Title: The neural mechanisms underlying confidence formation. Commentary on *Sherman, Seth, and Kanai (2016)*.

Corresponding author: Christina A. Van Heer.

Corresponding author contact information:

Melbourne School of Psychological Sciences

12th Floor, Redmond Barry Building

The University of Melbourne

Parkville, VIC 3010

Phone: +61 431 574 407

Email: christinavanheer@gmail.com

Acknowledgements: I thank David Sewell, Robert Hester, Philip Smith, Simon Lilburn, Elaine Corbett, Daniel Bennett, Bianca Levis, Gabriel Ong and Maggie Webb for helpful discussions, and/or comments on an earlier version of this manuscript.

1 An impressive feat of cognition is the brain's ability to maintain seamless perception, while
2 continually monitoring the observed state of the world. The brain capitalises on expectations
3 about perception to maintain energy efficient computation, coding the extent to which
4 incoming sensory information deviates from expectations as a prediction error (PE) signal
5 (Summerfield and de Lange, 2014). To an observer, unexpected events may also be
6 accompanied by changes metacognitive signals such as confidence (Meyniel et al., 2015).
7 Prior information (expectation) supporting a choice has been shown to increase confidence,
8 and indeed confidence closely tracks PE (Nassar et al., 2010; Meyniel et al., 2015). This
9 suggests that PE and confidence may represent the same underlying computation (Meyniel et
10 al., 2015a). However, the brain structures which are involved in representing prior and
11 observed information, and constructing a representation of confidence are unknown. This was
12 investigated in a recent functional magnetic resonance imaging (fMRI) study of 19 healthy
13 observers published in *The Journal of Neuroscience* by Sherman, Seth and Kanai (2016).

14
15 The authors developed a task that manipulated expectations about a perceptual event so that
16 observers experienced perception which was expected or unexpected. Observers detected the
17 presence or absence of a peripherally-displayed gabor patch, while performing a centrally
18 displayed visual search task - reporting target, 'T' presence or absence, where the target was
19 embedded in an array of 4 equidistant distractors ('L' letters overlaid by 'F'). To investigate
20 how prior information influenced confidence, observers were told before each block the
21 probability of a gabor patch appearing (25%, 50% or 75%), and this also predicted the event
22 (e.g. 25%, indicated a gabor would appear on 25% of trials). This information was displayed
23 with instructions to either 1) perform both gabor detection and visual search, task, or 2) only
24 gabor detection (ignoring the visual search task) to control attention. Observers reported their
25 confidence in their gabor detection judgement as either 'confident' or 'guess'. As the authors

were interested in the neural correlates of PE's, they defined two types of behavioral PE events for subsequent fMRI contrasts - 1) *congruent trials*: where expectations agree with the outcome (25%, gabor absent or 75%, gabor present) and 2) *incongruent trials*: where expectations differ from the outcome (25%, gabor present and 75%, gabor absent), and producing a larger PE. They then searched for brain areas where the Blood Oxygen Level Dependent (BOLD) response was higher for the incongruent condition than the congruent condition, and areas of significant activation were referred to as a PEreport region.

The authors provide considerable evidence towards rIFG being involved in confidence formation; rIFG activity was higher incongruent trials than congruent trials, and also for guess responses than confident trials. This supports the convergence of expectation and report information in rIFG. Secondly, they were able to link the PEreport signal in rIFG to the degree to which overall confidence is robust to the influence of observed information, by calculating the difference in the proportion of 'confident' judgements between the congruent condition and the incongruent condition, which they termed Δ Confidence. They found individuals whose overall confidence increased from the incongruent and congruent condition (high Δ Confidence score) had a *lower* PEreport BOLD. This indicates that high confidence states are linked to low PEreport in rIFG.

Sherman et al. (2016) were also able to neatly map variability in both their neural (rIFG signal) and behavioral (Δ Confidence and change in decision criteria) data to variability in signals in orbitofrontal cortex (OFC) and intracalcarine sulcus (ICS), using both psychophysiological interaction (PPI) and white matter (WM) density analyses. They show that the strength of predictive signals in OFC increased with both congruence and the predictability of prior information (25% and 75% condition), and also for observers whose

behavior was less influenced by observed information (low Δ Confidence score). Conversely, observers whose confidence was more influenced by observed information (high Δ Confidence score) showed marginal connectivity with ICS, suggesting that this behavior may be underpinned by a dominance of sensory signals from early visual areas.

While the authors have made an impressive attempt to account for individual variability in both the behavioral and neural data, there of course remains some notable caveats regarding the proposed mechanisms underlying confidence formation, and furthermore, the ability to infer these from neural data alone. Although activity in rIFG appears to encode expectation and observed information, and the extent to which confidence tracks these relative contributions, this does not preclude the possibility that rIFG activation reflects something other than confidence formation. Any cognitive process sensitive to an expectation deviation (the fMRI contrast used) may underlie this activity. One such candidate is response inhibition, which has been broadly associated with the recruitment of rIFG (Forstmann et al., 2008; Mulder et al., 2012; Aron et al., 2014). The authors argue against this account because they do not experimentally manipulate this, for example, by asking observers to withhold their response on specific trials (e.g. a go/no-go task). However, a direct manipulation is not necessary to induce response inhibition, given evidence for rIFG recruitment in “braking” responses, where observers slow their response, anticipating stopping, but do not completely stop. This account predicts observers whose confidence is more influenced by prior information should require minimal response inhibition for the congruent condition, and this may be reflected in lower rIFG activity. Interestingly, this is the pattern of results observed in the data. Alternatively, a PE signal in rIFG may be a ‘surprise’ signal, coding the degree to which a percept is unexpected. This may occur in rIFG or be transmitted from another region, which then results in response inhibition (Chikazoe et al., 2008). Another possibility is that

confidence is constructed in rIFG, and the metacognitive component derived from confidence signals a need for response inhibition in incongruent trials, and this results in lower rIFG activity. Additionally, while the data suggests OFC (and marginally, ICS) are involved in constructing the PEreport signal, it does not account for why other studies have reported signatures of confidence/uncertainty in OFC (De Martino et al., 2012; Lebreton et al., 2015). These regions may also be involved in representing confidence, and could suggest that confidence has an initial representation in rIFG and is transmitted to regions such as OFC, where it is combined with prior information. Alternatively, confidence may be encoded rIFG as a representation of uncertainty but is ‘read-out’ (Meyniel et al., 2015b), for report (i.e. response) in other prefrontal regions (Fleming and Dolan, 2014; Meyniel et al., 2015a). However, these accounts may not be readily distinguishable using fMRI because of its poor temporal resolution.

Moving forward, there are three main ways in which neural mechanisms proposed could be refined. Firstly, how are confidence and the other cognitive processes such as response inhibition, distinct from each other? There appears to be a degree of overlap between the neural structures underlying both confidence/metacognition and cognitive control (Fleming and Dolan, 2014), and the extent to which these processes can be disassociated may resolve this. Secondly, arbitrating between competing mechanistic accounts may be enhanced by undertaking a model based approach to analysing neural data, by uncovering latent processes that underpin behavior, which could then be related to neural data (Forstmann et al., 2011). In particular, Bayesian models describe how an observer integrates prior information about gabor expectations, with observed information, described by a likelihood function, to produce a posterior distribution of probability about particular outcomes. Formally, confidence in a Bayesian model can be described by the variance (width) of the distribution of beliefs.

Intuitively, the larger the distribution of beliefs, the lower confidence (Meyniel et al., 2015b). This model may generate readily testable predictions; the width (precision) of a posterior distribution should be wider for guesses, than confident responses, and be modulated by expectation (Nassar et al., 2010; Meyniel et al., 2015b; Sherman et al., 2016). These parameters could then be related to neural activity to find structures that might encode this information. Finally, while Sherman and colleagues have made progress in isolating the neural structures which support the emergence of confidence signal, we may need to move beyond attempts at localising processes to specific brain areas. Neural representations of confidence evolve over time, and in multiple brain areas (Meyniel et al., 2015b). By looking at the temporal evolution of confidence using high temporal resolution techniques like magnetoencephalography (MEG) and electroencephalography (EEG), we might be able to refine the directionality of neural processes which cannot be resolved in the present study, and look at how confidence changes representations of sensory information at different stages of processing, which could be combined with model based approaches.

In summary, the study by Sherman et al., (2016) presents a promising attempt to characterise how confidence is formed from prior and sensory information in the brain, suggesting that these signals converge in rIFG as a PE. While their data suggest encoding of neural processes which are required for the formation of confidence, the mechanisms underlying this are still unclear, and should be clarified by future research aiming to arbitrate between competing mechanisms underlying the formation of decision confidence.

References

- Aron A, Robbins T, Poldrack R (2014) Inhibition and the right inferior frontal cortex: one decade on. *Trends Cogn Sci* 18:177–185.
- Chikazoe, Jimura, Asari, Yamashita -i, Morimoto, Hirose, Miyashita, Konishi (2008) Functional Dissociation in Right Inferior Frontal Cortex during Performance of Go/No-Go Task. *Cereb Cortex*:146–152.
- Fleming S, Dolan R (2014) The Cognitive Neuroscience of Metacognition. :245–265.
- Forstmann B, Jahfari S, Scholte H, Wolfensteller U, Wildenberg W, Ridderinkhof K (2008) Function and Structure of the Right Inferior Frontal Cortex Predict Individual Differences in Response Inhibition: A Model-Based Approach. *J Neurosci* 28:9790–9796.
- Forstmann B, Wagenmakers E-J, Eichele T, Brown S, Serences J (2011) Reciprocal relations between cognitive neuroscience and formal cognitive models: opposites attract? *Trends Cogn Sci* 15:272–279.
- Lebreton M, Abitbol R, Daunizeau J, Pessiglione M (2015) Automatic integration of confidence in the brain valuation signal. *Nat Neurosci* 18:1159–1167.
- De Martino B, Fleming S, Garrett N, Dolan R (2012) Confidence in value-based choice. *Nat Neurosci* 16:105–110.
- Meyniel F, Schlunegger D, Dehaene S (2015a) The Sense of Confidence during Probabilistic Learning: A Normative Account. *PLoS Comput Bio* 11:e1004305.
- Meyniel F, Sigman M, Mainen Z (2015b) Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron* 88:78–92.
- Mulder M, Wagenmakers E-J, Ratcliff R, Boekel W, Forstmann B (2012) Bias in the Brain: A Diffusion Model Analysis of Prior Probability and Potential Payoff. *J Neurosci* :2335–2343.
- Nassar M, Wilson R, Heasly B, Gold J (2010) An Approximately Bayesian Delta-Rule

Model Explains the Dynamics of Belief Updating in a Changing Environment. *J Neurosci* 30:12366–12378.

Summerfield C, de Lange F (2014) Expectation in perceptual decision making: neural and computational mechanisms. *Nat Rev Neurosci*:816–816.