable cutoffs in the posterior probability space, as depicted in the inset. B) Solid lines represent the distributions from panel A, while dashed lines represent the expected distributions in a higher signal-to-noise environment. To remain consistent with the threshold placed at a cutoff of 0.75 on the posterior distributions, the confidence criteria (defined on the likelihood space) move "inward" (see the new criterion vs. baseline criterion in the figure). However, if the expectation for a high signal-to-noise environment is false, then the observer is using the dashed criteria (based on the observer's expectations) to judge stimuli characterized by the solid distributions. This means

that unbeknown to the observer, she is using a lower effective cutoff on the true posterior distribution (in the Figure above, the effective cutoff is ~0.6), naturally resulting in a higher proportion of high confidence responses.

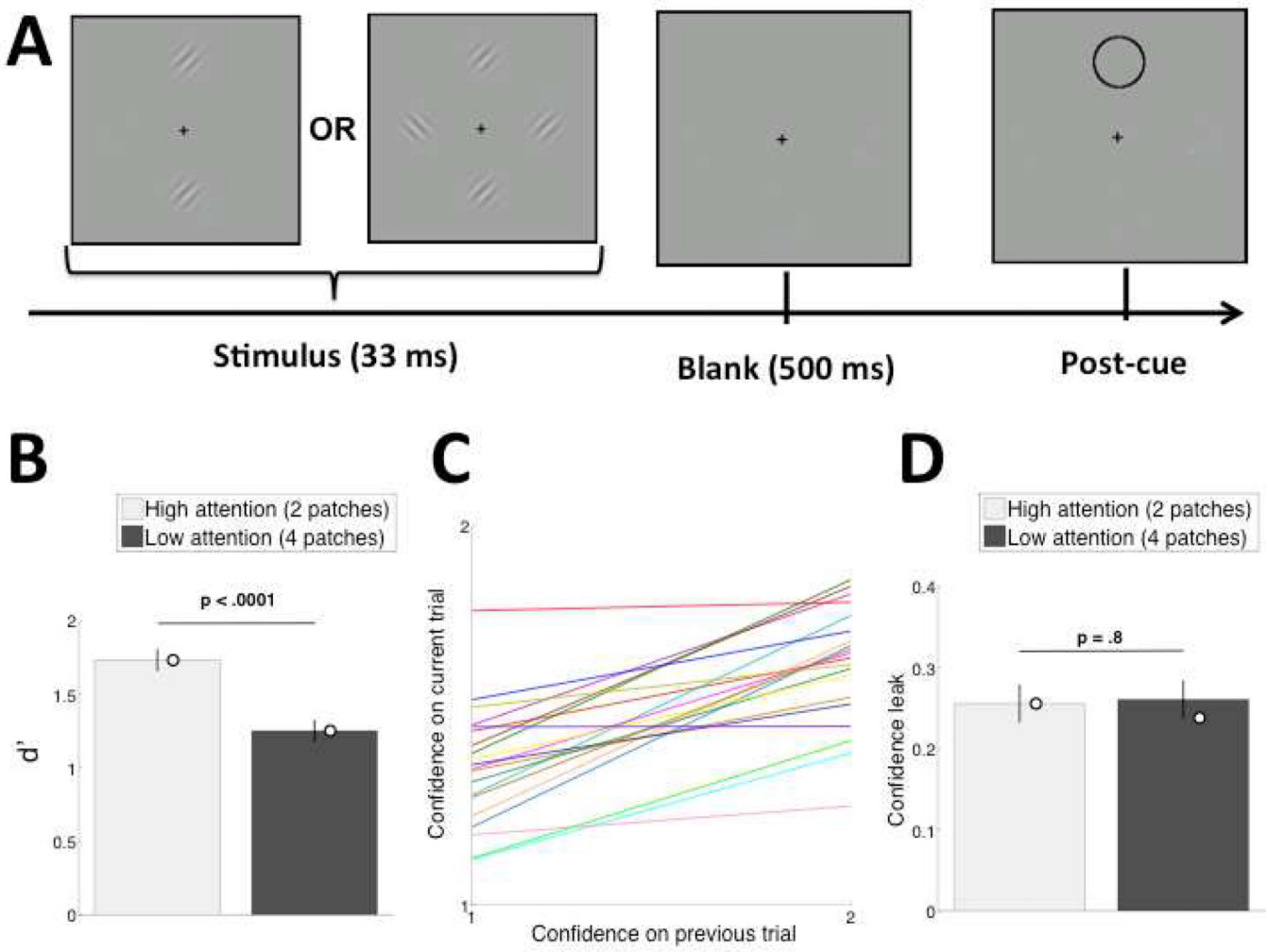

Figure 5.

Experiment 3. A) Observers were asked to decide on the orientation (clockwise vs. counter-clockwise) of briefly flashed Gabor patches. In different blocks either 2 (high attention condition) or 4 (low attention condition) patches were flashed, and observers indicated the orientation of a post-cued Gabor patch. B) Capacity *d'* was higher in the high-attention, 2-stimulus condition. C) Individual data showing that the average confidence on the current trial is positively correlated with the confidence on the previous trial for all 20 observers. D) Confidence autocorrelation was significantly positive (both p's < .001) for both the high and low attention conditions but was not significantly different between the two conditions (p = . 8).

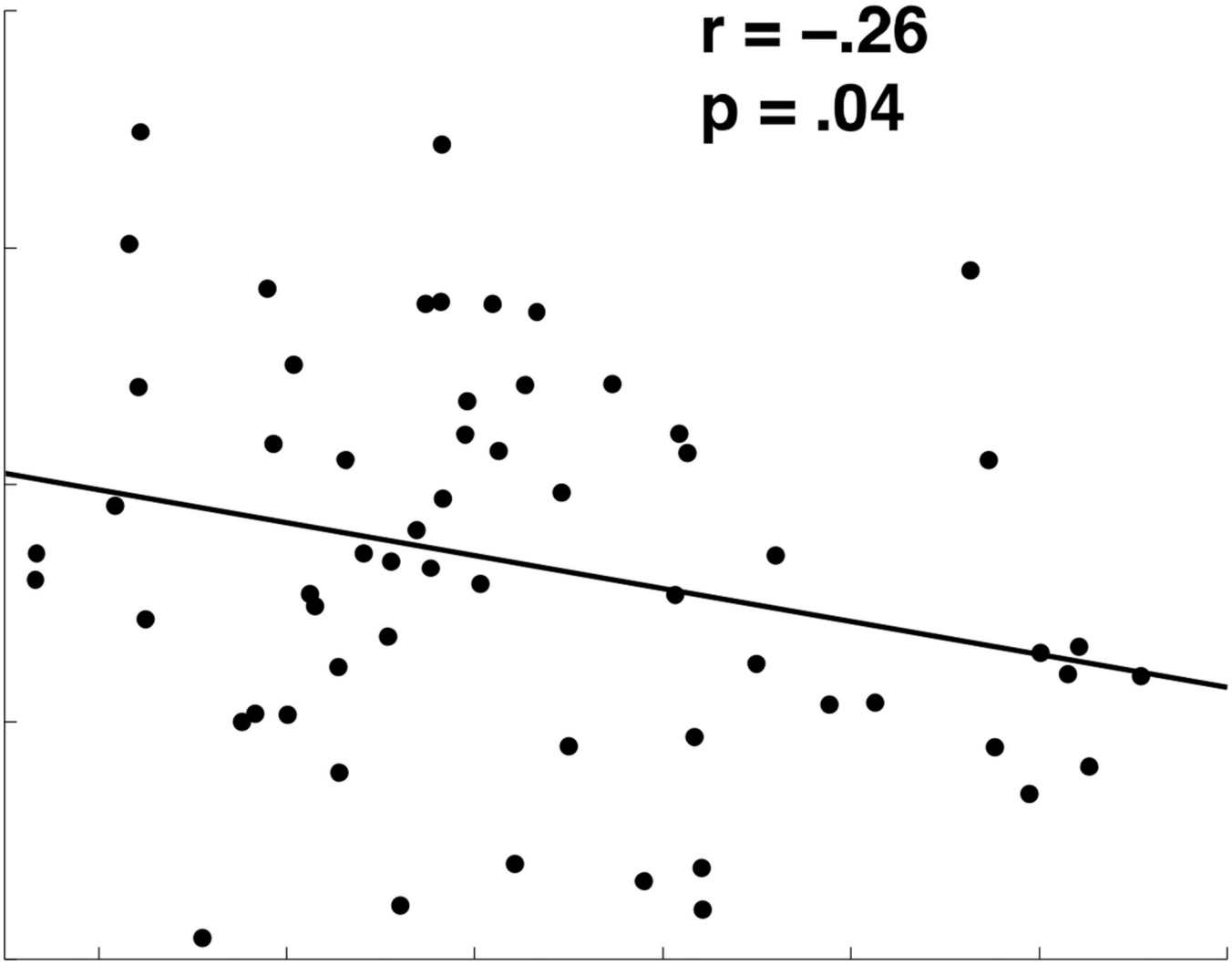

Figure 6.

The inter-task confidence leak correlates negatively with metacognition, measured as the area under the Type 2 ROC curve (Type 2 AUC), suggesting that the process of confidence leak affects negatively observers' metacognitive performance.

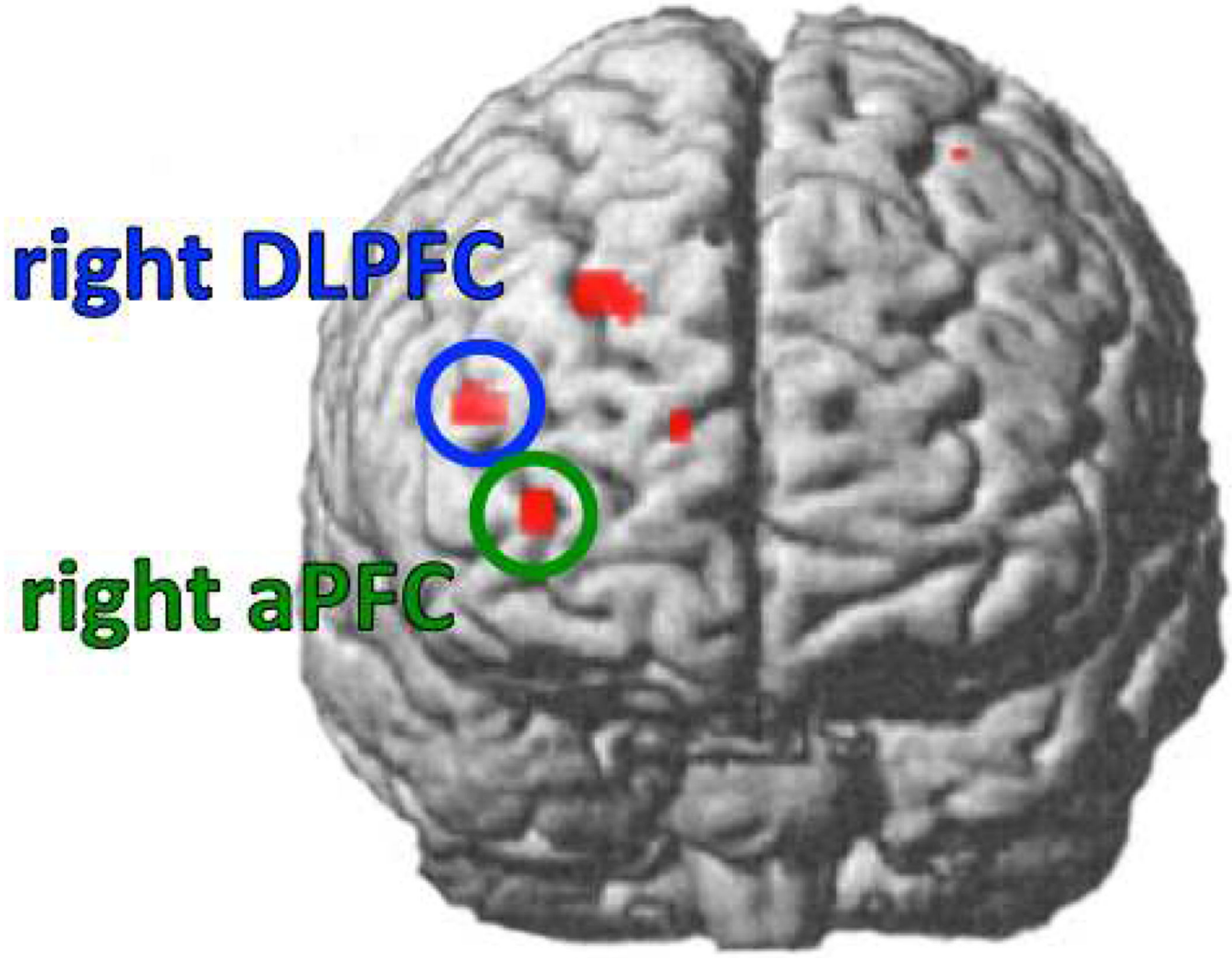

Figure 7.

Higher confidence leak scores are predicted by lower gray matter volume in right prefrontal cortex (PFC). The image shows a T map thresholded at $p = 0.005$ uncorrected for display purposes, though analyses were performed using small-volume correction for multiple comparisons. The regions in right DLPFC and aPFC are shown in blue and green, respectively.