Confidence can be automatically integrated across two visual decisions

David Aguilar-Lleyda, Mahiko Konishi, Jérôme Sackur, Vincent de Gardelle

Confidence can be automatically integrated across two visual decisions

David Aguilar-Lleyda [1*], Mahiko Konishi [2,3], Jérôme Sackur [2,4], Vincent de Gardelle [5]

[1] Centre d'Économie de la Sorbonne (CNRS & Université Paris 1 Panthéon-Sorbonne), 112 Boulevard de l'Hôpital, 75013 Paris, France
[2] Laboratoire de Sciences Cognitives et Psycholinguistique (LSCP), Département d'Études Cognitives de l'École Normale Supérieure, Centre National de la Recherche Scientifique, École des Hautes Études en Sciences Sociales, Paris Sciences et Lettres Research University, Paris, France
[3] ONERA, The French Aerospace Lab, Information Processing and Systems Department, 13661, Salon Cedex Air, France
[4] Laboratoire Interdisciplinaire de l'X, École Polytechnique, Palaiseau, France
[5] Paris School of Economics & CNRS

*corresponding author:
email: aguilarlleyda@gmail.com
address: Maison des Sciences Économiques, 112 Boulevard de l'Hôpital, 75013 Paris, France

Abstract:
Humans can estimate their confidence in making correct decisions, but these confidence judgments are biased by their other estimations, an effect known as confidence leak. However, it remains unclear whether this effect arises automatically. Here, we address this issue by having participants make two visual decisions and give confidence ratings for one or for both decisions within each trial. Using the well-known interaction between task difficulty and response accuracy as a proxy for confidence, we found that confidence ratings for one decision were greater when the other decision was also associated with greater confidence, even when the latter was not explicitly rated. For one of the two tasks, this confidence leak also occurred when participants knew in advance that no confidence rating would be required for the other task. Our results support the idea that confidence is not only automatically computed but also automatically integrated across decisions.

Keywords: confidence, metacognition, decision-making, vision, perception

Introduction

When we make a choice, we can often assess our confidence about it. This ability of evaluating our own decisions can serve many purposes. For instance, it can be used to guide learning in the absence of feedback (Guggenmos et al. 2016; Daniel and Pollmann 2012; Hainguerlot, Vergnaud, and de Gardelle 2018), to regulate information accumulation (van den Berg et al. 2016; Desender, Boldt, and Yeung 2018), or to compare different tasks (de Gardelle and Mamassian 2014; de Gardelle, Le Corre, and Mamassian 2016).

It has been long known that confidence can reflect the accuracy of the decision (Peirce and Jastrow 1884; Dallenbach 1913). However, confidence ratings are also affected by variables other than accuracy. Among them, confidence about the current decision may be attracted towards confidence ratings expressed for other decisions (Rahnev et al. 2015; Kantner et al. 2019). This effect, known as confidence leak, would be consistent with observers assuming that the quality of their perceptual evidence is relatively stable in time, as usual in natural situations. Observers would then exploit this regularity when evaluating their performance.

However appealing this general view is, its support only comes from experimental situations where both decisions required an explicit evaluation of confidence. In other words, it is not clear whether participants' confidence judgments rely on other decisions for which no explicit evaluation of confidence was made. If confidence leak depends on confidence being explicitly stated, then it would only matter in a handful of laboratory situations, excluding those situations where - as usual in life - confidence is not explicitly mapped onto a scale. Although there is evidence for confidence being computed automatically (Lebreton et al., 2015), this automatic evaluation could be an epiphenomenon without consequence. Our goal here is to evaluate whether this automatic computation of confidence also sets the context of further metacognitive evaluations. We hypothesize that our evaluation of a decision can be influenced by our confidence in another decision, even when the latter was not explicitly expressed.

One methodological difficulty to test our hypothesis is to experimentally estimate an unreported confidence judgment. Here, we relied on the previously demonstrated interaction between response accuracy and task difficulty, by which confidence should both increase for correct responses, and decrease for errors, as a task becomes easier (Kepecs et al. 2008; Sanders, Hangya, and Kepecs 2016). Specifically, in a dual-task paradigm, we evaluated whether confidence ratings for one task would be affected by the interaction between response accuracy and task difficulty for the other task. If it was, we could conclude the existence of a confidence leak from that other task to the task being rated.

We report three experiments in which we evaluate confidence leak in different settings. In Experiment 1, on each trial participants saw a perceptual stimulus for which they had to make two decisions, each followed by a confidence rating. Our goal was to replicate the original confidence leak finding, to provide a baseline against which to compare the other experiments. In Experiment 2, confidence was only required for one of the two tasks, randomly chosen on each trial. We evaluated the confidence leak from an unreported confidence, using the accuracy

x difficulty interaction as our proxy for this unreported confidence. In Experiment 3, we also told participants in advance which task would require a rating (this was kept constant within a block of trials as well) to eliminate any reason for participants to compute confidence for the non-rated task. Finding a confidence leak in this situation would provide clear evidence for the automatic nature of confidence integration across decisions.

Method

Participants

One-hundred and one healthy adults took part in three experiments (Experiment 1 = 34, Experiment 2 = 39, Experiment 3 = 28). They were recruited from the Laboratoire d'Économie Experimentale de Paris volunteer database. They reported normal or corrected-to-normal vision and gave written consent. They were naïve of the objective of the study. Each participant only took part in one experiment. Final payoff depended on performance and confidence (see supplementary material), with an average payoff of 16€. The study was approved by the Paris School of Economics ethics committee. Experiments were run in two sessions of up to 20 participants each. Final sample sizes depended on how many people attended the sessions.

Apparatus

Experiments were powered by MATLAB Psychtoolbox (Brainard 1997). Stimuli, appearing on a grey background, were viewed approximately 60cm away from the screen (17', 1024x768 pixel resolution, 60 Hz refresh rate).

Perceptual task

Trials began with a 500 ms fixation cross, followed by a 300 ms blank period, and then the stimulus for 1000 ms. The stimulus, adapted from Rahnev et al. (2015), was an array of 80 randomly positioned letters, each letter being an O or an X, colored blue or orange (see Figure 1). Each letter, in Arial font, occupied approximately 0.5º of visual angle. Elements were presented within a 10º wide imaginary square. After stimulus offset, participants reported whether there were more blue or orange elements (the color task), and more Os or Xs (the letter task). Across trials, the order of the two tasks was random, and so was the dominant element for each task. The response screen for each task presented two horizontally aligned boxes featuring an X and an O (letter task) or an orange and a blue square (color task), randomly allocated to the right or left box. Participants pressed the 'E' or 'R' key of the keyboard to select the left or right box (see Figure 1).

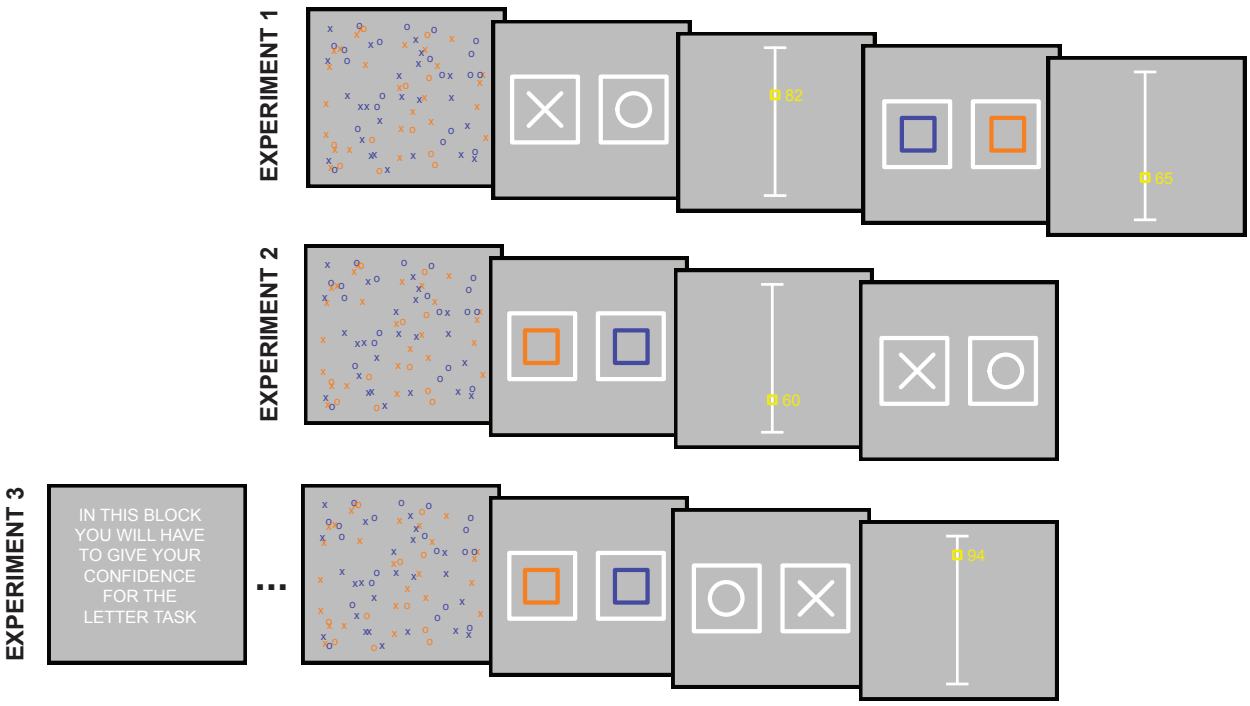

Figure 1. For all experiments, overview of a trial. Participants were presented with an array of Os and Xs, colored blue or orange, and indicated the predominant color and letter in the stimulus, by selecting the corresponding response boxes. The order of the two tasks was random within each trial. We required a confidence rating for both tasks within a trial in Experiment 1, for a task selected randomly on each trial in Experiment 2, and for a task announced to participants at the beginning of the block in Experiment 3.

Participants completed 4 parts of 4 blocks each, each block having 24 trials. A 15-second rest separated the blocks. After each part participants were given a self-timed rest, and informed of their accuracy (percentage of correct choices, pooling across both tasks) in the last part.

Confidence ratings for the different experiments

In Experiment 1, participants rated their confidence immediately after each choice. A white vertical confidence scale appeared together with a randomly positioned yellow cursor. The cursor was accompanied by an integer indicating the confidence rating, from 50 at the bottom to 100 at the top of the scale. Participants moved the cursor with their mouse, and clicked to select the desired rating. Participants were instructed that a rating of 50 would reflect total uncertainty (i.e. random choice), while 100 would reflect total certainty. After reporting confidence, the response screen for the second task appeared. After the second response, the confidence scale was presented for this second task.

In Experiment 2, confidence was asked for only one task within each trial, either the color or the letter task (counterbalanced across trials). Participants only knew which task required a

confidence rating when the confidence scale was presented. For the other task, no confidence scale was presented.

In Experiment 3, participants gave only one confidence rating per trial, but the task to be rated was known in advance: it was announced at the beginning of each block and kept constant during the block. The rated task was also counterbalanced across blocks.

Staircase and difficulty manipulation

On each trial, the proportion of items in the dominant category was controlled so that the task could be easy (90% expected performance) or hard (60% expected performance). Difficulty was controlled independently for each task, leading to a 2 x 2 factorial design. Within a block, each combination was presented equally often. The two difficulty levels were estimated for each participant in an initial psychophysical staircase of 96 trials (see supplementary material).

Statistics

We analysed our data with linear mixed models (LMM) using the lme4 (Bates et al. 2015) and lmerTest (Kuznetsova, Brockhoff, and Christensen 2017) packages in R, and report the results of ANOVA tests conducted on the fitted models. Using these models allows us to take into account every trial, instead of averaging them and giving the same weight to each level of the factor. In our experiments, this is particularly important when comparing correct trials and errors, which occur in highly different proportions. In our LMMs participants were always treated as a random intercept.

Results

Previous work claimed the presence of a confidence leak between two tasks by showing a positive correlation between the confidence ratings given for both. We started by trying to replicate these results in Experiment 1, where participants reported their confidence for both tasks within each trial. The correlation between the two confidence ratings across trials was positive for 29 out of 34 participants, with r values of participants being overall significantly larger than 0 ($t(33) = 7.000$, $p < 0.001$, 95% CI [0.132, 0.241], $d = 1.201$). In other words, in Experiment 1 we replicated the original finding of a confidence leak across tasks.

We also verified the expected relation between accuracy, difficulty and confidence. For each task (color and letter) separately, we fitted a LMM with rated confidence as dependent variable, and response accuracy (correct vs. error) and task difficulty (hard vs easy) in the task as fixed effects. The expected pattern was obtained: for both tasks we found not only significant main effects of response accuracy (color task: $F(13026) = 986.315$, $p < 0.001$; letter task: $F(13022) = 1288.025$, $p < 0.001$) and task difficulty (color task: $F(13022) = 19.814$, $p < 0.001$; letter task: $F(13021) = 34.761$, $p < 0.001$), but also a significant interaction between accuracy and difficulty (color task: $F(13024) = 399.138$, $p < 0.001$; letter task: $F(13022) = 464.092$, $p < 0.001$). When

performing this analysis for the remaining experiments, the interaction was always significant (all $p < 0.001$, see the supplementary material for the detailed information, as well as Figure 2A).

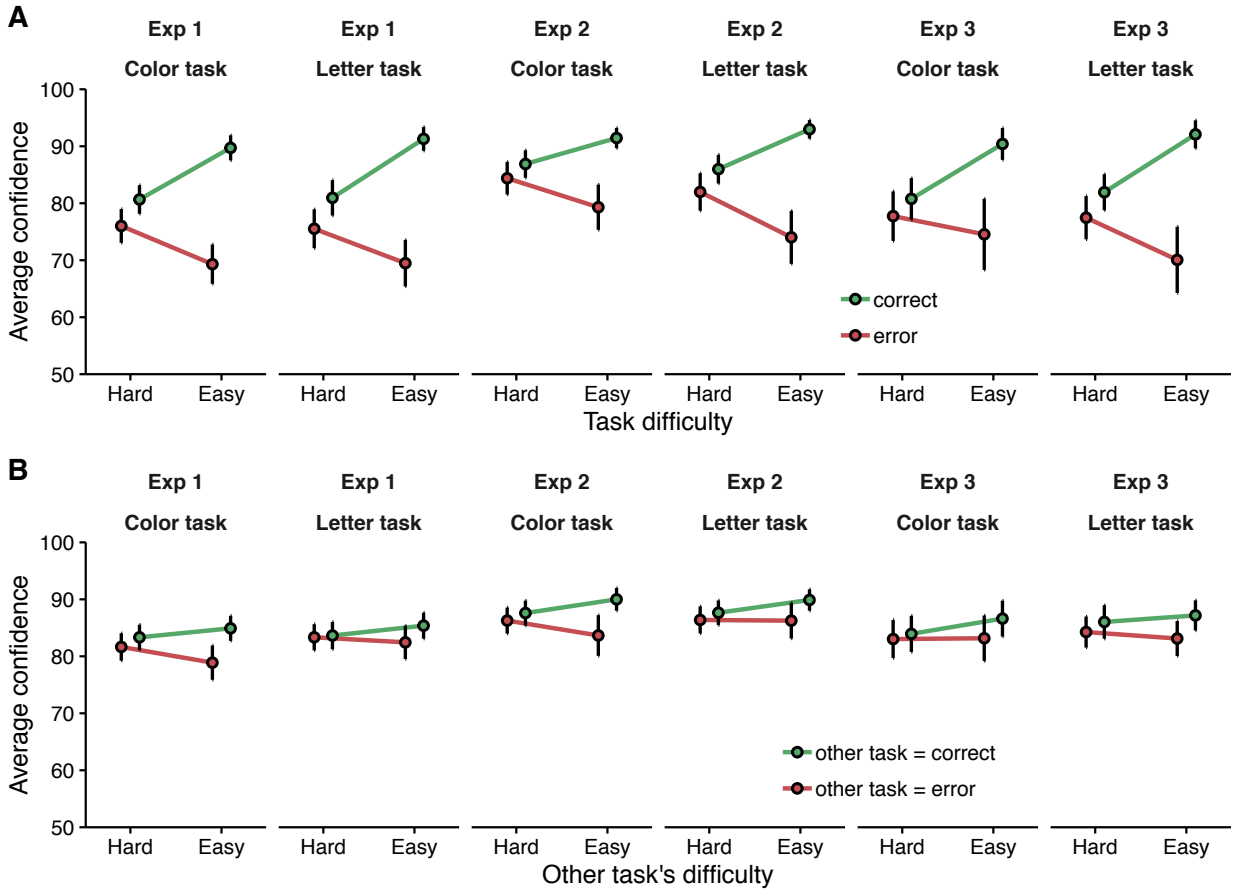

Figure 2. A. For each experiment and task, average confidence ratings across participants, as a function of task difficulty and response accuracy (color-coded). Error bars denote 95% confidence intervals. B. For each experiment and task, average confidence ratings across participants, as a function of task difficulty and response accuracy on the other task (color-coded). Error bars denote 95% confidence intervals.

Crucially, this interaction is the signature of confidence that we will rely on to evaluate confidence leak. In a nutshell, we reasoned that, if confidence ratings in one task were affected by confidence in the other task, then these ratings should be affected by the accuracy x difficulty interaction in the other task. Importantly, using this proxy, we can evaluate a confidence leak that may occur even when participants do not explicitly rate their confidence for task B.

To do so, we built a LMM where the confidence rating for a task was predicted by the response accuracy and task difficulty of the other task within the same trial, in addition to the response accuracy and task difficulty of the task at hand. By doing so, we ensure that our measure of confidence leak is not contaminated by a potential correlation in confidence between the two tasks, driven by a correlation in their performance. Table 1 shows, for each task, how

confidence was affected by the accuracy and difficulty and their interaction for the current task, as well as for the other task.

Critically, the accuracy x difficulty interaction based on the other task (rightmost columns in Table 1), was significant for both tasks in Experiment 1, for both tasks in Experiment 2, and for the color task in Experiment 3. In sum, we found evidence that ratings of confidence for one task can be influenced by the signature of confidence from another task (Experiment 1), even when confidence is not explicitly reported for that other task (Experiment 2), and even when participants know in advance that no confidence rating will be required for this other task (Experiment 3). Note however that the leak from color to letter was not significant in Experiment 3, and that this difference between the color-to-letter leak and letter-to-color leak in Experiment 3 was statistically significant, when including both tasks in the LMM (triple interaction task type x other accuracy x other difficulty, $F(10715) = 7.596$, $p = 0.006$).

		effect of task-relevant variables			**leak from the other task**		
		d.f.	**F**	***p***	**d.f.**	**F**	***p***
experiment 1		*color -> color*			*letter -> color*		
	Accuracy	13022.156	972.209	< 0.001	13019.511	47.947	< 0.001
	Difficulty	13019.363	21.491	< 0.001	13018.354	0.962	0.327
	Interaction	13020.628	392.766	< 0.001	13019.990	6.763	0.009
		letter -> letter			*color -> letter*		
	Accuracy	13018.869	1277.811	< 0.001	13021.884	6.644	0.010
	Difficulty	13017.920	36.327	< 0.001	13018.748	3.263	0.071
	Interaction	13019.263	457.565	< 0.001	13019.788	6.693	0.010
experiment 2		*color -> color*			*letter -> color*		
	Accuracy	7452.862	259.765	< 0.001	7447.707	78.842	< 0.001
	Difficulty	7446.741	2.789	0.094	7446.095	0.085	0.770
	Interaction	7447.594	106.703	< 0.001	7446.859	37.666	< 0.001
		letter -> letter			*color -> letter*		
	Accuracy	7449.883	463.810	< 0.001	7452.373	20.936	< 0.001
	Difficulty	7447.628	8.720	0.003	7446.821	12.788	< 0.001
	Interaction	7448.781	169.093	< 0.001	7447.551	3.889	0.049
experiment 3		*color -> color*			*letter -> color*		
	Accuracy	5346.829	208.313	< 0.001	5346.323	27.558	< 0.001
	Difficulty	5344.857	32.201	< 0.001	5343.594	3.172	0.075
	Interaction	5345.079	103.870	< 0.001	5344.230	19.743	< 0.001
		letter -> letter			*color -> letter*		
	Accuracy	5345.749	299.843	< 0.001	5347.009	19.493	< 0.001
	Difficulty	5343.561	18.505	< 0.001	5343.148	2.255	0.133
	Interaction	5343.562	138.020	< 0.001	5343.598	1.879	0.171

Table 1. For each experiment and task, results of the LMMs where confidence for a task is predicted by response accuracy and task difficulty for that task (task-relevant variables) and by response accuracy and task difficulty for the other task (leak from the other task).

In Experiment 3, participants may have devoted more cognitive resources to the task for which confidence had to be rated, simply because they were required more information for this task than for the other. To evaluate this possibility, we conducted an ANOVA on performance, with task type, task difficulty and whether confidence had been asked for that task or not, as factors. Performance was affected by task difficulty ($F(1,27) = 1510.331$, $p < 0.001$, $\eta_p^2 = 0.982$), but also by whether confidence was asked ($F(1,27) = 23.073$, $p < 0.001$, $\eta_p^2 = 0.461$) and by the interaction between these two factors ($F(1,27) = 17.288$, $p < 0.001$, $\eta_p^2 = 0.390$). No other main effects or interactions were significant. In short, performance was higher for the rated task, in particular for easier than for hard tasks, as illustrated in Figure 3, suggesting that participants may have dedicated more attention to the task that was associated with a confidence rating. In the case of the letter task, even when it received less attention this task still generated a leak.

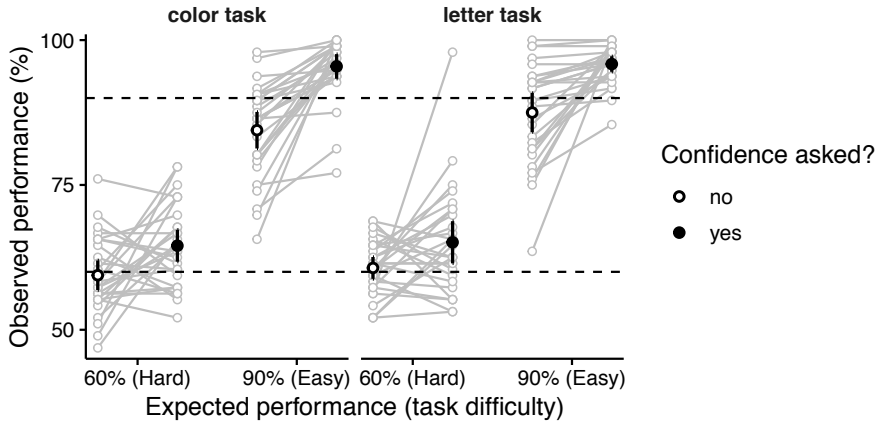

Figure 3. For each task of Experiment 3, observed performance as a function of the expected performance, split by whether confidence was asked for that task. Bigger dots represent average performance across participants expressed as a percentage, with error bars denoting 95% confidence intervals. Smaller dots represent individual participants. Top and bottom dashed lines help indicate where 90% and 60% observed performance would lay, respectively.

Discussion

Previous work suggests that confidence in one decision may leak into confidence in another decision, thus producing a confidence integration across decisions. However, evidence for this leak has only come from experimental situations where both decisions required an explicit judgment of confidence. In the present study, we show that even an unreported confidence can leak into the confidence on another decision. To do so, we rely on a proxy for confidence: specifically, we look at whether the confidence rating for a task depends on the interaction between response accuracy and task difficulty for the other task. We found that confidence in one dimension was influenced by confidence in the other dimension when the latter was reported (Experiment 1), but also when this other confidence was not reported (Experiments 2

and 3), and even when participants knew in advance that they would not have to report it (Experiment 3, although only from the letter to the color task). Beyond providing further evidence that confidence is computed automatically (Lebreton et al. 2015), we show that this computation of confidence is automatically integrated across decisions.

The finding that confidence leak is observed across two different tasks suggests that the representation involved is relatively abstract and task-independent, as argued before (Rahnev et al., 2015). The present study provides more evidence along the same lines, and furthermore shows that this computation of confidence affects our judgments even when it is not made explicitly. Moreover, we show that this influence from the non-rated task happens despite the fact that this task receives less resources, as indicated by a reduced perceptual performance.

We relied on the interaction between accuracy and difficulty as a proxy for the confidence in the non-rated task. As this interaction is indeed considered a signature of confidence (Kepecs et al. 2008; Sanders, Hangya, and Kepecs 2016), our approach ensured that the influence could be attributed to confidence in the non-rated task, and not simply to the difficulty or to the accuracy in the non-rated task. In addition, since our analyses incorporated accuracy and difficulty for the rated task, our results cannot be explained by a correlation of performance between the two tasks (e.g. due to arousal) that would introduce a correlation of confidence between the tasks masquerading as a confidence leak. Not requiring a confidence rating also ensured that the leak was not due to an anchoring effect by which participants would use similar regions within the displayed confidence scale for the two ratings.

In previous studies (Rahnev et al. 2015; Kantner et al. 2019; Mueller and Weidemann 2008), confidence has also been shown to be auto-correlated across consecutive trials, either within the same task, or between two different tasks. However, as detailed in the supplementary materials, our data indicated no widespread effect of such leak across trials. It may be that, in a dual task paradigm, any effect coming from the previous trial may vanish because of the greater delay between consecutive trials, or because of the influence of the other task within the same trial. As a matter of fact, in the aforementioned studies, confidence leak across consecutive trials has only been found when only one task was presented per trial.

One obvious question that remains is why in Experiment 3 confidence for the letter task was clearly integrated into confidence for the color task, but not vice versa. A possible answer lies in how each task may be solved. We can speculate that a decision for the color task is reached at a glance, without having to identify whether each individual element is blue or orange, whereas the letter task requires a more active engagement in terms of visual search. This greater deliberation or effort could have made confidence for the letter task more salient, and thus more likely to leak into color confidence, but also more impervious from influence by the color task. While the present data does not allow us to test these hypotheses, further studies could clarify the cognitive processes engaged during the two tasks, and could evaluate how participants form confidence on these tasks (Reyes and Sackur 2014). One exciting question for future research is whether the saliency of confidence computation, and thus the strength of confidence leak, would depend on the automaticity of the tasks themselves. This suggested link between

automaticity and availability for the monitoring system echoes theoretical proposals made in the context of metacognition and consciousness studies (Cleeremans, 2006).

References

Bates, D., M. Maechler, B. Bolker, and S. Walker. 2015. “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software*.

Berg, Ronald van den, Ariel Zylberberg, Roozbeh Kiani, Michael N. Shadlen, and Daniel M. Wolpert. 2016. “Confidence Is the Bridge between Multi-Stage Decisions.” *Current Biology: CB* 26 (23): 3157–68.

Brainard, D. H. 1997. “The Psychophysics Toolbox.” *Spatial Vision* 10 (4): 433–36.

Cleeremans, A. (2006). Conscious and unconscious cognition: A graded, dynamic perspective. In Q. Jing, M.R. Rosenzweig, G. d’ Ydewalle, H. Zhang, H.-C. Chen & K. Zhang (Eds.), Progress in Psychological Science Around the World Volume I: Neural, Cognitive, and Developmental Issues (Proceedings of the 28th International Congress of Psychology), Hove, UK: Psychology Press, pp. 401- 418.

Dallenbach, Karl M. 1913. “The Relation of Memory Error to Time Interval.” *Psychological Review* 20 (4): 323.

Daniel, Reka, and Stefan Pollmann. 2012. “Striatal Activations Signal Prediction Errors on Confidence in the Absence of External Feedback.” *NeuroImage* 59 (4): 3457–67.

Desender, Kobe, Annika Boldt, and Nick Yeung. 2018. “Subjective Confidence Predicts Information Seeking in Decision Making.” *Psychological Science* 29 (5): 761–78.

Gardelle, Vincent de, François Le Corre, and Pascal Mamassian. 2016. “Confidence as a Common Currency between Vision and Audition.” *PloS One* 11 (1): e0147901.

Gardelle, Vincent de, and Pascal Mamassian. 2014. “Does Confidence Use a Common Currency across Two Visual Tasks?” *Psychological Science* 25 (6): 1286–88.

Guggenmos, Matthias, Gregor Wilbertz, Martin N. Hebart, and Philipp Sterzer. 2016. “Mesolimbic Confidence Signals Guide Perceptual Learning in the Absence of External Feedback.” *eLife* 5 (March). https://doi.org/10.7554/eLife.13388.

Hainguerlot, Marine, Jean-Christophe Vergnaud, and Vincent de Gardelle. 2018. “Metacognitive Ability Predicts Learning Cue-Stimulus Associations in the Absence of External Feedback.” *Scientific Reports* 8 (1): 5602.

Kantner, Justin, Lisa A. Solinger, David Grybinas, and Ian G. Dobbins. 2019. “Confidence Carryover during Interleaved Memory and Perception Judgments.” *Memory & Cognition* 47 (2): 195–211.

Kepecs, Adam, Naoshige Uchida, Hatim A. Zariwala, and Zachary F. Mainen. 2008. “Neural Correlates, Computation and Behavioural Impact of Decision Confidence.” *Nature* 455 (7210): 227–31.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune Haubo Bojesen Christensen. 2017. “lmerTest Package: Tests in Linear Mixed Effects Models.” *Journal of Statistical Software* 82 (13). http://orbit.dtu.dk/portal/en/publications/id(2e6e11c6-6a44-43df-b809-d96b897d9bf2).html.

Lebreton, Maël, Raphaëlle Abitbol, Jean Daunizeau, and Mathias Pessiglione. 2015. “Automatic Integration of Confidence in the Brain Valuation Signal.” *Nature Neuroscience* 18 (8): 1159–67.

Mueller, Shane T., and Christoph T. Weidemann. 2008. “Decision Noise: An Explanation for Observed Violations of Signal Detection Theory.” *Psychonomic Bulletin & Review* 15 (3): 465–94.

Peirce, Charles Sanders, and Joseph Jastrow. 1884. “On Small Differences in Sensation.”

https://philarchive.org/archive/PEIOSD.
Rahnev, Dobromir, Ai Koizumi, Li Yan McCurdy, Mark D'Esposito, and Hakwan Lau. 2015. "Confidence Leak in Perceptual Decision Making." *Psychological Science* 26 (11): 1664–80.
Reyes, Gabriel, and Jérôme Sackur. 2014. "Introspection during Visual Search." *Consciousness and Cognition* 29 (October): 212–29.
Sanders, Joshua I., Balázs Hangya, and Adam Kepecs. 2016. "Signatures of a Statistical Computation in the Human Sense of Confidence." *Neuron* 90 (3): 499–506.

Declaration of competing interests

The authors declare no competing interests.

Supplementary information

Training and staircase

Before the main part of the experiment, analyzed in the results section, participants went through two initial parts.

At the beginning of the experiment, participants were familiarized with the stimulus and the way to give responses. To start with the easiest setting, they practised with trials for which only one choice and subsequent confidence had to be given. Participants completed a short block of 5 trials concerning only the color task, and another block of 5 trials concerning only the letter task, with the order randomized across participants.

After this short training, participants completed 96 trials very much resembling the main part of the experiment. On each trial, they gave the responses for both tasks and, depending on the experiment, rated their confidence for both or one task. However, there was a crucial difference with the subsequent main part of the experiment: for each task, we implemented a staircase procedure. On each trial, each task's difficulty was updated based on the last trial's response for that task. Difficulty was altered by changing the proportion of elements (blue over orange, or O over X). For the first trial, 64 of the 80 elements belonged to the dominant category. On subsequent trials, a correct response would decrease the dominant category (make the task more difficult) by one element, while an error would increase (make the task easier) it by 4 elements. In order to avoid participants tampering with the staircase procedure to make final difficulties easier, the staircase trials were disguised as an initial part of 4 blocks. Except for the adjustment of the stimulus, this part mimicked the design of the other parts of the experiment, including the specific details of the confidence rating for that experiment.

Once the staircase trials had finished, we used the data from this part to obtain the proportion of elements that would be used in the main part when presenting easy and hard trials. For the color task, we estimated the psychometric function representing the probability of responding blue as a function of the number of blue elements, fitted with a cumulative Gaussian. To calculate the number of elements leading to a 90% expected performance (easy condition), we took 40 (50% of the elements present on a trial) and added the the semi-difference between the number of blue elements for which the psychometric curve predicted a 90% and 10% of blue choices, rounded to its nearest integer. This procedure assumes that participants were not biased towards responding more blue or orange. A similar procedure was used to obtain the proportion of elements for the hard condition, and for both conditions of the letter task.

To demonstrate that the staircase procedure worked, Figure S1 shows, for all task types and experiments, expected performance against observed performance in the main part of the experiment. As can be seen, they were well-matched.

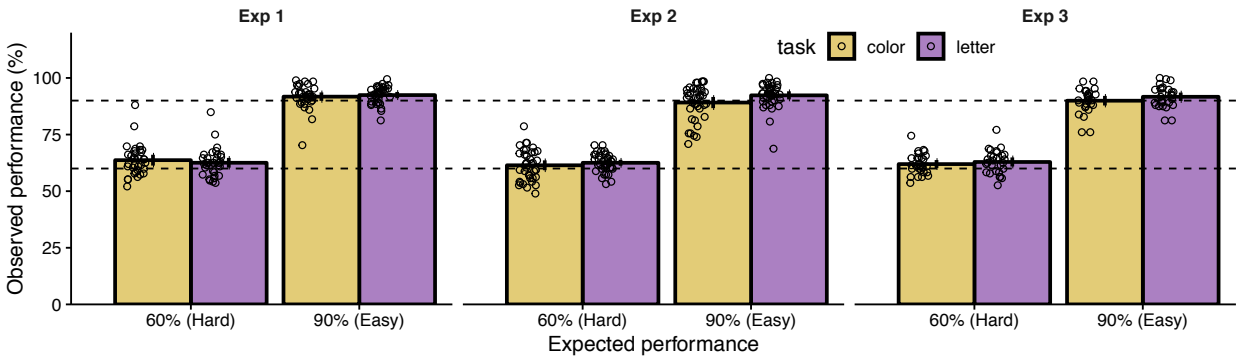

Figure S1. For each experiment, observed performance as a function of expected performance, both in percentage, split by task type. Bars represent averages across participants, with error bars denoting 95% confidence intervals. Dots represent individual participants. Top and bottom dashed lines help indicate where 90% and 60% observed performance would lay, respectively.

Payoff

At the end of the experiment, participants were paid proportionally to both their performance in the perceptual tasks and to how accurately their confidence ratings reflected their performance. In Experiment 1, for each of the 16 blocks, a trial was randomly selected. Then, also at random, we chose the color or the letter task. We compared the confidence rating that had been given for that task, with a random number picked from a uniform distribution ranging from 50 to 100. If the random number was smaller than the confidence rating, payoff depended on performance: 1€ was given if the choice for that task within that trial had been correct, and 0€ if it had been an error. If the random number was bigger than the confidence rating, payoff depended on a lottery: we compared the previous random number with another number was sampled from a uniform distribution between 0 and 100, and 1€ was given if the former was bigger than the latter, and 0€ otherwise.

In experiments 2 and 3 the payoff system was slightly different. Since only one confidence rating was given per trial, this was always what was compared with the initial random number. If payoff depended on performance, we then randomly chose whether it would depend on the correctness of the color or the letter task.

This payoff system was adapted from (Massoni, Gajdos, and Vergnaud 2014). Its objectives are two. First, making participants be focused at all time, since any trial could be picked for payoff. Second, forcing participants to give a confidence rating that accurately reflected their belief in their choices to be correct. If they expressed a high confidence, they would be more likely to be paid based on their choice, whereas a low confidence would more easily lead to a lottery. In the long run, accurately reporting confidence would be translated into a higher payoff. This system and its rationale were explained to participants before the experiment.

Correlation of confidence between tasks for Experiment 1

In the main text we already reported how, in Experiment 1, the rating between the color and the letter task within a trial were positively correlated. Figure S2 offers graphic support for that finding.

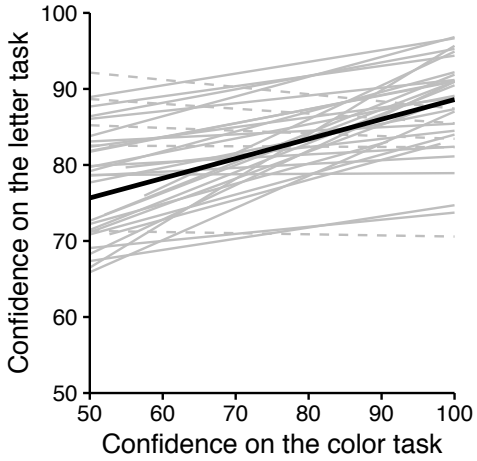

Figure S2. For Experiment 1, confidence on the color task against confidence on the letter task. Each grey line corresponds to a linear regression for one participant, made using the confidence ratings for all trials. For visualization purposes, the few participants with a negative r value are depicted with a dashed line. The black line corresponds to a linear regression using the whole data across participants.

Confidence rating as a function of accuracy and difficulty of that task, split by experiment

Here we report the results of the ANOVA on the LMM that we conducted to investigate how a confidence rating changed based on the parameters of the task for which that rating had been given. Unlike in the main text, where we collapsed our dataset across experiment, here we give the results individually for each experiment, showing that the pattern does not change. Indeed, note that the interaction between accuracy and difficulty is always very significant.

Table S1. For each experiment, results of the ANOVA on an LMM where confidence for a task is predicted as a function of that task's type, response accuracy for that task on that trial, and difficulty for that task on that trial.

Experiment number	Task type	Parameter (always from that task)	Degrees freedom	F	*p*
1	Color	Accuracy	13026	986.315	< 0.001
		Difficulty	13022	19.814	< 0.001
		Accuracy * Difficulty	13024	399.138	< 0.001
	Letter	Accuracy	13022	1288.025	< 0.001
		Difficulty	13021	34.761	< 0.001
		Accuracy * Difficulty	13022	464.092	< 0.001
2	Color	Accuracy	7456	276.900	< 0.001
		Difficulty	7450	2.037	0.154
		Accuracy * Difficulty	7451	108.895	< 0.001
	Letter	Accuracy	7453	478.969	< 0.001
		Difficulty	7451	7.788	0.005
		Accuracy * Difficulty	7452	174.081	< 0.001
3	Color	Accuracy	5350	216.795	< 0.001
		Difficulty	5348	30.124	< 0.001
		Accuracy * Difficulty	5348	108.154	< 0.001
	Letter	Accuracy	5349	306.084	< 0.001
		Difficulty	5347	18.378	< 0.001
		Accuracy * Difficulty	5347	139.311	< 0.001

Analyses on the across-trial confidence leak

Our reported analyses so far have tried to determine the presence of a confidence leak from one task to another within the same trial. However, previous work has documented a leak from the confidence reported for tasks on the previous trial. Thus, we also tested for the existence of such leaks in our data. We explored whether the rating for a task was affected by confidence on the same task for the previous trial, but also by confidence on the other task for the previous trial. We approached this by expanding the LMM used to check for a confidence leak in the results section of the main text. The added fixed effects were accuracy and difficulty for the same task on the previous trial, and those two parameters for the other task on the previous trial. We then identified any significant interaction between the two parameters for one of the tasks. The interaction between accuracy and difficulty of the last trial's same task was only

significant for the color task of Experiment 2 (F(7131.904) = 7.547, p = 0.006), while it was marginally significant for the color task of Experiment 1(F(12470.525) = 3.594, p = 0.058). The interaction for the parameters of the last trial's other task was significant for the color task of Experiment 2 (F(7131.884) = 4.077, p = 0.044), and marginally significant for the letter task of the same experiment (F(7126.753) = 3.045, p = 0.081). These results show that although there was evidence for across-trial confidence leak for some experiments and tasks, this was not as widespread as the within-trial leak. Figure S3 illustrates how confidence on a task depended on the parameters for the previous trial's same (Figure S3A) and other (Figure S3B) tasks.

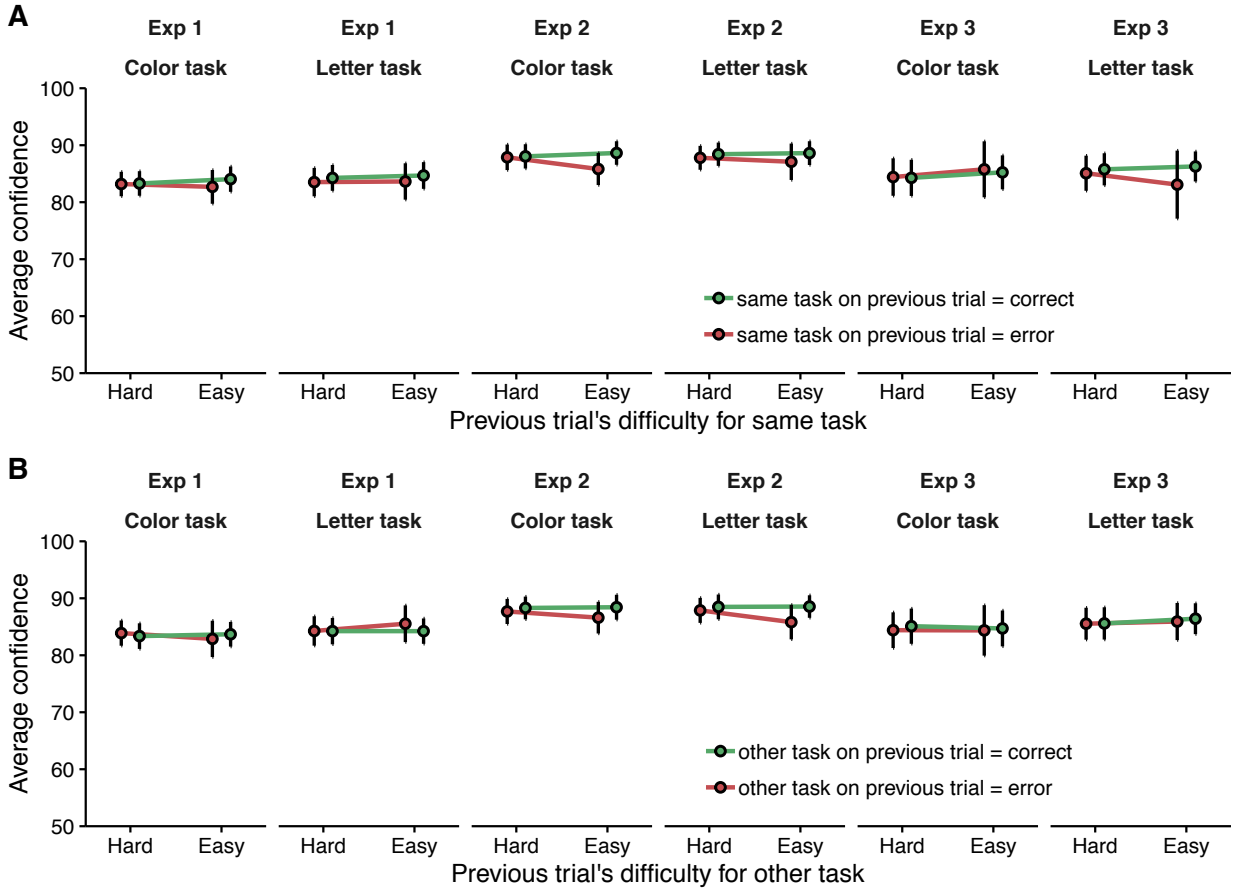

Figure S3. A. For each experiment and task, average confidence ratings across participants, as a function of the difficulty of the same task type on the past trial, and split according to the accuracy of the same task type on the past trial (color-coded). Error bars denote 95% confidence intervals. B. For each experiment and task, average confidence ratings across participants, as a function of the difficulty of the other task type on the past trial, and split according to the accuracy of the other task type on the past trial (color-coded). Error bars denote 95% confidence intervals.

Rated confidence as a function of the order in which the rated task was responded to

In our 3 experiments we always presented two tasks per trial. An interesting question not asked in the main text is whether confidence ratings changed as a function of the order for which the corresponding task had to be done: first or second within the trial. A confidence rating for the first-responded task was only preceded by that task. A rating after the second-responded task was preceded also by the other task, and in Experiment 1 also by its confidence rating. It could be that confidence suffered a memory decay, such that, for a second-responded task, confidence on that decision had somewhat waned. We investigated this possibility by taking the LMM used in the main text to test for the confidence leak, and adding as a factor the order in which that rating's task had been responded.

The main effect of order was significant or marginally significant for all experiments and tasks (Exp. 1, color task: $F(13009.625) = 3.921$, $p = 0.048$; Exp. 1, letter task: $F(13009.509) = 8.112$, $p = 0.004$; Exp. 2, color task: $F(7436.940) = 2.836$, $p = 0.092$; Exp. 2, letter task: $F(7437.229) = 4.029$, $p = 0.045$; Exp. 3, color task: $F(5335.440) = 3.501$, $p = 0.061$; Exp. 3, letter task: $F(5335.899) = 6.958$, $p = 0.008$). The direction of this effect revealed that, indeed, confidence tended to be higher for first-responded tasks. Figure S4 gives a hint of this tendency.

In order to see whether order changed the confidence leak, we checked for a significant triple interaction among accuracy for the other task, difficulty for the other task, and order. Only in Experiment 2 for the color task did this interaction approach significance ($F(7347.829) = 3.016$, $p = 0.083$). In summary, there was a tendency for confidence ratings to be higher for first-responded tasks than for second-reported tasks, but overall that order difference did not affect the confidence leak.

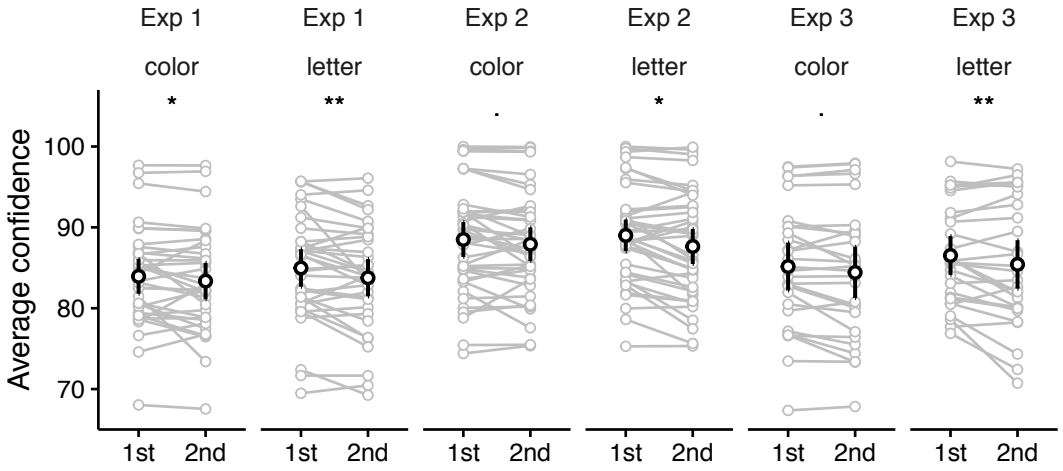

Figure S4. For all experiments and task types, average confidence as a function of whether the confidence rating was given for the first- or for the second-responded task. Individual participants are depicted in grey. Black, bigger dots represent across-participant averages, with 95% confidence intervals. Asterisks above the plotted data denote the significance level for the main effect of order given by the ANOVA on the LMM described in this section. A dot instead of an asterisk denotes marginal significance.

References

Massoni, Sébastien, Thibault Gajdos, and Jean-Christophe Vergnaud. 2014. “Confidence Measurement in the Light of Signal Detection Theory.” *Frontiers in Psychology* 5 (December): 1455.