Confidence as a priority signal

David Aguilar-Lleyda[1]*, Maxime Lemarchand[1], Vincent de Gardelle[2]

[1]Centre d'Économie de la Sorbonne (CNRS & Université Paris 1 Panthéon-Sorbonne)
[2]Paris School of Economics & CNRS
* Correspondence should be sent to aguilarlleyda@gmail.com

When dealing with multiple tasks, we often find ourselves in the problem of establishing the order in which to tackle them. Here we show that confidence, the subjective feeling in the accuracy of our decisions or actions, plays an active role in task ordering. Our participants categorized perceptual stimuli along two dimensions and reported the confidence on each decision, while we manipulated the difficulty over each dimension independently. We found an effect of confidence in priority: participants tended to address first the dimension they were more confident in. In subsequent studies, we replicated this finding using task difficulty and response accuracy as proxies for confidence: participants first responded to the easiest of the two tasks, and their first choice was also more likely to be correct. We showed that this effect was not due to a difference in response availability between dimensions. We also extended our finding to situations of prospective confidence, and to situations involving non-perceptual (mental calculation) decisions. Our results support the role of confidence as a priority signal, thus strengthening the evidence for confidence having an active role shaping our future behavior.

It is frequent to start a day at work by compiling in a to-do list all the tasks to complete that day. But once the list is written, where do we start? Which task shall we tackle first, and which tasks can be postponed? Of course, this planification problem is sometimes solved by taking into account external constraints, such as scheduled meetings, imminent deadlines, or the limited availability of specific tools or collaborators. In other circumstances, however, we are free to decide in which order to complete our tasks. In such unconstrained situations, humans may not perform their tasks in a random order, and they might instead exhibit some systematic preference for doing one task before another. Here, we suggest that confidence may play a role in this prioritization problem. In a situation where two tasks have to be done, individuals may address first the one about which they feel more confident. In a situation where two completed tasks have to be reported, individuals may start by the one in which they were more confident. In both situations, confidence would be a priority signal.

Different considerations make this hypothesis plausible. Firstly, even in the absence of external rewards, the feeling that we have completed a task successfully, and that we can cross an item off our to-do list, may be intrinsically rewarding. Confidence corresponds to this feeling of success in the task. This feeling is generally valid in perceptual[1,2] and memory tasks[3], where confidence typically correlates with performance, although some dissociations have been documented [4–6]. Therefore, in order to obtain rewards as soon as possible, people could seek to complete first the task in which they are more confident[7]. Another reason to expect that humans would use confidence to prioritize one task over another is to reduce their mental workload[8]. Indeed, having multiple pending tasks can divide our attention and diminish our cognitive resources, what would impact performance especially

on complex tasks. In order to face the most demanding tasks with a reduced mental workload, agents could complete first the less demanding tasks, in which they are more confident. To do so, they could compare their confidence across the different tasks at hand[9,10], and prioritize them accordingly.

To evaluate whether confidence guides task prioritization, we used a simple experimental paradigm involving elementary perceptual decisions. On each trial (see Figure 1A), participants (n=10) were presented an array of colored letters and they had to report both the dominant color (orange or blue) and the dominant letter (X or O) in the stimulus. Immediately after they had to give a confidence rating for each choice, on a subjective probability scale of success ranging from 50% (i.e. pure guess) to 100% (i.e. completely certain of being correct). By calibrating the stimulus before the main experiment (see methods), for each task we obtained an easy and a hard condition, which corresponded to an expected 90% and 60% of correct responses, respectively. As Figure 1B shows, observed performance in the experiment closely matched expected performance.

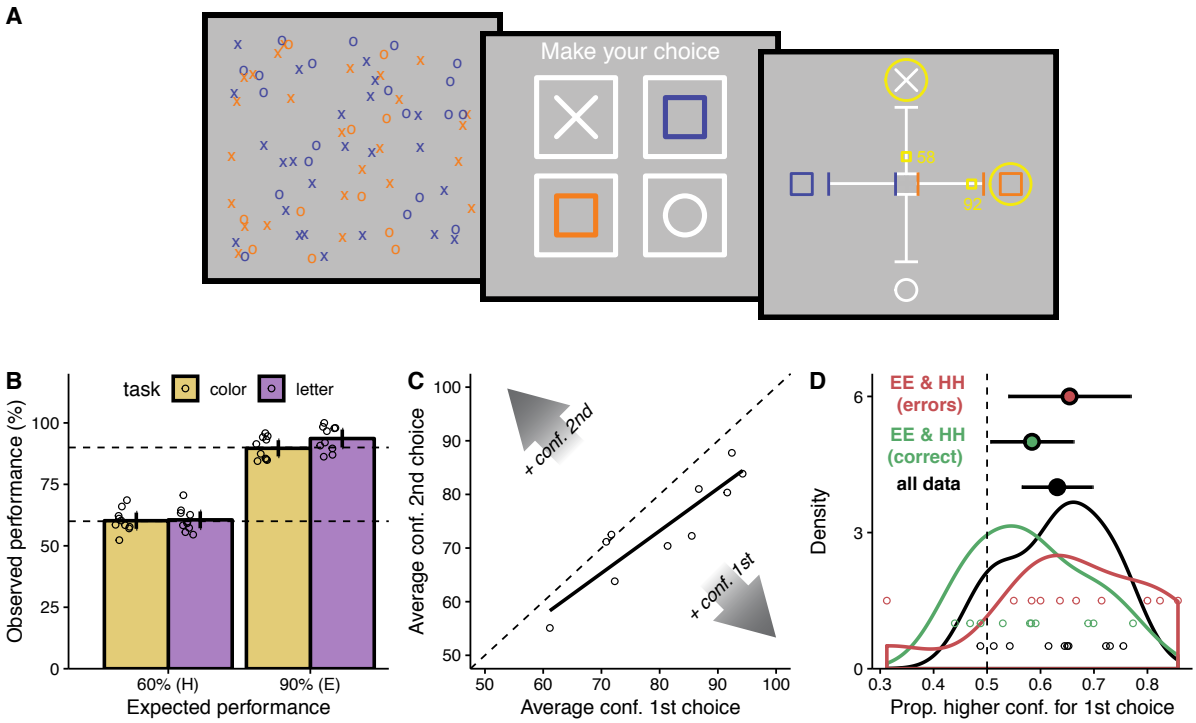

Figure 1 | Experiment 1: design and results. A. Overview of a trial. An array of 80 elements appeared for 1s. Each element could be an O or an X, colored blue or orange. After the stimuli disappeared, a response screen appeared. Four boxes were presented, each containing one of the four elements related to the choice. Participants gave their response to the color task by clicking on the box containing either the blue or the orange square, and to the letter task by clicking on the box containing either the O or the X. Next two scales, one for the color task and another for the letter task, appeared on the screen disposed as a cross. Participants could then rate their confidence on each task's decision, from 50 (total uncertainty) to 100 (total certainty). **B.** 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. **C.** Average confidence rating given for the second choice within a trial, as a function of that for the first choice. Dots represent individual participants, and the black, solid line represents a linear fit of the data. Participants below the dashed, identity line expressed more confidence in their first choices, and vice versa. **D.** Density of the proportion of trials where the first choice had a higher confidence rating than the second choice. The black density corresponds to the whole dataset, the green density corresponds only to those EE and HH trials where both choices were correct, and the red density corresponds only to those EE and HH trials where

both choices were errors. Densities are drawn by using each participant's data, with each participant's average shown as an individual dot at the bottom. Across-participant averages are displayed as big dots with error bars showing 95% confidence intervals. The dashed line corresponds to equal confidence ratings for first and second choices. The sample size for Experiment 1 was 10 participants, each participant completing 400 trials.

Critically, although participants were not asked to report their judgments in a specific order, we expected that they would report first the judgment associated with a higher confidence. This was indeed the case on average across trials: confidence in the first choice was higher than confidence in the second choice (Figure 1C, $t(9) = 4.569$, $p = 0.001$, 95% CI [3.516, 10.413], $d = 1.445$). It was also the case within each trial, as the proportion of trials in which the confidence associated to the first choice was greater than that associated to the second was systematically greater than 0.5 (black line in Figure 1D, $t(9) = 4.479$, $p = 0.002$, 95% CI [0.565, 0.698], $d = 1.416$). Furthermore, this pattern held even when both dimensions had the same difficulty level, either when we included only trials in which both responses were correct (green line in Figure 1D, $t(9) = 2.441$, $p = 0.037$, 95% CI [0.506, 0.662], $d = 0.772$) or trials where both responses were errors (red line in Figure 1D, $t(9) = 3.049$, $p = 0.014$, 95% CI [0.540, 0.769], $d = 0.964$). Note that, by construction, this later analysis confirms that participants' priority was driven by confidence per se, and not simply by task difficulty or response accuracy.

To strengthen this point, we conducted four more analyses on this data. First, in a logistic regression model, we found that responding to the color task first was significantly predicted from the difference in reported confidence between the two tasks ($\beta = 0.048$, S.E. = 0.002, $p < 0.001$), even when the difference in actual accuracy ($\beta = 0.002$, S.E. = 0.001, $p = 0.022$) and in expected performance ($p = 0.180$) between the two tasks were included as predictors in the same regression. Second, we compared this regression model with a simpler one that only included the difference between tasks in expected performance and in actual accuracy, and we found that the model including confidence explained the data significantly better than the model without confidence (log-likelihood: full model = -2266.989, no confidence model = -2529.497; $\chi^2(1) = 525.015$, $p < 0.001$). Third, in a different but similar approach, we could predict priority from confidence ($p < 0.001$) in a regression in which we included as an offset the prediction of priority by expected performance and actual accuracy. These analyses were conducted on the full data and are presented in more details in supplementary information 3. The previous analyses show that, despite the correlation between confidence and difficulty, we can demonstrate that confidence affects priority above and beyond task difficulty (expected accuracy) and actual accuracy. Our last analysis focused on a narrow subset of our data: namely, those trials where an error had been given a higher confidence rating than a correct response. Our goal was to know if, in these trials, there was a tendency to prioritize (i.e. respond to first) the task where the error was made over the task of the correct response. We took all those trials (pooling all participants together, due to the limited number of trials per participant) and performed a binomial exact test where the alternative hypothesis was that the proportion of trials where the error was prioritized was greater than 0.5. The test reached significance (actual proportion = 0.623, n = 472 trials, $p < 0.001$). Results were similar if only taking into account EE and HH trial types (proportion = 0.664, n = 217 trials, $p < 0.001$). Such finding revealed that, when confidence and accuracy were dissociated, the former still drove priority. This not only reinforces the idea that the role of confidence cannot be reduced to that of accuracy, but also suggests that the role of confidence is preponderant.

In the experiment above, participants may have associated their confidence ratings with task priority simply because they had to explicitly report their confidence. Our previous result (i.e. 'easy-first' bias in task prioritization) would be strengthened if it could be replicated by using an experimental paradigm that eliminated the possibility of it being caused by an implicit demand effect. In our second experiment, thus, we tested our hypothesis without asking for confidence ratings, by relying instead on task difficulty as a proxy for confidence. Indeed, in our first experiment we observed that reported confidence was greater for easy tasks than for hard tasks, and also greater for correct responses than for errors (for details see supplementary information 2.2.), as typically found in the confidence literature[1,3,11].

We therefore asked a new set of participants (n=20) to perform the same dual task without confidence ratings. Figure 2A depicts the proportion of times the first response was related to the color task, in each of the four possible kind of trials: easy color - easy letter (EE), easy color - hard letter (EH), hard color - easy letter (HE), and hard color - hard letter (HH). If confidence truly guided the priority of the responses, participants should exhibit a bias towards responding to the easy task before the difficult task. We found that, indeed, the color judgment was reported first on 42% of EH trials and on 37% of HE trials. This difference in proportion was highly consistent across participants ($t(19) = 3.059$, $p = 0.006$, 95% CI [0.018, 0.097], $d = 0.684$), as illustrated by Figure 2B. In other words, participants' task prioritization was affected by task difficulty, as a proxy for confidence, such that participants exhibited an easy-first bias. For completeness, we also confirmed that the same results were obtained in our previous study when using expected difficulty as a proxy for confidence (see supplementary information 2.3.).

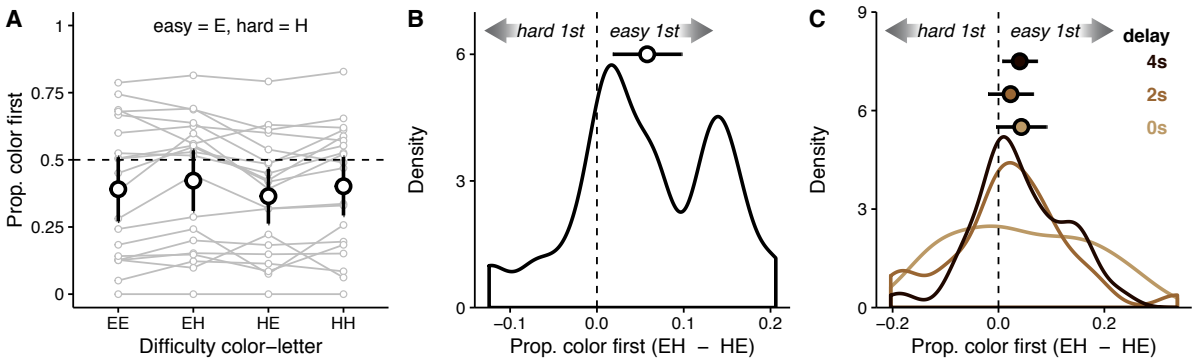

Figure 2 | Experiment 2 & Experiment 3: results. A. For Experiment 2, proportion of trials where the color task was first responded to, as a function of the type of trial (EE = easy color, easy letter; EH = easy color, hard letter; HE = hard color, easy letter; HH = hard color, hard letter). Each set of four dots connected by lines represents one participant. Bigger dots represent average proportions across participants, and error bars denote 95% confidence intervals. The dashed line corresponds to no preference towards responding first to the color or the letter task. **B.** For Experiment 2, density of the difference between proportion color task chosen first for EH trials and proportion color task chosen first for HE trials. The density is drawn by using the difference of proportions for each participant. The across-participant average is displayed as a big dot with error bars showing 95% confidence intervals. The dashed line corresponds to no difference between EH and HE trials. **C.** For Experiment 3, density of the difference between proportion color task chosen first for EH trials and proportion color task chosen first for HE trials. The three densities, one for each delay, are drawn by using the difference of proportions for each participant. Across-participant averages are displayed as big dots with error bars showing 95% confidence intervals. The dashed line corresponds to no difference between EH and HE trials. The sample size for Experiment 2 was 20 participants, each participant completing 400 trials. The sample size for Experiment 3 was 30 participants, each participant completing 400 trials.

In our third experiment, we wanted to discard one possible interpretation of our results in terms of response availability. In our previous experiments, the response screen was presented immediately after the stimuli disappeared, so one could argue that only the decision for the easy task had been reached by then. If that was true, then the easy-first priority effect could be due to response availability rather than confidence. To tackle this concern, we modified our previous experiment by introducing a delay (0s, 2s, or 4s, randomized across trials) between the stimulus offset and the onset of the response screen. We reasoned that, if the easy-first effect was only driven by response availability at the response screen onset, then it should disappear for long delays. We ran this experiment with new participants (n=30) and then calculated again the proportion of trials in which color was first responded to in EH trials and in HE trials, now for each delay condition. We found that the main effect of trial type was significant overall ($F(1,29) = 6.098$, $p = 0.020$, $\eta_p^2 = 0.174$), with an easy-first bias replicating our previous studies. And although the main effect of delay was also significant ($F(2,29) = 12.760$, $p = 0.020$, $\eta_p^2 = 0.306$), it did not interact with trial type ($p = 0.679$, see also Figure 2C). This data thus ruled out the concern that participants responded to the easy task first because it was the only task for which a decision was available by the time the response screen was presented.

We note nonetheless that it is still possible that participants answered the easy task first because it was the first for which a decision was made, even if both decisions had been reached by the time the response screen appeared. This possibility is difficult to eliminate when investigating priority and confidence about past stimuli, decisions and responses. However, it can be circumvented by relying on prospective estimates of difficulty and confidence, instead of retrospective ones. Prospective confidence corresponds to the anticipated probability of success one has before engaging in a task, based for instance on prior experience with the same task[12,13]. We implemented this idea in our next experiment. First we familiarized participants with two tasks, one easy and one hard. Then, we let them choose the order in which they would face the same two tasks in a subsequent test phase. In other words, unlike in our previous experiments where we measured priority from the ordering of the two responses, in this new design we could measure priority as the choice of engaging in one task before the other, very much like in the to-do list example.

With our fourth experiment, we thus evaluated whether task prioritization could be driven by prospective confidence as well. This experiment was conducted on a new sample of participants (n=31), which completed 32 experimental blocks, each constituted of a familiarization and a test phase (Figure 3A). In the familiarization phase, participants completed two sets of trials. Each set contained 6 trials, and was associated with a name and a task (either color or letter). Importantly, one set contained only hard trials, while the other set contained only easy trials. Difficulty was calibrated before the experiment (Figure 3B), and the order of the easy and hard set was random. After the familiarization with both sets was completed, the test phase started. Participants were presented with the names of both sets, and they could choose the order in which they wanted to complete 4 trials of each set. They knew that performance in these test trials would determine their final payoff. Critically, participants chose more often to face the easy set first (Figure 3C, $t(30) = 2.284$, $p = 0.030$, 95% CI [0.503, 0.557], $d = 0.410$). These results thus replicate the easy-first bias found in our previous experiments, and demonstrate that prospective confidence too can establish the priority of one task over another. Interestingly, whether the two sets involved

different tasks or the same task did not affect this priority effect (p = 0.256). This suggests that the comparison of the two sets involved relatively abstract representations, as was found for confidence comparison[9,14].

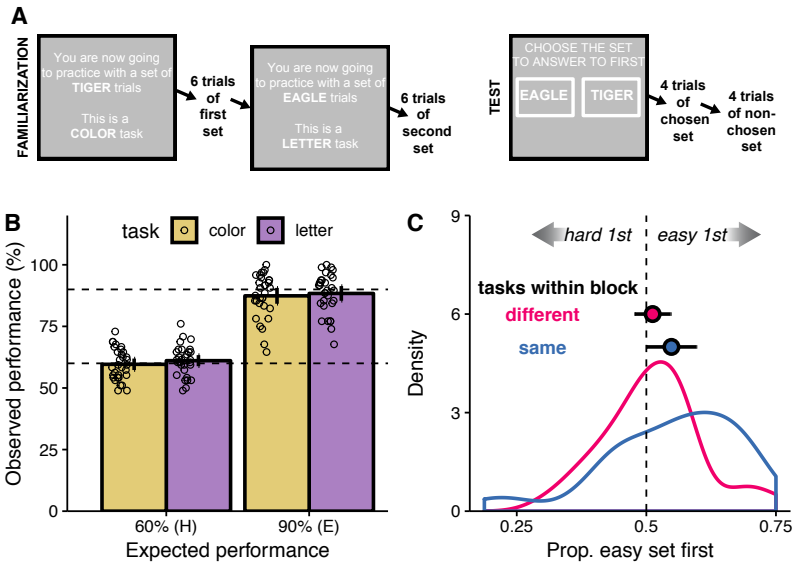

Figure 3 | Experiment 4: design and results. A. Overview of one block. The familiarization phase made participants practice with 6 trials of two different sets, with each set introduced by a screen stating its name and the type of task participants would have to respond to. During the test phase participants had to complete 4 trials of each set. At the beginning of the phase, a screen with the names of both sets was presented, and participants clicked on the name of the set they wanted to start with. **B.** For the familiarization phase, 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. **C.** Density of the proportion of blocks where participants chose to complete first the 4 trials from the set with an easy task. One of the displayed densities corresponds to blocks where both tasks were the same, the other to those where they were different. Both densities are drawn by using each participant's data. Across-participant averages are displayed as big dots with error bars showing 95% confidence intervals. The dashed line corresponds to easy and hard tasks being chosen equally often. The sample size for Experiment 4 was 31 participants, each participant completing 400 trials and making 32 set order choices.

Finally, we aimed at evaluating the generalizability of our findings beyond perceptual decisions. For this, we conducted another experiment to test whether the easy-first priority bias found in our previous studies would also manifest itself when considering simple problems of mental calculation. On each trial, two schematics of mental calculation problems were presented (Figure 4A). One was a priori more difficult than the other, in the sense that it included a multiplication or a division (e.g. "x+y/z"), instead of only additions or subtractions (e.g. "x+y-z"). The problems were otherwise matched in the number of elements and operations they involved. Participants had to click on one of the schematics to reveal the actual problem, which they then had to solve. Immediately after, participants had to solve the non-chosen problem. Although there was no difference in performance between easy and hard problems (p = 0.874), the response times for the hard problems were significantly longer (Figure 4B, t(83) = 9.079, p < 0.001, 95% CI [2.843, 4.438], d = 0.991), proving that they were more demanding. Most participants predominantly chose to solve the easy problem first (Figure 4C, t(83) = 5.247, p < 0.001, 95% CI [0.580, 0.677], d = 0.572). This experiment thus replicated and extended beyond perceptual tasks our previous findings of an easy-first bias in task prioritization.

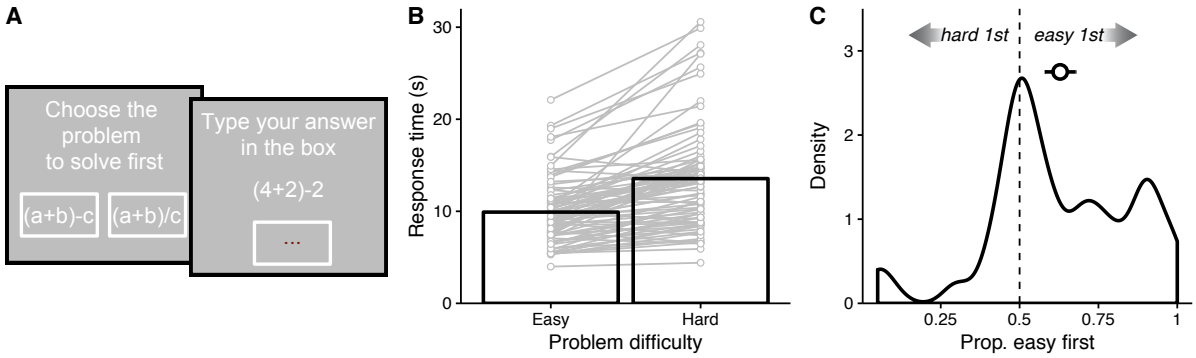

Figure 4 | Experiment 5: design and results. A. Overview of one trial. Two schematics of simple mental calculation problems were presented, one easier to solve than the other. Participants clicked on the one they wanted to solve first. The actual problem was then revealed, and participants had to type their response. Then they had to do the same with the other problem. **B.** Average response time, in seconds, as a function of the difficulty of the problem. Bars show the average across participants, while dots connected by lines represent each individual participant. **C.** Density of the proportion of blocks where participants chose to respond first to the easy problem. The density is drawn by using the proportion for each participant. The across-participant average is displayed as a big dot with error bars showing 95% confidence intervals. The dashed line corresponds to easy and hard problems being chosen equally often. The sample size for Experiment 5 was 84 participants, each participant completing 20 problem pairs.

To summarise, we found that, when facing task prioritization problems, participants exhibited a systematic bias towards addressing first the task in which they had greater confidence. We replicated the original finding in four more experiments. Converging evidence came from different measures, using task difficulty and observed accuracy as proxies for confidence, but also using actual confidence ratings. In terms of generalizability, we demonstrated that this effect was present in both situations involving retrospective and prospective confidence, and both situations involving perceptual and non-perceptual tasks, such as mental calculation.

Situations in which we face multiple tasks present a huge diversity, and task prioritization is likely determined by a varied set of factors, most of which were not manipulated in our design. For instance, prioritization can be driven by personal interest and the familiarity with the task[15]. We did also not manipulate the duration of each task, although there is evidence from real-life situations where people tended to approach first the shortest tasks[16]. Although in our experiments the two tasks were presented simultaneously, previous research has shown that task prioritization also depends on the time at which task-relevant information is presented[17]. We therefore only claim that confidence is one among many forces driving task prioritization. In our data, we find a large inter-individual variability in the general tendency to prioritize the color or the letter task (Figure 2A, see also supplementary information 5.1.), and we found that the locations of the different choice options on the screen could also bias task priority (see supplementary information 5.2.). Whether this variability is due to strategies or heuristics explicitly chosen by participants, or to differences in their perceptual or decision processes, remains to be investigated. Another factor that deserves more scrutiny is the reward associated with each task, which should affect task priority. In our experiments, we purposefully avoided any difference in reward between both tasks within a trial. We also did not instruct or incentivize participants to prioritize the more confident or the more accurate of the two decisions. In sum, nothing in the instructions or the reward structure pushed participants to prioritize one task over the other. Some of the aforementioned features could

be manipulated in future studies, and whether humans change their priorities under such manipulations is an open empirical question.

Another important issue for future research is whether this strategy of approaching the easy task first is adaptive. A priori, reducing the mental workload before tackling more difficult tasks could be beneficial. In our data, however, we found no clear evidence that addressing the easy task first led to better performance. Overall, participants were neither better nor worse when responding first to the easy task than when responding first to the hard task (see supplementary information 6.1.). To further investigate this issue, we conducted an additional experiment comparing a free response order condition (which replicated Experiment 2) to an imposed order condition in which the choice for the hard task was required either first or second (see supplementary information 6.2.). We found that participants' performances were similar between these two conditions, suggesting that, in our paradigm, the confidence-driven prioritization does not lead to an advantage in task performance. In other words, while we cannot claim that there is no advantage behind confidence-driven prioritization, our manipulation could not find it.

Our study highlights a new way in which confidence actively shapes our behavior. It therefore contributes to a growing literature showing how confidence is not a mere commentary on performance but actually affects individual or collective decisions. Previous research showed how, when agents interact, the confidence with which an advice is expressed affects the use of this information by others[18–20] and how well the group will perform[21]. At the individual level, confidence may serve as a teaching signal when feedback is unavailable[22–25]. It also influences the amount of resources we engage in a task[26,27]. Here we show that confidence does not only determine how we do a task, but also when we plan to do it.

Understanding how people decide to perform one task before another may have practical consequences for the management of individuals and organizations. This motivated some studies that focused on different applied scenarios. Some of this previous work already found that individuals address easier tasks first when they can decide their task schedule. For instance, students appear to prioritize easier course assignments over difficult ones[28], and the expected easiness of finding the relevant information affects their prioritization of a web-search task over another[15]. Physicians at an emergency department do prioritize easy patient cases, especially under high workload contexts[7]. The present study makes several unique contributions to this research topic: we extend this result to situations involving immediate decisions, we link the easy-first bias to confidence, and we offer a strict control over performance in the task via our psychophysical procedures. Finally, although we did not find a clear consequence from the easy-first bias on performance, previous research has shown that in another context this strategy was associated with both short-term benefits and long-term costs. Indeed, prioritizing easy cases allowed physicians to treat more patients per unit of time, but across individuals it was associated with a lower productivity and less revenue for the hospital[7]. Understanding whether and how similar short-term benefits and long-term costs may arise in other contexts, and in particular in situations where individuals have to perform multiple perceptuo-motor tasks at the same time, constitutes an exciting topic for future research.

Methods

Participants

Participants (Exp1=10, Exp2=20; Exp3=30; Exp4=31; Exp5=84) were healthy adults recruited from the Laboratoire d'Économie Experimentale de Paris. They reported normal or corrected-to-normal vision, and provided informed consent. Participation was compensated based on performance and duration, at an average rate of 12€ per hour. The Paris School of Economics ethics committee approved the experimental procedures after determining that they complied with all the ethical regulations.

Apparatus

Experiments were run using MATLAB Psychtoolbox[29] on a computer screen (17', 1024x768 pixel resolution, 60 Hz refresh rate) viewed from approximately 60 cm. Stimuli appeared on a grey background.

Perceptual task (experiments 1-4)

For experiments 1-4, the task consisted in reporting the dominant letter and the dominant color of an array of 80 letters (Xs or Os), colored blue or orange (see Figure 1A). Each trial started with a 500 ms fixation cross, followed by a 300 ms blank period, and then by the stimulus, which was presented for 1 s. Each letter, in Arial font, occupied approximately 0.5º (degrees of visual angle). Elements were presented within a 10º wide imaginary square. These stimuli were based on the code available at https://github.com/DobyRahnev/Confidence-leak and used in a previous publication[30].

After the stimuli disappeared, the response screen was presented, which consisted in 4 square boxes that contained a blue square, an orange square, a white O and a white X (see Figure 1A). The 4 boxes were randomly arranged in a 2x2 grid on each trial. Participants gave their response by clicking on a color box and a letter box, in their preferred order. A first response on one dimension (e.g. clicking on the X box) made the corresponding boxes disappear, leaving only the boxes for the other dimension (e.g. the orange box and the blue box), which required a second response.

Participants completed 4 parts of 4 blocks each, with 25 trials per block. Between block and block, participants could rest for 15 seconds, and at the end of each part they could take self-timed breaks.

Manipulation of task difficulty (experiments 1-4)

For each perceptual task, difficulty was manipulated by changing the proportion of the dominant over the non-dominant feature. After a short training, we used psychophysical staircases to estimate, separately for each task, the proportion of items that should be used for responses to be 90% correct (*easy* condition) or 60% correct (*hard* condition). Details of the training and staircase procedure are presented in supplementary information 1.1. We then used these parameters in the main experimental part to manipulate difficulty on the letter (easy vs. hard) and color (easy vs. hard) tasks, orthogonally within each participant, in

a 2x2 factorial design. We collected 100 trials for each of the 4 conditions, with trials of all conditions intermixed within blocks.

Experiment 1: confidence ratings

In Experiment 1, after participants made their two choices, they had to rate their confidence on each of them. Two confidence scales were simultaneously presented, one horizontal for the letter task and one vertical for the color task (see Figure 1A), or vice-versa (randomized across trials). Each scale featured the two response categories at its ends (which corresponded to 100% confidence) and an empty region at its center (near which confidence was at 50%). The two choices participants had made on that trial were surrounded by a yellow circle. Participants gave their confidence levels on both tasks by clicking on the corresponding scales, and then validated the ratings by clicking at the center of the screen. Participants were told that maximum uncertainty (no preference between their response and the other option within the same task) corresponded to a 50% confidence, and absolute certainty on the choice to a 100% confidence.

Experiment 2: choice-only

Trials in Experiment 2 closely resembled those in Experiment 1, except for the fact that no confidence had to be rated. Thus, the trial finished after participants made a choice for each task on the response screen.

Experiment 3: varying delays

In Experiment 3, we manipulated the time between stimulus offset and the presentation of the response screen, which was set to 0, 2 or 4 seconds (counterbalanced across trials).Considering the response times for our previous experiments (see supplementary information 2.4.), 4s ensured that participants had reached both decisions by the time the response screen was presented.

Experiment 4: anticipated difficulty in prospective confidence

In Experiment 4, participants only responded to one task on each trial. Therefore, the response screen uniquely contained the two relevant boxes. The main part was organized in 32 blocks. All blocks included a familiarization and a test phase (Figure 3A). During the familiarization phase, participants completed two sets of 6 trials. Each set was associated to a single task (color or letter), a difficulty level (easy or hard), and an animal name (randomly chosen without replacement from a predefined list of common names). Based on staircases ran before the main phase, the proportion for the task-relevant dimension was experimentally controlled, such that for each block, one set contained difficult trials and the other easy trials (for the non-relevant dimension, the proportion was always set at 50%). This manipulation was not revealed to participants but they could learn the difficulty of each set by performing the task during the familiarization phase. After the familiarization, the test phase started. Participants saw a screen with two horizontally aligned boxes presenting (in a random position) the animal names of the two sets just completed. 4 trials of each set had to be completed, with reward in that block depending on performance in that part. Participants decided the order in which to complete the test trials of each set by clicking on the name of the set they desired to complete first. We used animal names to avoid any obvious order

bias (set A vs B, or 1 vs 2) for this decision (for the complete list of set names, see supplementary information 7.1.). Across blocks, we manipulated factorially the task (color/letter) used for the first set in the block, the task (color/letter) used for the second set, and the order of the difficulty levels used for the two sets (easy first vs. hard first). This design resulted in 8 conditions, presented in a randomized order as a sequence 4 times per participant.

Experiment 5: anticipated difficulty in a mental calculation task

The stimuli used in Experiment 5 were different from those used in the rest of experiments. On each trial, two horizontally aligned boxes appeared (see Figure 4A), each containing a simple formula (e.g. “a+2b”) . Our formulas involved letters connected by basic operators (+, -, *, /) and were sometimes nested by brackets. Both problems were similar except for the operators connecting their respective letters. One of the problems (easy) contained additions and subtractions, while the other (hard) contained multiplications and divisions. The easy and hard box were randomly positioned on the left or the right of the screen on every trial. Participants had to solve both problems in their prefered order. By clicking on one box, the letters of the formula in that box were replaced by one-digit integers. Participants then typed their response, which appeared inside a box below the problem. They could change it as they wanted before pressing the ENTER key to validate it. Then, the non-clicked on problem appeared for participants to give their response through the same method. Each participant completed the same 20 pairs of questions (listed in supplementary information 7.2.), but in a random order. The initial presentations of formulae ensured participants could not try to solve the problems before making a choice.

Payoff

Performance on the main part of the experiments determined the final amount of money participants won. In experiments 2 and 3, for each of the 16 blocks, a random trial was chosen. Within that trial, the choice for either color or letter was also randomly picked. If that choice had been correct, the participant received 1 €, and 0 € otherwise. This payoff method ensured participants kept a stable performance and did not relax their attention or automatized their responses during some trials.

In Experiment 1, the payoff system was slightly different. Once a trial and a choice had been selected within a block, a number was selected from a uniform distribution between 0 and 100. If the number was lower than the confidence given for that choice, the payoff depended on performance. If the number was higher, another random number was sampled from a uniform distribution between 0 and 100. If this number was lower than the confidence value, 1€ was given, and 0€ otherwise. This system has previously[31] been adopted to ensure that participants try to accurately rate their confidence, since their payoff depends on it. On top of this, participants were awarded a fixed 3€ just for the added length reporting confidence gave to the experiment.

In Experiment 4, payoff was similar to Experiment 2, but given the more numerous and shorter blocks, each of them could be rewarded with only 0.5€. The rest between blocks was also shorter, with a duration of 10 s.

Experiment 5 had a payoff system which mimicked that of the previous experiments. At the end of the experiment, two trials were chosen randomly. For each of these selected trials, the answer to one of the problems was picked. If the answer was correct, participants received 1€, and nothing otherwise.

Before starting each experiment, participants were carefully informed about the procedure. The payoff structure and its objective were made clear, and the fact that reward did not depend on their response times was also pointed out.

Statistics

All reported t-tests and binomial exact tests are two-tailed. All ANOVA are repeated measures. For t-tests, effect size is reported by using Cohen's d_z, written here as d, which was calculated using Rosenthal's formula[32]: the t statistic was simply divided by the square root of the sample size. For ANOVA, effect size is reported by using partial eta squared (η_p^2). For each factor, this was calculated by taking the sums of squares of the condition, and dividing it by the sums of squares of the condition plus the sums of squares of the error[33]. Logistic regressions were performed with R's function glmer, by using maximum likelihood and a logit link function.

Code availability

The code used to produce the analyses and the figures that appear on this study are available from the corresponding author upon reasonable request.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Peirce, C. S. & Jastrow, J. *On Small Differences of Sensation*. (1885).
2. Fleming, S. M., Dolan, R. J. & Frith, C. D. Metacognition: computation, biology and function. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 1280–1286 (2012).
3. Dallenbach, K. M. The relation of memory error to time interval. *Psychol. Rev.* **20**, 323–337 (1913).
4. Wilimzig, C., Tsuchiya, N., Fahle, M., Einhäuser, W. & Koch, C. Spatial attention increases performance but not subjective confidence in a discrimination task. *J. Vis.* **8**, 7

(2008).

5. Zylberberg, A., Barttfeld, P. & Sigman, M. The construction of confidence in a perceptual decision. *Front. Integr. Neurosci.* **6**, (2012).

6. Rahnev, D. A., Maniscalco, B., Luber, B., Lau, H. & Lisanby, S. H. Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. *J. Neurophysiol.* **107**, 1556–1563 (2012).

7. Kc, D. S., Staats, B. R., Kouchaki, M. & Gino, F. Task Selection and Workload: A Focus on Completing Easy Tasks Hurts Long-Term Performance. *SSRN Electronic Journal* (2017). doi:10.2139/ssrn.2992588

8. Moray, N. *Mental Workload: Its Theory and Measurement*. (Springer Science & Business Media, 2013).

9. de Gardelle, V. & Mamassian, P. Does confidence use a common currency across two visual tasks? *Psychol. Sci.* **25**, 1286–1288 (2014).

10. de Gardelle, V. & Mamassian, P. Is confidence amodal? Confidence comparison across dimensions of a visual stimulus. *J. Vis.* **13**, 57–57 (2013).

11. Kepecs, A., Uchida, N., Zariwala, H. A. & Mainen, Z. F. Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455**, 227–231 (2008).

12. Wright, G. Changes in the realism and distribution of probability assessments as a function of question type. *Acta Psychol.* **52**, 165–174 (1982).

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

14. de Gardelle, V., Le Corre, F. & Mamassian, P. Confidence as a Common Currency between Vision and Audition. *PLoS One* **11**, e0147901 (2016).

15. Spink, A., Park, M. & Koshman, S. Factors affecting assigned information problem ordering during Web search: An exploratory study. *Inf. Process. Manag.* **42**, 1366–1378 (2006).

16. Ibanez, M., Clark, J. R., Huckman, R. S. & Staats, B. R. Discretionary Task Ordering:

Queue Management in Radiological Services. *Management Science* (2017). doi:10.2139/ssrn.2677200

17. Sigman, M. & Dehaene, S. Dynamics of the central bottleneck: dual-task and task uncertainty. *PLoS Biol.* **4**, e220 (2006).
18. Zalesny, M. D. Rater confidence and social influence in performance appraisals. *J. Appl. Psychol.* **75**, 274–289 (1990).
19. Paese, P. W. & Kinnaly, M. Peer Input and Revised Judgment: Exploring the Effects of (Un) Biased Confidence. *J. Appl. Soc. Psychol.* **23**, 1989–2011 (1993).
20. Zarnoth, P. & Sniezek, J. A. The Social Influence of Confidence in Group Decision Making. *J. Exp. Soc. Psychol.* **33**, 345–366 (1997).
21. Massoni, S. & Roux, N. Optimal group decision: A matter of confidence calibration. *J. Math. Psychol.* **79**, 121–130 (2017).
22. Hainguerlot, M., Vergnaud, J.-C. & de Gardelle, V. Metacognitive ability predicts learning cue-stimulus associations in the absence of external feedback. *Sci. Rep.* **8**, 5602 (2018).
23. Zylberberg, A., Wolpert, D. M. & Shadlen, M. N. Counterfactual Reasoning Underlies the Learning of Priors in Decision Making. *Neuron* **99**, 1083–1097.e6 (2018).
24. Guggenmos, M., Wilbertz, G., Hebart, M. N. & Sterzer, P. Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *Elife* **5**, (2016).
25. Daniel, R. & Pollmann, S. Striatal activations signal prediction errors on confidence in the absence of external feedback. *Neuroimage* **59**, 3457–3467 (2012).
26. van den Berg, R., Zylberberg, A., Kiani, R., Shadlen, M. N. & Wolpert, D. M. Confidence Is the Bridge between Multi-stage Decisions. *Curr. Biol.* **26**, 3157–3168 (2016).
27. Desender, K., Boldt, A. & Yeung, N. Subjective Confidence Predicts Information Seeking in Decision Making. *Psychol. Sci.* **29**, 761–778 (2018).
28. Puffer, S. M. Task-Completion Schedules: Determinants and Consequences for Performance. *Hum. Relat.* **42**, 937–955 (1989).
29. Brainard, D. H. The Psychophysics Toolbox. *Spat. Vis.* **10**, 433–436 (1997).

30. Rahnev, D., Koizumi, A., McCurdy, L. Y., D'Esposito, M. & Lau, H. Confidence Leak in Perceptual Decision Making. *Psychol. Sci.* **26**, 1664–1680 (2015).

31. Massoni, S., Gajdos, T. & Vergnaud, J.-C. Confidence measurement in the light of signal detection theory. *Front. Psychol.* **5**, 1455 (2014).

32. Rosenthal, R. (1991). Meta-analytic procedures for social research. Newbury Park, CA: SAGE Publications, Incorporated.

33. Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Frontiers in psychology, 4, 863.

Acknowledgements

This work was supported by the Agence Nationale de la Recherche [grant ANR-16-CE28-0002 to VdG]. We thank Maxim Frolov and Audrey Azerot for helping in data collection. We thank Mahiko Konishi, Jean-Christophe Vergnaud, Jérôme Sackur, Bruno Berberian and Pascal Mamassian for discussions.

Author contributions

DAL and VdG conceived the study and designed the experiments. DAL programmed the stimuli. DAL, ML and VdG collected and analysed the data. DAL and VdG wrote the manuscript. All authors approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.