Supplementary materials

Post-decision wagering after perceptual judgments reveals bi-directional certainty readouts

Caio M. Moreira, Max Rollwage, Kristin Kaduk, Melanie Wilke, Igor Kagan

Archived data

The data for this article are available at the Open Science Framework, https://osf.io/ys8cj

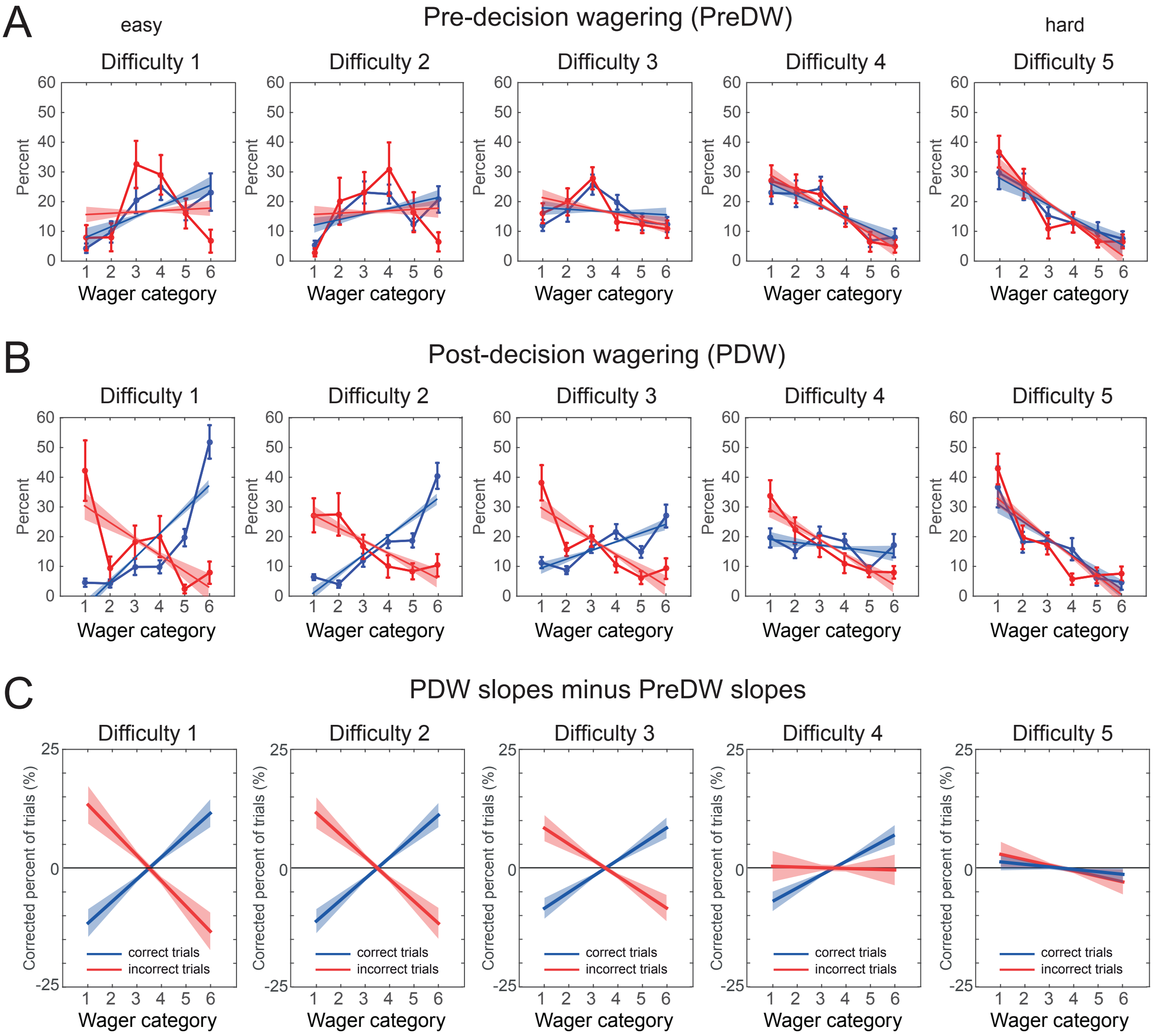
S.1 Supplementary Results

S.1.1 Wager-specific proportions of trials and slopes for the five difficulty levels

The Supplementary Figure S1A and Supplementary Figure S1B illustrate – for PreDW and PDW, respectively – the wager-specific proportions of correct and incorrect trials separately for the five difficulty levels. These proportions resulted in the slope-based measures illustrated in the third row of the Supplementary Fig. S1C and in the Figure 10. To test, within each difficulty level, if slope-correct was different from slope-incorrect depending on the trial type (PreDW or PDW), we performed two-way ANOVAs for repeated measures. In none of the difficulty levels PreDW slope-correct was significantly different from PreDW slope-incorrect ($p>0.05$; Supplementary Fig. S1A), indicating reliable baselines for the five difficulty levels separately.

Next, we performed paired t-tests within each difficulty level to test for significant differences between PDW and PreDW slopes. The t-tests revealed that subjects' PDW slopes-correct were different from their PreDW slopes-correct at all difficulty levels ($p<0.01$), except for the highest difficulty level 5 ($p>0.05$). Additionally, PDW slope-incorrect was not different from PreDW slope-incorrect for the difficulty levels 4 and 5 ($p>0.05$), while subjects were able to read out certainty of being incorrect at the difficulty levels 1, 2 and 3 ($p<0.01$).

The Supplementary Figure S1 reveals two important points that are considered in our analyses. Firstly, the baseline condition (PreDW trials) showed the general (non-trial-specific) effect of expected perceptual difficulty assessments. For instance, the realization of increased family difficulty made subjects to use low wagers increasingly but independently of the correctness of the trials. Without considering this baseline measurement, we would not be able to distinguish this adaptive strategy (wagering low for hard trials) from the use of trial-specific certainty during PDW trials. Even if the task would contain only one difficulty level, or maybe especially in those cases, the baseline measurement is essential to quantify certainty readouts taking into account individual biases. Secondly, during PreDW trials (especially during expected easier trials, partially predictable by the family difficulty) subjects chose more often middle wagers (wagers 3 and 4). Since linear fits captured wagering trends regardless of the effect of such behavior, our choice for using the slopes of linear fits (instead of the best-fitted curves) proved to be valuable for establishing useful baselines for the slope-based measurements.



Supplementary Figure S1. (A) Means and standard errors of linear fits for correct trials (blue lines and shaded bands) and incorrect trials (red lines and shaded bands) for (A) PreDW (baseline) and (B) PDW, fitted to the data: means and standard errors of wager-specific proportion of correct (blue curves) and incorrect (red curves) trials, for each difficulty level. (C) Means and standard errors of PDW slope-correct minus PreDW slope-correct (blue line and band) and of PDW slope-incorrect minus PreDW slope-incorrect (red line and band) for each difficulty level (same data as in Figure 10). Measurements represent averages across subjects.

The frequency of the use of each wager (means and standard errors from the lowest to the highest wager) in PDW trials (wagers 1 to 6: 19±2% 11±1% 16±1% 16±2% 13±1% 25±3%) and in PreDW trials (wagers 1 to 6: 15±2% 19±3% 23±3% 19±2% 11±2% 13±3%) showed that subjects used the entire wager scale.

S.1.2 The increase in Type 1 performance after wagering high in PreDW

Supplementary Figure S1A also revealed that there was an increase in Type 1 performance after wagering high in PreDW trials. For the low difficulties 1 and 2, this effect was significant for the highest wager 6, in isolation, and also reached significance after correcting for multiple comparisons across all 6 wagers for the difficulty 2 (difficulty 1: $t_{16}=2.45$, $p=0.026$; difficulty 2: $t_{16}=3.10$, $p=0.0068$; $p=0.041$, Bonferroni-corrected). We interpret this effect as a form of attentional mobilization, leading to an increase in performance after wagering high. This finding resembles a similar influence of subjective beliefs about own competency that were induced by manipulated social-comparative performance feedback and fictional research findings (Zacharopoulos et al., 2014).

S.1.3 Perceptual decision criteria for the five difficulty levels

We calculated the perceptual decision criterion separately for the five different difficulty levels using Signal Detection Theory approach. This information is relevant because subjects might develop different spatial biases for different difficulties (e.g. select more often the image on the right side in harder trials), making it impossible to compare different difficulty levels. The two-way ANOVA for repeated measures revealed that perceptual decision criterion did not differ from zero ($F_{1,16}=0.009$, $p=0.93$) or among the difficulty levels ($F_{4,64}=0.248$, $p=0.91$), suggesting that subjects identified the match on the right or left side of the screen with the same probability at all difficulty levels.

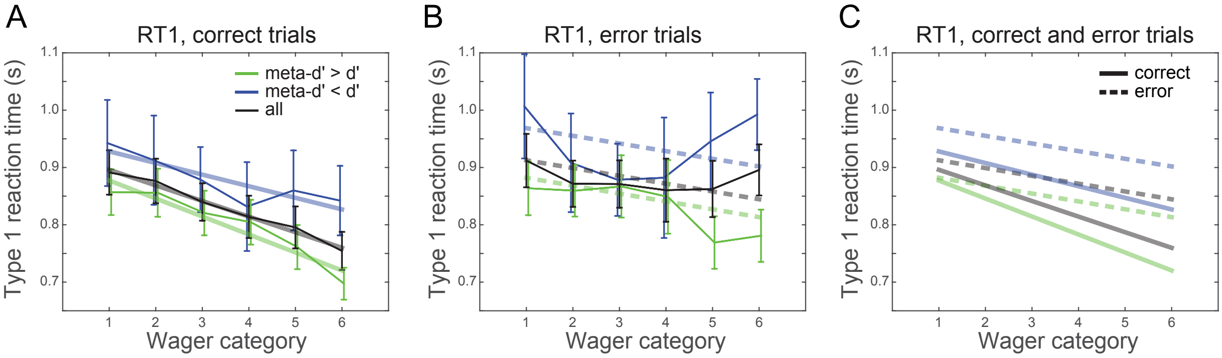
S.2 Supplementary Discussion

Different aspects of the task design might prompt or limit post-decisional evidence accumulation and, consequently, bi-directional certainty readouts. We discuss these aspects below together with our results.

Time pressure: Yeung and Summerfield (2012) and others have proposed that late drifts towards the correct option, not considered by the time subjects commit to the wrong option, generate extra evidence used for error detection. In the study of Charles et al. (2013), for example, subjects who needed to report Type 1 decisions within 1 s showed higher metacognitive efficiency (meta-d'/d', probably because the time pressure lowered d' but not meta-d'), than those who had twice the time to report their perceptual decisions. These results indicate that time pressure over Type 1 decisions increased the use of extra information during Type 2 decisions. In our experiment, the average readout of certainty of being incorrect and the predominance of subjects with meta-d'>d' suggest that the applied time pressure (1.5 s) was enough to restrict evidence accumulation before Type 1 decisions and favor the use of extra evidence from post-decisional accumulation or a parallel metacognitive processing.

Memory: memory-based tasks such as DMTS might promote the use of short-term memory as another source of extra information for the Type 2 decisions. Since the basic information required for Type 1 decisions, the previously presented sample, can only be accessed through top-down memory retrieval, such mechanism – which does not depend exclusively on continuous input of sensory evidence – might continue to provide information also after Type 1 decisions (Magnussen and Greenlee, 1999; Yu et al., 2015).

Propriosensory evidence: the propriosensory evidence related to the manual report of the Type 1 decision itself can serve as another source of post-decisional information. It is known that Type 1 reaction times correlate with certainty (Fetsch et al., 2014; Kiani and Shadlen, 2009). Therefore, subjects could, in principle, read out those reaction times – instead of or in addition to the perceptual evidence – in order to judge their Type 1 performance during the Type 2 decisions. In accordance with this reasoning, Fleming et al. (2015) modified subjects' Type 2 decisions by manipulating the activity of motor areas prior to metacognitive reports. In our experiment, however, the distribution of wager-specific Type 1 reaction times (RT1) was unidirectional (decreasing towards the highest wager, Figure 9A). Even if RT1s were bi-directionally distributed, e.g. similar to metacognitive Type 2 RT2 inverted U-shape, this would only allow reading out certainty level, but not *certainty direction* (i.e. correct or incorrect). Additional possibility could be that distributions and ranges of correct and incorrect RT1s differed enough to provide a probabilistic readout of both certainty level and direction. This does not seem to be the case in our data (see Supplementary Figure S2). While the slopes fitted to incorrect wager-specific RT1s were shallower than for correct RT1s (RT1 slope-incorrect = -0.0136; RT1 slope-correct = -0.0301; $t_{16}=-2.148$, $p<0.05$), their distributions overlapped substantially, especially in the low wagers. For example, in the high metacognitive efficiency group, both correct and incorrect RT1 distributions were unidirectional, with negative slopes originating from the same range in the lowest wager. These patterns render it unlikely to enable inferring the direction of certainty from a single trial RT1 'sample'. In summary, the readouts of RT1 in our experiment could not provide the bi-directional certainty readouts apparent in slope-based measurements and wager-specific Type 2 reaction times.



Supplementary Figure S2. Means and standard errors of wager-specific perceptual reaction times (RT1) across all subjects (black curve), for the high metacognitive efficiency group (meta-d'>d', green curve) and for the low metacognitive efficiency group (meta-d'<d', blue curve), separated to correct trials (A) and incorrect trials (B). The linear regression slopes (thick lines) are means of the slopes that were fitted to the individual subject averages. Note that several subjects did not select high wagers in incorrect (error) trials; therefore data for those wagers represent only a subset of subjects while slopes include all subjects (in subjects with missing wagers, the slopes were fitted to available wagers). This led to a partial mismatch between the data and fitted slopes in B. (C) Direct comparison of correct trial (solid lines) and error trial (dashed lines) slopes.

Rewards and punishments: lastly, the use of PDW might favor evidence accumulation because it motivates subjects to fully explore their sources of information through gains and losses (i.e. profit more when correct and avoid losses when incorrect). We suggest that the monetary motivational aspect of PDW also contributed to the predominance of subjects who showed high metacognitive efficiency (11 out of 17 subjects).

Supplementary References

Charles, L., Van Opstal, F., Marti, S., & Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. NeuroImage, 73, 80–94. http://doi.org/10.1016/j.neuroimage.2013.01.054

Fetsch, C. R., Kiani, R., Newsome, W. T., & Shadlen, M. N. (2014). Effects of Cortical Microstimulation on Confidence in a Perceptual Decision. Neuron, 84(1), 239. http://doi.org/10.1016/j.neuron.2014.09.020

Fleming, S. M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T., & Lau, H. (2015). Action-specific disruption of perceptual confidence. Psychological Science, 26(1), 89–98.

Kiani, R., & Shadlen, M. N. (2009). Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex. Science, 324(5928), 759–764. http://doi.org/10.1126/science.1169405

Magnussen, S., & Greenlee, M. W. (1999). The psychophysics of perceptual memory. Psychological Research, 62(2-3), 81–92.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1594), 1310–1321. http://doi.org/10.1098/rstb.2011.0416

Yu, S., Pleskac, T. J., & Zeigenfuse, M. D. (2015). Dynamics of postdecisional processing of confidence. Journal of Experimental Psychology: General, 144(2), 489–510. http://doi.org/10.1037/xge0000062

Zacharopoulos, G., Binetti, N., Walsh, V., & Kanai, R. (2014). The effect of self-efficacy on visual discrimination sensitivity. *PloS one*, *9*(10), e109392. https://doi.org/10.1371/journal.pone.0109392