Confidence ratings do not distinguish imagination from reality

Nadine Dijkstra, Matan Mazor & Stephen M. Fleming

Perceptual reality monitoring refers to the ability to distinguish internally triggered imagination from externally triggered reality. Such monitoring can take place at perceptual or cognitive levels – for example, in lucid dreaming perceptual experience feels real but is accompanied by a cognitive insight that it is not real. We recently developed a paradigm to reveal perceptual reality monitoring errors during wakefulness in the general population, showing that imagined signals can be erroneously attributed to perception during a perceptual detection task. In the current study, we set out to investigate whether people have insight into perceptual reality monitoring errors by additionally measuring perceptual confidence. We used hierarchical Bayesian modelling of confidence criteria to characterize metacognitive insight the effects of imagery on detection. Over two experiments, we found that confidence increased in tandem with perceptual detection during congruent imagery, indicating a failure of reality monitoring not only at a perceptual, but also at a metacognitive level. These results further show that such failures have a perceptual rather than a decisional origin. Interestingly, offline queries at the end of the experiment revealed global, task-level insight, which was uncorrelated with local, trial-level insight as measured with confidence ratings. Taken together, our results demonstrate that confidence ratings do not distinguish imagination from reality during perceptual detection. Future research should further explore the different cognitive dimensions of insight into reality judgements and how they are related.

Perceptual reality monitoring—inferring whether sensory signals reflect reality or imagination—operates at perceptual and metacognitive levels (Dijkstra et al., 2022). Generally, perceptual experiences of reality correlate with metacognitive beliefs that such experiences indeed reflect reality. For example, during dreams or hallucinations, imagined content is perceived as real and also believed to reflect reality (Siclari et al., 2017; Zmigrod et al., 2016). However, there are cases where perception and metacognition of reality diverge. For example, in Charles Bonnet Syndrome, a condition in which visual impairment is associated with the development of hallucinations, patients generally have insight into the unreality of their experiences (Menon et al., 2003). Another example is lucid dreaming, where perceptual experience still 'feels real' but people are aware that they are in fact dreaming (Baird et al., 2019; Corlett et al., 2014; Konkoly et al., 2021).

One way to objectively characterize dissociations between perceptual and metacognitive processes is to ask how people evaluate confidence in their percepts (Fleming & Daw, 2017; Fleming & Lau, 2014). Metacognitive insight into effects of imagery on perception can then be measured as shifts in confidence: if participants *perceive* illusory objects when imagining and at the same time *know* they have this tendency, they might report seeing illusory objects but be less confident in these reports. If, on the other hand, they don't know they have this tendency, subjective confidence should exhibit the same qualitative effects as perceptual decisions.

More precisely, if confidence shifts in tandem with biases in perception, this indicates an absence of metacognitive insight. In line with this idea, recent studies of perceptual illusions (such as motion and colour afterimages) have documented confidence shifts in tandem with idiosyncratic perceptual biases (Gallagher et al., 2019; Mamassian & de Gardelle, 2022). Indeed, changes in confidence have been proposed as a diagnostic feature of truly perceptual (as opposed to decision-level) biases (Gallagher et al., 2019). In turn, recent computational models have suggested that perceptual confidence reflects a probability that one's response is self-consistent, rather than objectively correct (Boundy-Singer et al., 2022; Mamassian & de Gardelle, 2022). In contrast, other studies have documented cases in which participants' confidence shows tell-tale signs of insight into a first-order decision bias, with so-called "counterfactual" confidence being used to update prior beliefs (Zylberberg et al., 2018). Similarly, cases in which metacognitive sensitivity is greater than performance have also been documented (Scott et al., 2014), indicating that some evidence may be accessible to confidence judgements that is not incorporated into perceptual decisions.

In this study we aimed to characterize whether participants have insight into perceptual reality monitoring errors by investigating perceptual confidence in a scenario where both imagery and perception are at play. We have previously found that simultaneously imagining congruent stimuli during perceptual detection leads to an increase in presence responses, indicating that imagined signals are sometimes mistaken for perception (Dijkstra, Kok, et al., 2021; Dijkstra, Mazor, et al., 2021; Dijkstra & Fleming, 2023). Here, we investigated to what extent participants are metacognitively aware of these reality monitoring errors by additionally measuring confidence. We operationalise (lack of) insight here as the extent to which confidence shifts in tandem with imagery-induced biases in perception. Within signal detection theoretic models of confidence, such shifts can be quantified as the distance between decision criteria and confidence criteria – if such a distance remains invariant under shifts of decision criteria, this would reflect a lack of insight. To evaluate this hypothesis we extended a hierarchical Bayesian model of confidence ratings (Fleming, 2017) to include a prior over the distances between decision and confidence criteria, enabling us to infer metacognitive insight about perceptual reality monitoring errors.

Mistaking imagination for reality during perceptual detection

In a first experiment, 102 participants performed a perceptual detection task on gratings that gradually appeared within dynamic noise while simultaneously imagining either the same grating (congruent) or a grating orthogonal to the one they were detecting (incongruent) or nothing (no imagery; Fig. 1A). In line with previous findings, we observed a significant decrease in decision criterion (that is, an increased tendency to report stimulus presence) specifically for congruent imagery (*M* = 0.01, *SD* = 0.73) compared to no imagery (*M* = 0.22, *SD* = 0.57, *t*(101) = 3.10, *p* = 0.0025, *d* = 0.32, 95% *CI* difference = 0.07 – 0.34; Fig. 1B). While the criterion for congruent imagery was numerically lower than that for incongruent imagery (*M* = 0.16, *SD* = 0.61), this difference did not reach significance (*t*(101) = 1.98, *p* = 0.0502, *d* = 0.23, 95% *CI* difference = -0.0002 – 0.31). Finally, there was no significant difference in criterion between incongruent imagery and no imagery (*t*(101) = -0.96, *p* = 0.341, *d* = -0.09, 95% *CI* difference = -0.16 – 0.06). Instead, and consistent with previous findings, there was a significant decrease in d' during incongruent imagery (*M* = 2.22, *SD* = 1.54) compared to congruent imagery (*M* = 2.63, *SD* = 1.76; *t*(101) = -2.50, *p* = 0.014, *d* = -0.25, 95% *CI* difference = -0.74 - -0.09). Together, these results demonstrate that participants were more likely to indicate perceptual presence when simultaneously imagining the same stimuli while incongruent imagery merely decreased performance (Fig. 1D). These results are in line with previous studies and are interpreted to reflect perceptual reality monitoring errors: participants are more likely to say there was a grating on the screen when the same stimuli were both imagined and perceived (Dijkstra, Kok, et al., 2021; Dijkstra, Mazor, et al., 2021; Dijkstra & Fleming, 2023).

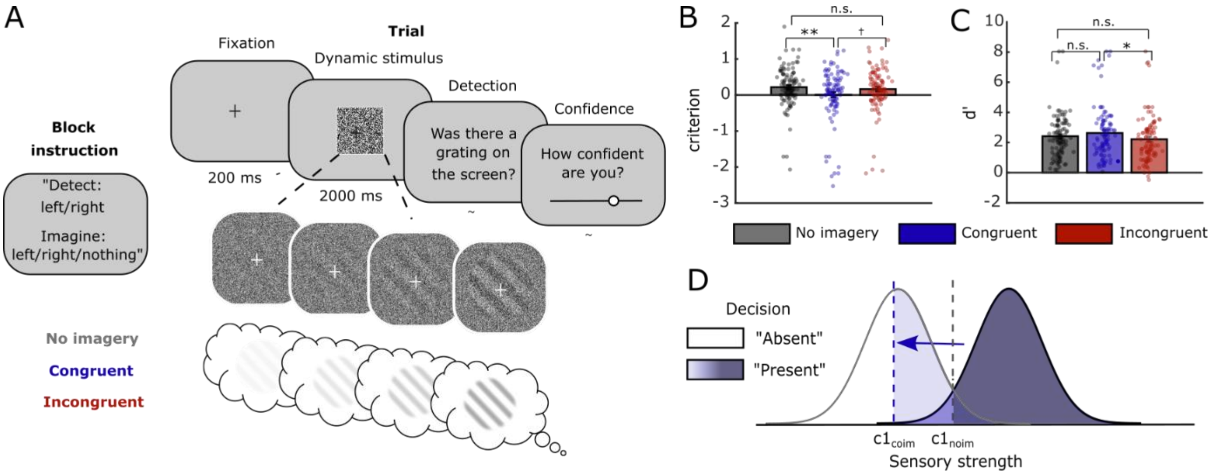

Figure 1. Experimental design, decision level responses and model. (**A**) Participants were instructed to detect oriented gratings in noise while simultaneously imagining the same grating (congruent), a grating perpendicular to the to-be-detected stimulus (incongruent) or nothing (no imagery). After each trial, participants indicated whether a stimulus had been presented on the screen and after that indicated the confidence in their answer from 'complete guess' to 'absolutely certain' by moving a slider with their mouse. (**B**) Decision criterion was significantly lower during congruent imagery compared to no-imagery and marginally lower during congruent imagery compared to incongruent imagery. There was no significant difference in criterion between incongruent imagery and no imagery. (**C**) In contrast, there was no effect of d' during congruent imagery, but there was a significant decrease in d' during incongruent imagery. (**D**) Signal detection theory (SDT) model of (congruent) imagery increasing perceptual presence responses by decreasing the decision level criterion. Within SDT, a decrease in criterion is equivalent to an increase in sensory strength. $c1_{noim}$ = first-order criterion during no-imagery; $c1_{coim}$ = first-order criterion during congruent imagery. * $p < 0.05$; ** $p<0.005$; † $p<0.06$; n.s. $p>0.1$.

Quantifying insight into perceptual reality errors

We next aimed to characterize insight into these reality monitoring errors using a signal detection theoretic model of metacognition. As noted above, within a signal detection framework, congruent imagery leads to a more liberal decision criterion and more “presence” responses. Subjects who are metacognitively aware of this criterion shift should be less confident in their “presence” responses when they are imagining the target stimulus, because a liberal criterion means they are more likely to commit a false alarm. Similarly, they should be more confident in their “absence” responses when imagining the target stimulus, because a liberal criterion means they are less likely to miss presented stimuli. In contrast, if subjects have no metacognitive insight into the effects of imagery on perception, their confidence ratings in “presence” and “absence” responses should be equivalent to those given for veridical perception.

These effects can be quantified by tracking the position of their metacognitive (confidence) criteria: the cut-off points at which participants rate their decision as high versus low confidence (Fig. 2A-C). If participants have no insight into the influence of imagery on their responses, metacognitive criteria should shift in tandem with the decision-level criterion. Counter-intuitively, if confidence criteria shift in tandem with the decision criterion, the two will cancel each other out and there will be no change in confidence ratings compared to no imagery conditions (Fig. 2A). In contrast, if participants are aware that imagery increases perceptual presence, metacognitive criteria should remain closer to the no-imagery condition, leading to imagery-induced presence responses being associated with lower confidence (Fig. 2B). Finally, participants might have an incorrect insight and believe that engaging in imagery in fact decreases perceptual presence. This would instead lead to an overshoot in the shift in metacognitive criteria, leading to higher confidence presence responses during imagery.

In order to characterize these relative shifts in metacognitive criteria, we extended the hierarchical meta-d’ model (Fleming, 2017; Maniscalco & Lau, 2012). Specifically, we modelled the second-order (metacognitive) criteria on confidence as free parameters and allowed asymmetries between the distances from the decision criterion to the negative (‘absence’) and positive (‘presence’) confidence criteria, allowing us to model a full range of insight profiles (cf. Fig. 2A and Fig. 2B-C). We first validated this new analysis pipeline on simulated data (Fig. 2D-G). We generated confidence ratings for three scenarios: one in which the confidence criteria moved with the decision-level criterion shift, indicating no insight (Fig. 2D), one in which the confidence criteria remained at the no-imagery values, indicating insight (Fig. 2E) and one in which the confidence criteria moved beyond the decision-level criterion shift, indicating incorrect insight (Fig. 2F). In all scenarios, our model was able to successfully recover the confidence criteria.

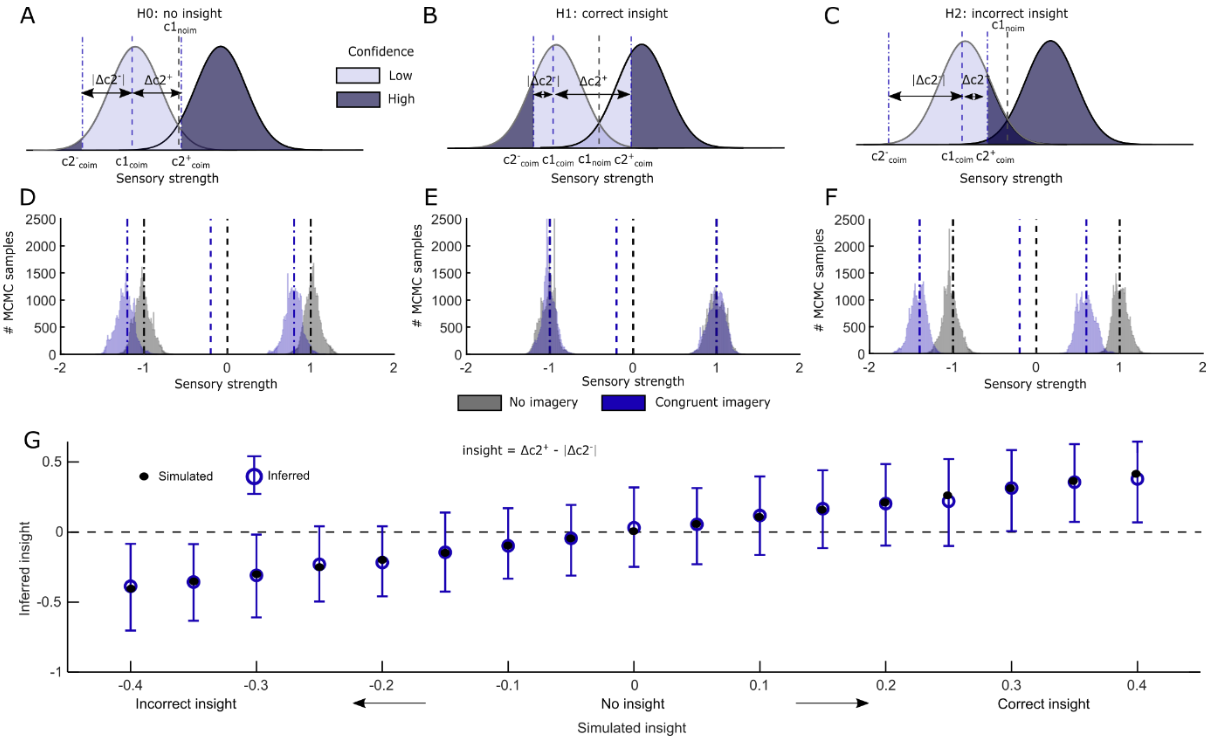

Figure 2. Modelling insight into perceptual reality monitoring. (**A**) If participants have no insight into their first-order criterion shift, we would expect confidence criteria to move with the decision-level criterion shift so that, on average, presence responses are more confident. coim = congruent imagery; c1 = first-order criterion; $c2^{-}$ = negative second-order (metacognitive) criterion; $c2^{+}$ = positive second-order criterion; Δc2+/- = c2 relative to c1 (i.e. c2-c1). (**B**) In contrast, if participants have insight into the fact that imagery increases perceptual presence, their confidence criteria would remain closer to the no-imagery criterion, leading to a decrease in confidence for presence responses. (**C**) Finally, if participants incorrectly believe that their imagery decreases perceptual presence, they might overcompensate by decreasing their confidence criteria, leading to an overly exuberant increase in confidence for presence responses. (**D**) Simulation of no insight during imagery within the HMeta-d model: second-order confidence criteria ($c2^{+/-}$) move along with the first-order criterion shift. Dashed lines indicate first-order criteria, dashed-dot lines indicate second-order criteria; MCMC = Markov Chain Monte Carlo. (**E**) Simulation of insight during imagery: confidence criteria during imagery are similar to during no imagery. (**F**) Simulation of wrong insight during imagery: confidence criteria move beyond the first-order criterion shift. (**G**) Parameter recovery of confidence criteria shift across a number of insight levels (similar to D-F) and associated 95% high-density interval (HDI) of MCMC samples.

Within this framework, insight is quantified as the asymmetry between the positive and negative confidence criteria relative to the first-order decision criterion: if there is no insight, the confidence criteria will move in tandem with the decision criteria, leading to symmetrical distances between c1 and c2$^+$ and between c1 and c2$^-$ (Fig. 2A). In contrast, in the case of a correct insight, the confidence criteria will remain close to the no-imagery case despite the decrease in first-order criterion. This will result in a larger distance between c2$^+$ and c1 than between c2$^-$ and c1, leading to a decrease in confidence for presence responses (Fig. 2B). Finally, if the insight is incorrect and the belief is that imagery decreases perceptual presence, the metacognitive criteria will move in the other direction, leading to a smaller distance between c2$^+$ and c1 than between c2$^-$ and c1 (Fig. 2C). Taken together, a difference close to zero between the positive Δc2+ and negative Δc2- first-to-second order distances indicates no insight, a difference above zero indicates correct insight and a difference below zero indicates incorrect insight. In a final simulation, we showed that our analysis pipeline was able to accurately recover insight across a wide range of simulated parameter values (Fig. 2G).

Confidence ratings reveal absence of insight into perceptual reality errors

After successful validation of our analysis pipeline, we next turned to our empirical confidence ratings. We first ran a repeated-measures ANOVA on the confidence ratings and found no effects of condition (Fig. 3A), suggesting that participants were not adjusting their confidence criteria when engaged in simultaneous imagery. To directly investigate the relationship between first and second-order criterion shifts, we fitted our extension of the hierarchical meta-d model to the data. These results show that the asymmetry between the negative and positive confidence criteria was centred around 0, indicating no insight (Fig. 3B&C; *M* = -0.07, 95% *HDI* = -0.33 – 0.22). The Savage-Dickey density ratio indicated a BF_{01} of 86.62, which means that it is ~87 times more likely that there is no difference between the negative and positive second-level criterion than that there is a difference. In other words, confidence criteria moved along with the decision criterion shift, suggesting that participants did not have metacognitive insight into the fact that imagery was increasing their perceptual presence responses.

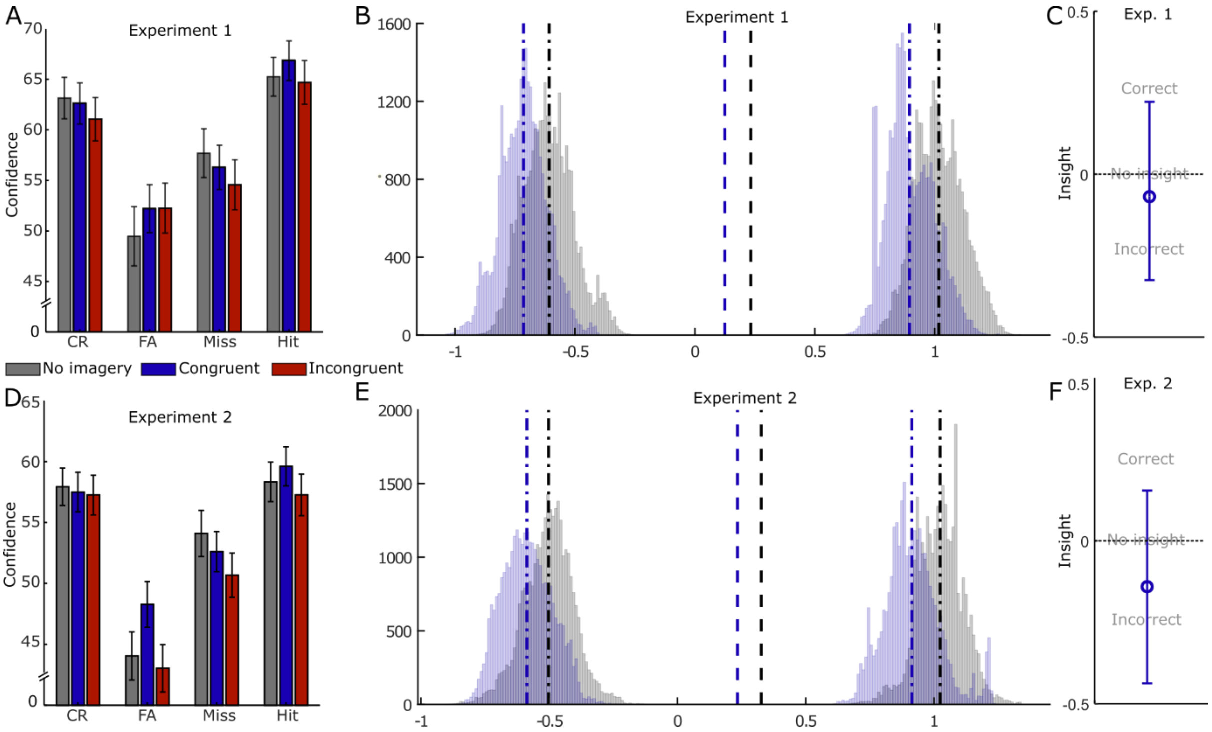

Figure 3. Metacognitive insight during source mixing. (**A**) Confidence ratings per condition and trial type for experiment 1 (CR = correct rejection; FA = false alarm). Only data for participants with all trial types in all conditions are shown (*N* = 59). A repeated-measures ANOVA indicated no effect of condition (*F*(104,2) = 1.039, *p* = 0.36; or of response (*F*(57,1) = 0.156, *p* = 0.694); nor an interaction between the two (*F*(104,2) = 2.821, *p* = 0.068). There was a significant effect of input (*F*(57,1) = 41.13, *p* < 0.001) as well as an interaction between input and response (*F*(104,2) = 112.254, *p* < 0.001), indicating that confidence was higher when a stimulus was presented and when the response was correct. (**B**) Posterior probability estimates of second-level criteria during no imagery and imagery for experiment 1. MCMC = Markov Chain Monte Carlo. (**C**) Estimate of empirical metacognitive insight for experiment 1 with associated HDI of MCMC samples. (**D**) Same as (A) for experiment 2. In contrast to experiment 1, there were significant effects of condition, see main text for more details. In line with experiment 1, there was also a significant effect of input (*F*(91,2) = 65.08, *p* = <0.001, η_p^2 = 0.417) and an interaction between input and response (*F*(91,1) = 162.31, *p* = <0.001, η_p^2 = 0.64) indicating that confidence was higher when a stimulus was presented and when the response was correct. In contrast to experiment 1, there was also a significant main effect of response (*F*(91,2) = 4.26, *p* = 0.042, η_p^2 = 0.045) showing that absence responses were more confident than presence ones. (**E-F**) same as (B-C) for experiment 2.

To replicate these results, we ran a second experiment that was mostly identical to the first but also included global insight questions at the end (see section below). After exclusion, data from 111 participants were analysed. We first again replicated the decision-level effect, showing a decrease in criterion for congruent imagery (*M* = 0.23, *SD* = 0.50) compared to no imagery (*M* = 0.35, *SD* = 0.53, *t*(110) = 2.48, *p* = 0.015, *d* = 0.22, 95% *CI* difference = 0.02 – 0.21), but not for incongruent imagery (*M* = 0.32, *SD* = 0.56, *t*(110) = 0.813, *p* = 0.42, *d* = 0.06, 95% *CI* difference = -0.04 – 0.10). In contrast, there was a significant decrease in d' for incongruent imagery (*M* = 1.56, *SD* = 1.25), compared to no imagery (*M* = 1.72, *SD* = 1, *t*(110) = 2.16, *p* = 0.033, *d* = -0.14, 95% *CI* difference = 0.01 – 0.32), but not for congruent imagery (*M* = 1.75, *SD* = 1.27, *t*(110) = -0.27, *p* = 0.79, *d* = 0.02, 95% *CI* difference = -0.2 – 0.15). These results again show that participants were more likely to indicate perceptual presence when imagining the same stimulus whereas imagining a different stimulus merely decreased performance.

We next ran a repeated-measures ANOVA on the confidence ratings of the participants that did not have any empty cells (that is, who committed both misses and false alarms; N = 92). In contrast to experiment 1, we found a significant effect of condition on confidence (*F*(182,2) = 6.64, *p* = 0.002, η_p^2 = 0.068). However, post-hoc tests revealed that this effect was driven by a decrease in confidence in the incongruent imagery condition (*M* = 52.1, *SD* = 15.2) compared to both no imagery (*M* = 53.5, *SD* = 15.5, *t*(91) = 2.28, *p* = 0.79, *d* = 0.02, 95% *CI* difference = -0.2 – 0.15) and congruent imagery (*M* = 1.75, *SD* = 1.27, *t*(110) = -0.27, *p* = 0.79, *d* = 0.02, 95% *CI* difference = -0.2 – 0.15). There was no significant difference between congruent imagery and no imagery in confidence ratings (*t*(91) = 1.322, *p* = 0.188, *d* = -0.138, 95% *CI* difference = -0.25 – 0.74). This decrease in confidence during the incongruent condition presumably reflected the decrease in performance in that condition.

Interestingly, there was also a significant interaction between condition and response (Huyn-Feldt corrected *F*(172.48,1.9) = 3.937, *p* = 0.023, η_p^2 = 0.041). Post-hoc analyses revealed that confidence of presence responses, not absence responses, was higher for congruent imagery (*M* = 53.94, *SD* = 17.2) compared to no imagery (*M* = 51.19, *SD* = 17.83, *t*(91) = 2.71, *p* = 0.008, *CI* = 0.73 – 4.78, *d* = 0.16) and incongruent imagery (*M* = 50.15, *SD* = 17.9, *t*(91) = 4.3, *p* = 0.0004, *CI* = 2.04 – 5.55, *d* = 0.22). We next directly quantified metacognitive insight using our extension of the hierarchical meta-d model. In line with the results from experiment 1, we again found evidence for a lack of insight, indicating that the metacognitive criteria moved along with the decision-level criterion (Fig. 3E&F, *M* = -0.14, 95% *HDI* = -0.44 – 0.15, BF_{01} = 62.89).

Post-experiment queries reveal global insight into perceptual reality monitoring errors

Taken together, our results suggest that people do not have insight into the influence mental imagery has on their perception at the level of individual trials. However, one intriguing possibility is that they do have insight at a more global level, but that this awareness is not incorporated into local within-task confidence ratings. Dissociations between local and global metacognition have previously been reported in relation to several dimensions of mental ill health (Bhome et al., 2022; Seow et al., 2021) as well as in normal aging (McWilliams et al., 2023).

To investigate this possibility in the context of our study, we included global insight questions at the end of both experiments. At the end of experiment 1, we asked the open-ended question “Do you think imagining the gratings influenced whether you saw a grating?”. Of the 51 participants who answered this question, 24 (47.1%) thought that imagery did have an influence, 19 (37.2%) were unsure and only 8 (15.7%) thought that imagery did not influence was they saw (Fig. 4A). Due to the qualitative nature of this question and the fact that only a small number of participants answered it, we were unable to link global insight to local insight in experiment 1.

To address this, after the second experiment, we interrogated global insight using a structured questionnaire. We asked participants to indicate how they thought imagining the gratings influenced their detection responses relative to not imagining by moving a slider ranging from 0 = ‘made me less likely to say I saw a grating’ to 100 = ‘made me more likely to say I saw a grating’ with the centre of the scale indicating 50 = ‘no effect’. This question was asked both in relation to imagining the same grating (congruent) or imagining a different grating (incongruent).

Global insight ratings indicated that, overall, participants accurately indicated that congruent imagery made it more likely for them to report presence compared to no imagery (*M* = 68.14, *SD* = 19.79, *t*(110) = 9.65, *p* < 0.0001, *CI* = 64.41 – 71.86, *d* = 0.92) whereas for the incongruent imagery the ratings were not significantly different from the centre, ‘no effect’ point (*M* = 46.63, *SD* = 25.92, *t*(110) = -1.37, p = 0.17, *CI* = 41.75 – 51.51, *d* = 0.13). This suggests that participants did have global insight into the effect of imagery on perceptual presence responses. However, there was no significant correlation between the extent to which participants believed imagery increased their presence responses and how much it actually did, measured by the decision-level criterion shift (r = 0.12, p = 0.2). Furthermore, there were no significant correlations between global insight and local confidence ratings in any of the conditions (all absolute r's < 0.15, all p's > 0.11). These results suggest that the insight judgement at the end of the experiment might have relied on different cognitive mechanisms than the decisions and confidence ratings elicited during the perceptual detection task.

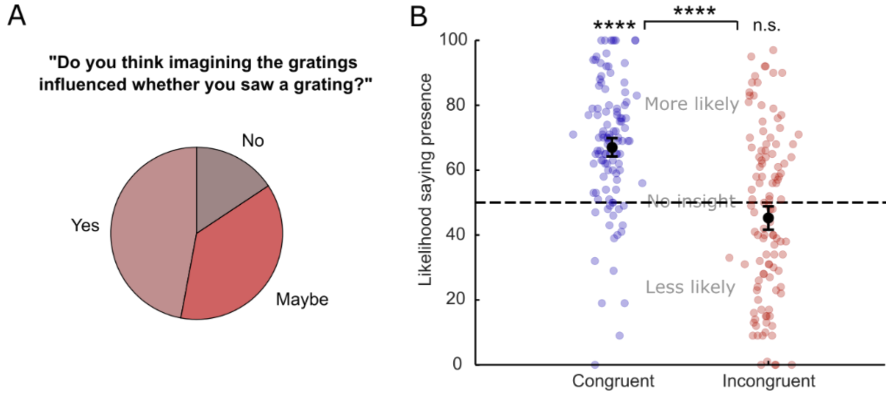

Figure 4. Post-experiment queries indicate insight (**A**) Proportion of participants indicating that they thought imagery influenced their perceptual response after experiment 1. N = 51. (**B**) Responses to the question whether imagery made participants more or less likely to indicate perceptual presence, separate for congruent (blue) and incongruent (red) conditions, after experiment 2. Dots represent individual participants. **** p < 0.0001; n.s. = non-significant.

Discussion

In this study we set out to investigate insight into perceptual reality monitoring errors. To this end, we asked participants to indicate confidence in a perceptual detection task while they simultaneously also imagined the stimuli they had to detect. In line with previous studies, we first observed that participants more often reported seeing a stimulus during congruent imagery, indicating that imagery was sometimes mistaken for perception. We extended a hierarchical Bayesian model of confidence to characterize metacognitive insight into these perceptual reality monitoring errors. Over two experiments, we showed that confidence criteria moved in tandem with the decision criterion shift, indicating a lack of awareness of the effect of imagery on perceptual presence responses. In other words, participants were equally confident in veridical and imagery-induced presence responses. In experiment 2, there was even an *increase* in confidence for presence responses during congruent imagery. Interestingly, however, offline queries indicated some level of global insight into the influence of imagery on perceptual detection, but this insight was unrelated to perceptual decisions and confidence ratings. Together, our results demonstrate a lack of local insight into mistaking imagination for reality.

The observation that congruent imagery increases both perceptual presence responses and confidence in those responses is in line with the idea that imagery can function as perceptual evidence (Dijkstra, Kok, et al., 2021; Dijkstra, Mazor, et al., 2021; Pearson et al., 2008). Specifically, while an imagery-induced increase in 'presence' responses can be equally explained by a stronger perceptual signal or a shift in the decision criterion, the corresponding shifts in confidence criteria break this indeterminacy in favour of a perceptual account (Gallagher et al., 2019). This idea is further supported by neuroimaging studies showing that imagery is associated with perception-like neural representations throughout the visual cortex (Albers et al., 2013; Dijkstra et al., 2019; Pearson, 2019; Ragni et al., 2020). Sensory activation during imagery tends to be much weaker than during externally triggered perception, which might be why we generally do not mistake our imagination for reality (Dijkstra & Fleming, 2023; Koenig-Robert & Pearson, 2021). However, our results indicate that in ambiguous contexts such as threshold perceptual detection, imagery-induced sensory signals might be erroneously judged as real.

Furthermore, in experiment 2, we found that congruent imagery was associated with an increase in high confidence false alarms. High confidence false alarms have recently been proposed as a proxy for studying hallucinations in non-human animals, with studies in mice linking these percepts to elevated striatal dopamine (Schmack et al., 2021). In that context, hallucinations were induced by either manipulating perceptual expectation or reward expectation. To what extent imagery-induced hallucinations rely on similar mechanisms is unclear. One possibility is that imagery functions as a perceptual expectation, increasing the prior for imagined content. However, contrary to the usual (Bayesian) function of expectation, imagery tends to be used to generate perceptual information about stimuli that we know are not present in the environment (Kosslyn et al., 2001). Future research is necessary to investigate how these different types of hallucinations are related.

Our findings demonstrate the multi-dimensional nature of insight. Despite the fact that perceptual reality judgements and confidence ratings in those judgements indicated an absence of insight, participants were nevertheless able to accurately indicate that congruent imagery would make them more likely to report perceptual presence in response to post-experiment questioning, suggesting some more global form of insight. However, these offline responses were unrelated to online perceptual reality judgements or confidence ratings, suggesting that these two forms of insight are driven by different factors. Offline insight judgements might for example rely more on abstract knowledge rather than direct perceptual experience during the task. This dissociation might in turn be related to a distinction between perception of reality and beliefs about reality (Dijkstra et al., 2022).

One recent study aimed to directly dissociate perceptual reality and beliefs about reality in the context of a visual illusion (Mihali et al., 2022). Prior to the experiment, participants were informed about how the illusion worked and then, in separate blocks, had to indicate their perceptual experiences ('what do you see') and their beliefs ('what do you believe is presented on the screen') while rating confidence in both. The results indicated that confidence ratings always tracked the first-order decision but that, in contrast to perceptual judgements, the belief judgements reflected insight into the visual illusion (Mihali et al., 2022). In our study, we asked participants 'was there a grating on the screen?' which could refer either to their belief or to their perception. Future research is necessary to further disentangle these two levels of reality monitoring during simultaneous imagery and perception.

Similar dissociations between different levels of insight have been documented in relation to functional cognitive disorders – conditions in which patients believe that their cognitive functioning has declined despite intact performance on cognitive tasks (Pennington et al., 2015). A recent study of patients with functional memory disorder showed that despite a global belief of low memory performance, online confidence ratings during a memory task demonstrated intact metacognitive sensitivity (Bhome et al., 2019). The authors hypothesized that this dissociation might be due to a disconnection syndrome in which global priors are unable to influence local confidence ratings. A similar mechanism might be at play in the current context, where the drivers of offline insight are unable to influence local perceptual reality judgements on individual trials.

In conclusion, by extending a hierarchical Bayesian model of metacognition to characterize confidence judgements during a simultaneous imagery and perceptual detection task, we reveal an absence of local insight into mistaking imagination for reality. However, at a global level, participants were able to indicate how imagery influenced their perceptual responses at the end of the experiment. Future research is necessary to investigate these different levels of insight and how they relate to disorders of perceptual reality monitoring.

Materials and Methods

This study was pre-registered at https://osf.io/n39tm/. We initially planned for single-subject analyses and after first data collection, we realized that there were a high number of participants that did not have any false alarms or misses in one of the conditions. Therefore, we next tried to reduce this number by staircasing the decision-level criterion in a second data collection (updated pre-registration). However, after further consideration, we eventually adopted a hierarchical model which allowed us to relax these criteria and include all participants. Moreover, because the staircasing procedure was orthogonal to our main question, we decided to combine the two data-sets for the final analyses to achieve maximum statistical power. To ensure the validity of our results we replicated our findings in an independent sample in experiment 2.

Participants experiment 1. 130 participants (half with the first staircasing procedure and half with the second) were recruited using Prolific (www.prolific.co) and completed the study online. Informed consent was obtained from each participant included in the study. The experiment took approximately 50 mins to complete and participants were paid £7.50 (£9 hourly rate, which is more than the pre-registered hourly rate of £7.50 due to an update of the default within Prolific). All procedures were approved by the University College London ethics committee. Data from 8 participants was not obtained due to technical issues. Participants were furthermore excluded if (1) their mean detection accuracy over conditions was below chance; (2) they answered the imagery check (see below) correctly on less than 2 blocks in any of the conditions; (3) they indicated in the debrief questions that they did not imagine the gratings as instructed; (4) if, for a given detection response (yes/no) they used the exact same confidence rating in more than 90% of the trials. 6 participants were moved due to below 55% detection accuracy, 13 due to too few correct imagery blocks, 1 due to indicating they hadn't imagined as instructed and 0 due to too little variance in confidence ratings. In total, 102 participants (mean age = 30.4, *SD* = 8.5) were included in the final analyses.

Experimental design and procedure experiment 1. To explain the concept of mental imagery to participants in a systematized way, they started the study by filling out a selection of the Vividness of Visual Imagery Questionnaire (VVIQ2; (Marks, 1995). We chose to focus only on the 'shop' part of the questions to save time and because this part of the questionnaire has been shown to lead to the highest vividness scores in the general population (Aphantasia Network; https://aphantasia.com). This part instructed the participant to "Think of a shop which you often go to. Consider the picture that comes before your mind's eye." And then answer questions like "How vivid is the overall appearance of the shop from the opposite side of the road?" Which participants have to rate from "No image at all, you only "know" that you are thinking of the object" to "Perfectly clear and as vivid as real seeing" on a scale from 1 to 5. After the VVIQ2, participants practiced detecting gratings in dynamic noise (Fig. 1) for 6 trials or until they had at least 75% accuracy.

After this, the staircasing procedure started. For the first data collection, we staircased the visibility of the grating only based on performance – aiming for a 70% detection accuracy. We staircased only one orientation to save time and because previous experiments showed that the threshold visibility values. The staircase contained 120 trials and accuracy was calculated after every 10 trials. Visibility was increased if accuracy was below 65 and decreased if it was above 75. After this, the concept of confidence was explained to participants and they practice indicating confidence in their decision. Finally, participants practiced imagining the gratings in noise and rated their imagery vividness afterwards on a scale from 1 to 5 for 10 trials per orientation before continuing to the main task.

For the second data collection, we additionally aimed to staircase the criterion so that we would obtain sufficient number of misses and false alarms. Furthermore, we now included the confidence ratings already within the staircase. Criterion was staircased by presenting prompts: if there were no misses in a mini-block of 16 trials, we presented "Remember that sometimes noise might look like a grating and gratings are present on 50% of the trials", conversely, if there were no false alarms, we presented "Remember that the gratings are hard to detect and there is a grating present on 50% of the trials". This was repeated until the accuracy was between 50 and 100 and participants had at least one hit and one miss, or after 5 blocks of 16 staircase trials had passed, whichever came first. Finally, participants practiced imagining the gratings in noise.

The main experimental design is shown in Figure 1. Participants used the "F" and "G" keys with their left hand to perform the detection task and the mouse with their right hand to indicate their confidence on a continuous scale. Response-key mappings for the detection task - which key, "F" or "G", corresponds to presence and which key to absence – were randomized over participants. To ensure participants properly indicated their confidence, they had to move the confidence slider on each trial, at least a small amount, before the experiment continued. To ensure that participants accurately followed the imagination instructions, after each block we asked participants what they were instructed to imagine. We only analysed data of blocks that were answered correctly and of participants that answered correctly in most blocks (see above). At the end of the experiment we asked the following open-ended questions: (1) the participant's age; (2) if they imagined the gratings as instructed; (3) if they thought that imagining the grating influenced whether they saw one and if yes, how (only for the second data collection); and (4) any if they had any further comments.

Participants experiment 2. We performed a power calculation based on the results from experiment 1 to determine the number of participants for experiment 2. Assuming an effect size of 0.32, 103 participants would be required to reach a power of 90% with a two-sided alpha level of 0.05. Taking into account drop-out, 130 participants gave informed consent and completed the study online. Data recruitment, collection and exclusion criteria were identical to the first experiment. Data from 3 participants were not collected due to technical issues, 8 participants were removed due to below 55% detection accuracy, 6 due to too few correct imagery blocks, 0 due to indicating they hadn't imagined as instructed and 2 due to too little variance in confidence ratings. In total, 111 participants (mean age = 33.16, *SD* = 9.89) were included in the final analyses.

Experimental design and procedure experiment 2. The design and experiment were identical to both data collections of experiment 1 except that the staircasing procedure was now a mix between the two. Specifically, in experiment 2 the confidence rating was included in the staircasing procedure and the visibility was only staircased based on accuracy. The visibility first quickly went down to a level that corresponded to about threshold performance in the previous sample and was then fine-tuned depending on participant's response: going up if accuracy was below 60% and going down if it was above 80%. At the end of the main experiment, a global insight question was asked as follows:

"On some blocks you imagined [the same]/[a different] grating as the one you had to detect. Relative to the blocks where you didn't have to imagine anything, how did this affect your tendency to say there was a grating on the screen?"

Relative to not imagining anything, imagining [the same]/[a different] grating…"

The answer was indicated using a slider that ranged from: "Made me much less like to report seeing a grating" to "Made me much more likely to report seeing a grating" with "Had no effect" in the middle.

Data analyses. We analysed the first-order decision responses using standard signal detection theory (Green & Swets, 1966). Detection sensitivity (d') and criterion (c) were calculated separately for the imagery and no-imagery trials as follows:

$$d' = z(H) - z(FA)$$

$$c = \ -0.5 \ \times [z(H) + z(FA)]$$

where z indicates the inverse of the cumulative normal distribution, H is the hit rate (the proportion of present trials for which the participant reported presence), and FA is the false alarm rate (the proportion of absent trials for which the participant reported presence). Detection sensitivity d' is a measure of detection performance, with greater values indicating better performance. Criterion c is a measure of participant's bias towards responding 'yes' (present) or 'no' (absent), irrespective of whether a stimulus is present or not. Greater values of c indicate a more conservative criterion, indicating a greater tendency towards reporting absence. Hit rates of 1 or false alarm rates of 0 lead to biased estimations of d' and c. To correct for this, in those cases of extreme values we added a count of 0.5 to the relevant cell (Hautus, 1995).

Confidence ratings were first analysed with a repeated-measures ANOVA with input (present vs absent) × response ('absent' vs 'present') × condition (no imagery vs congruent vs incongruent) as within-subject variables. Only participants with at least one observation in each cell were included for this analysis (N = 59). Next, we extended the hierarchical meta-d' model (Fleming, 2017) to allow inference on metacognitive insight into perceptual reality monitoring errors. Because the generative multinomial model used by HMeta-d naturally handles zero cell counts, and the hierarchical structure pools data over participants, we were able to use all participants for this analysis. Furthermore, in contrast to standard frequentist approaches, the Bayesian framework allowed us to quantify evidence in favour of the *absence* of metacognitive insight. Prior to fitting the model, we split the confidence ratings of each subject and each condition into low and high by first z-scoring the data and counting all ratings above 0 as high and below 0 as low.

Within the original HMeta-d model, the group level parameter of interest is M-ratio M, the ratio between second-order $meta - d$' and first-order decision d':

$$M = \frac{meta - d'}{d'}$$

where $meta - d'$ is calculated by estimating what the first-order d' would be based on the observed confidence ratings alone and d' is a fixed parameter directly calculated from the decision data (Fleming, 2017; Maniscalco & Lau, 2012). An M below 1 indicates that there is information loss when going from decision to confidence rating whereas an M above 1 indicates that additional information is incorporated at the confidence rating stage.

Here, in contrast, we were interested in metacognitive insight into changes in decision criterion *c* rather than performance *d'*. Within the original HMeta-d model, the absolute distances between the decision criterion and the negative and positive confidence criteria are modelled as being symmetric (the priors for both positive and negative criteria are shared). This means that a shift in the decision criterion leads to an accompanying (symmetric) shift in the confidence criteria under the default priors (cf. Fig. 2A vs Fig. 2B). Here, in contrast, in order to allow for independent shifts in insight, we included separate task-specific group-level priors over the negative $c2$- and positive $c2$+ confidence criteria (μ_{c2-t}, σ_{c2-t} and μ_{c2+t}, σ_{c2+t} in Fig. 5). Furthermore, we modelled second-order criteria relative to the first-order criterion to allow for a direct comparison between the positive and negative $c2$ (Fig. 5). Asymmetries between positive and negative $c2$ distances would indicate some form of metacognitive insight (cf. Fig. 2).

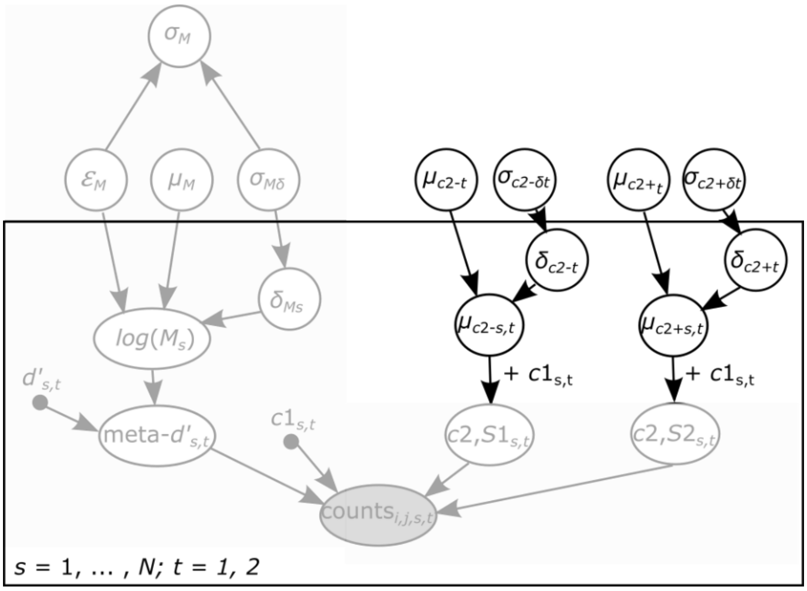

Figure 5. Current extension of the HMeta-d model. Probabilistic graphical model for estimating metacognitive insight into criterion c shifts. A full description of the original HMeta-d model, here indicated in light grey, can be found in Fleming (2017). We extended the model to include group-level priors (mean and variance) over the difference between decision $c1$ and confidence $c2$ criteria, separately for negative (absence) $c2$- (μ_{c2-t} and σ_{c2-t}) and positive (presence) $c2$+ (μ_{c2+t} and σ_{c2+t}) responses. Point estimates for type 1 d' and criterion are represented as black dots. The box encloses participant-level parameters subscripted with *s* whereas parameters outside the box represent group-level parameters. Task specific parameters are subscripted with t. We employ the scheme suggested by (Matzke et al., 2014), such that the mean and variance of $log(Ms)$ are scaled by a redundant multiplicative parameter ξM. The posterior on σM can then be recovered by adjusting for the influence of this additional random component

To perform statistical inference on these estimates, we used the Savage-Dickey density ratio to calculate Bayes Factors (Wagenmakers et al., 2010). The Savage-Dickey density ratio is defined as:

$$BF_{01} = \frac{p(D|H0)}{p(D|H1)} = \frac{p(\theta = 0|D, H1)}{p(\theta = 0|H1)}$$

Which states that the ratio of the probability of the null hypothesis of no difference $H0{:}\ \theta = 0$ versus the alternative hypothesis of there being a difference $H1{:}\ \theta \neq 0$ is given by dividing the height of the posterior at $\theta = 0$ by the height of the prior at $\theta = 0$. To this end, we created a dummy variable in the model which directly sampled from the prior. We next calculated the difference between positive and negative $c2$ distances – our measure of insight - for both prior and posterior samples and obtained a Bayes Factor by comparing the two.

Code and data availability

All code and data are published at https://github.com/NadineDijkstra/METPRM

References

Albers, A. M., Kok, P., Toni, I., Dijkerman, H. C., & De Lange, F. P. (2013). Shared representations for working memory and mental imagery in early visual cortex. *Current Biology : CB*, *23*(15), 1427–1431. https://doi.org/10.1016/j.cub.2013.05.065

Baird, B., Mota-Rolim, S. A., & Dresler, M. (2019). The cognitive neuroscience of lucid dreaming. *Neuroscience and Biobehavioral Reviews*, *100*, 305–323. https://doi.org/10.1016/j.neubiorev.2019.03.008

Bhome, R., McWilliams, A., Huntley, J. D., Fleming, S. M., & Howard, R. J. (2019). Metacognition in functional cognitive disorder- a potential mechanism and treatment target. *Cognitive Neuropsychiatry*, *24*(5), 311–321. https://doi.org/10.1080/13546805.2019.1651708

Bhome, R., McWilliams, A., Price, G., Poole, N. A., Howard, R. J., Fleming, S. M., & Huntley, J. D. (2022). Metacognition in functional cognitive disorder. *Brain Communications*, *4*(2). https://doi.org/10.1093/braincomms/fcac041

Boundy-Singer, Z. M., Ziemba, C. M., & Goris, R. L. T. (2022). Confidence reflects a noisy decision reliability estimate. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-022-01464-x

Corlett, P. R., Canavan, S. V., Nahum, L., Appah, F., & Morgan, P. T. (2014). Dreams, reality and memory: Confabulations in lucid dreamers implicate reality-monitoring dysfunction in dream consciousness. *Cognitive Neuropsychiatry*, *19*(6), 540–553. https://doi.org/10.1080/13546805.2014.932685

Dijkstra, N., Bosch, S. E., & van Gerven, M. A. J. (2019). Shared Neural Mechanisms of Visual Perception and Imagery. *Trends in Cognitive Sciences*, *23*, 18–29. https://doi.org/10.1016/j.tics.2019.02.004

Dijkstra, N., & Fleming, S. M. (2023). Subjective signal strength distinguishes reality from imagination. *Nature Communications*, *14*, 1627. https://doi.org/10.1038/s41467-023-37322-1

Dijkstra, N., Kok, P., & Fleming, S. (2021). Imagery adds stimulus-specific sensory evidence to perceptual detection. *PsyArXiv*. https://doi.org/10.31234/OSF.IO/76AD2

Dijkstra, N., Kok, P., & Fleming, S. M. (2022). Perceptual reality monitoring: Neural mechanisms dissociating imagination from reality. *Neuroscience and Biobehavioral Reviews*, *135*, 104557. https://doi.org/10.1016/j.neubiorev.2022.104557

Dijkstra, N., Mazor, M., Kok, P., & Fleming, S. (2021). Mistaking imagination for reality: Congruent mental imagery leads to more liberal perceptual detection. *Cognition*, *212*, 104719. https://doi.org/10.1016/j.cognition.2021.104719

Fleming, S. M. (2017). HMeta-d: Hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, *2017*(1). https://doi.org/10.1093/nc/nix007

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114. https://doi.org/10.1037/rev0000045

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*(JULY), 443. https://doi.org/10.3389/fnhum.2014.00443

Gallagher, R. M., Suddendorf, T., & Arnold, D. H. (2019). Confidence as a diagnostic tool for perceptual aftereffects. *Scientific Reports*, *9*(1), 1–12. https://doi.org/10.1038/s41598-019-43170-1

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. In *John Wiley* (Vol. 5). https://doi.org/10.1016/0022-460x(67)90197-6

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d′. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51. https://doi.org/10.3758/BF03203619

Koenig-Robert, R., & Pearson, J. (2021). Why do imagery and perception look and feel so different? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1817), 20190703. https://doi.org/10.1098/rstb.2019.0703

Konkoly, K., Appel, K., Chabani, E., Mironov, A. Y., Mangiaruga, A., Gott, J., Mallett, R., Caughran, B., Witkowski, S., Whitmore, N., Berent, J., Weber, F., Pipa, G., Türker, B., Maranci, J.-B., Sinin, A., Dorokhov, V., Arnulf, I., Oudiette, D., … Paller, K. (2021). Real-Time Dialogue between Experimenters and Dreamers During rem Sleep. *Current Biology*, *31*, 1–11. https://doi.org/10.2139/ssrn.3606772

Kosslyn, S. M., Ganis, G., & Thompson, W. L. (2001). Neural Foundations of Imagery. *Nature Reviews Neuroscience*, *2*(9), 635–642. https://doi.org/10.1038/35090055

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course.* Cambridge university press.

Mamassian, P., & de Gardelle, V. (2022). Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychological Review*, *129*, 976–998. https://doi.org/10.1037/rev0000312

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430. https://doi.org/10.1016/J.CONCOG.2011.09.021

Marks, D. F. (1995). New directions for mental imagery research. *Journal of Mental Imagery*, *19*(3–4), 153–167. https://doi.org/10.1177/0963721414532287

McWilliams, A., Bibby, H., Steinbeis, N., David, A. S., & Fleming, S. M. (2023). Age-related decreases in global metacognition are independent of local metacognition and task performance. *Cognition*, *235*. https://doi.org/10.1016/j.cognition.2023.105389

Menon, G. J., Rahman, I., Menon, S. J., & Dutton, G. N. (2003). Complex visual hallucinations in the visually impaired: The Charles Bonnet Syndrome. *Survey of Ophthalmology*, *48*(1). https://doi.org/10.1016/S0039-6257(02)00414-9

Mihali, A., Broeker, M., Ragalmuto, F., & Horga, G. (2022). Introspective inference counteracts perceptual distortion. *BioRxiv*, 2021.11.13.468497.

Pearson, J. (2019). The human imagination: The cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience*. https://doi.org/10.1038/s41583-019-0202-9

Pearson, J., Clifford, C. W. G., & Tong, F. (2008). The functional impact of mental imagery on conscious perception. *Current Biology : CB*, *18*(13), 982–986. https://doi.org/10.1016/j.cub.2008.05.048

Ragni, F., Tucciarelli, R., Andersson, P., & Lingnau, A. (2020). Decoding stimulus identity in occipital, parietal and inferotemporal cortices during visual mental imagery. *Cortex*, *127*, 371–387. https://doi.org/10.1016/j.cortex.2020.02.020

Schmack, K., Bosc, M., Ott, T., Sturgill, J. F., & Kepecs, A. (2021). Striatal dopamine mediates hallucination-like perception in mice. *Science*, *372*(6537). https://doi.org/10.1126/SCIENCE.ABF4740

Scott, R. B., Dienes, Z., Barrett, A. B., Bor, D., & Seth, A. K. (2014). Blind Insight: Metacognitive Discrimination Despite Chance Task Performance. *Psychological Science*, *25*(12), 2199–2208. https://doi.org/10.1177/0956797614553944

Seow, T. X. F., Rouault, M., Gillan, C. M., & Fleming, S. M. (2021). How Local and Global Metacognition Shape Mental Health. *Biological Psychiatry*, *90*(7). https://doi.org/10.1016/j.biopsych.2021.05.013

Siclari, F., Baird, B., Perogamvros, L., Bernardi, G., LaRocque, J. J., Riedner, B., Boly, M., Postle, B. R., & Tononi, G. (2017). The neural correlates of dreaming. *Nature Neuroscience*, *20*(6), 872–878. https://doi.org/10.1038/nn.4545

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189. https://doi.org/10.1016/J.COGPSYCH.2009.12.001

Zmigrod, L., Garrison, J. R., Carr, J., & Simons, J. S. (2016). The neural mechanisms of hallucinations: A quantitative meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, *69*, 113–123. https://doi.org/10.1016/j.neubiorev.2016.05.037

Zylberberg, A., Wolpert, D. M., & Shadlen, M. N. (2018). Counterfactual Reasoning Underlies the Learning of Priors in Decision Making. *Neuron*, *99*(5), 1083-1097.e6. https://doi.org/10.1016/j.neuron.2018.07.035