Published in final edited form as:

Cogn Neuropsychiatry. 2019 May ; 24(3): 165–177. doi:10.1080/13546805.2019.1606706.

Factor One, Familiarity and Frontal Cortex: A challenge to the two-factor theory of delusions

Philip R. Corlett^{1,*}

¹Department of Psychiatry, Yale University, School of Medicine. New Haven, CT, USA

Abstract

Introduction: Two-factor theory suggests delusions require two neuropsychological impairments, one in perception (which furnishes content), and a second in belief evaluation (that augurs formation and maintenance). Capgras delusion; the belief that one's loved one has been replaced by an imposter, then entails two independent processes; first a lack of skin conductance response to familiar faces so the loved one feels different. This has been demonstrated in four patients with damage to ventromedial prefrontal cortex (vmPFC) but who do not have delusions. Thus two-factor theorists demand a second factor: a change in belief evaluation, which is associated with damage to the right dorsolateral prefrontal cortex (rDLPFC).

Methods: Literature review of foundational and related papers on the cognitive neuropsychology of delusions, perception and belief.

Results: The four vmPFC patients appear together in another publication, uncited by two-factor theorists, in which the full extent of their damage is documented. These four cases not only lack skin responses to familiar faces, but lack responses to salient psychological stimuli more generally, which challenges factor one. They also have damage outside vmPFC, including damage to rDLPFC, which challenges factor two.

Conclusion: Two factor theory is found lacking and should be reappraised.

Keywords

Delusions; 2-factor Theory; Modularity; Neuropsychology; Lesions

However beautiful the strategy, you should occasionally look at the results

Ian Gilmour/Winston Churchill

Over recent years, *Cognitive Neuropsychiatry* has become the premier journal for the debate on the neurobiology and cognitive psychology of delusions.

My own work associates aberrant prediction error – the mismatch between belief-based expectation and experience - with the formation of delusions in schizophrenia (Corlett et al., 2007). Whilst the relationship between prediction error and delusions has been replicated with different tasks (Romaniuk et al., 2010; Schlagenhauf et al., 2009), my work has still

*Corresponding author. philip.corlett@yale.edu.

been criticized in the pages of this journal (Griffiths, Langdon, Le Pelley, & Coltheart, 2014). Those criticisms have been rebutted (Corlett & Fletcher, 2015), but the debate around delusions, their underlying mechanisms, and those of the mind and brain more broadly, continues (Tappin & Gadsby, 2019; Williams, 2018).

Recent empirical work on the resting-state functional connectivity of lesion locations identified from neurological patients with delusions (connectivity measured in healthy control data) has suggested that a two-factor explanation of delusions (driven by independent impairments in belief and perception, see below) may be more appropriate (Darby, Laganieri, Pascual-Leone, Prasad, & Fox, 2017). This work was not conclusive (McKay & Furl, 2017).

Taken together, 2-factor and prediction error accounts of delusions have garnered much interest and spurred continued debate and exchanges (see for example, <http://imperfectcognitions.blogspot.com/2014/03/explaining-delusional-belief-two-factor.html>, <http://imperfectcognitions.blogspot.com/2014/03/a-reply-to-max-on-explaining-delusions.html?m=1>). I think this arises partly because the arguments have clinical import and because delusions are inherently fascinating: they concern our beliefs which we cherish and often find hard to relinquish. There is a broader appeal here too, in that, whilst the debates focus on delusions, at their heart, they concern how best to map the mind and brain to one-another, which is of central relevance to the enterprise of cognitive neuropsychiatry.

In this piece, I critically evaluate the evidence for and logic of the two-factor theory of delusions. Crucially, I have recently uncovered a new piece of evidence in the literature which I believe undermines two-factor theory (Coltheart, 2010; Coltheart, Langdon, & McKay, 2011; Davies, Coltheart, Langdon, & Breen, 2001). More broadly, I will consider the assumptions underlying the theory – in particular, the assumed uni-directional flow of information from an encapsulated perceptual module into a cognitive belief-evaluation module. This assumption is challenged by recent data that suggest perception can sometimes be penetrated by cognition.

What is two-factor theory?

Two-factor theory was devised to explain the monothematic delusions that arise following traumatic brain injury, such as Capgras (that a loved one, familiar conspecific or object has been replaced by an imposter that looks the same but does not feel the same). The two-Factor explanation holds that people experience the lack of familiarity for their loved one (Factor 1) and they adopt the belief that they are an imposter due to a belief evaluation deficit (Factor 2) (Coltheart, 2010; Coltheart et al., 2011; Davies et al., 2001).

A second factor was deemed necessary because of evidence from four neuropsychological patients who lack familiarity responses to familiar faces, but do not have delusions (Tranel, Damasio, & Damasio, 1995) – these four people are one focus of this piece. These individuals are characterized by having damage to their ventromedial prefrontal cortex (vmPFC), and their familiarity response is operationalized as their lack of a skin conductance response (SCR) to famous faces like Marilyn Monroe or Margaret Thatcher.

Since these patients but do not have delusions, 2-factor theorists suggest the need for a second, delusion-generating factor. Crucially, 2-factor theory is based on the observation that these four individuals all have circumscribed damage in bilateral ventromedial prefrontal cortex (Tranel et al., 1995).

People with neurological delusions usually have damage to dorsolateral prefrontal cortex, often on the right (Coltheart, 2010; Coltheart et al., 2011; Davies et al., 2001). In empirical work, using transcranial magnetic stimulation (TMS), Coltheart and colleagues find that right DLPFC disruption with TMS (but not left) increases the objective hypnotizability of participants (they are susceptible to more tells of having been hypnotized) which the authors interpret as an increase in the malleability of their beliefs (Coltheart et al., 2018). Note that these data have nothing directly to do with delusions *per se*, only in so far as hypnosis is an appropriate model of delusions. Two-Factor theory for Capgras thus suggests that vmPFC damage subtends Factor 1 (the loss of familiarity response to familiar faces, so ‘That doesn’t feel like my wife’) and rDLPFC damage underlies Factor 2 (inappropriately adopting the belief that my wife is an imposter) (Coltheart, 2010; Coltheart et al., 2011; Davies et al., 2001).

Other Monothematic Delusions

There are other monothematic delusions that are a consequence of head injury, and many have been subjected to two-factor theory analysis. For example, Cotard (I believe I am dead), Fregoli (I believe that strangers are people familiar to me in disguise), Mirrored-Self Misidentification (the person I see reflected in the mirror is not me), and Somatoparaphrenia (This isn’t my arm, it belongs to someone else). Two-factor theory posits that each of these cases shares a common factor: dysfunction in belief evaluation, as a result of damage to the right hemisphere, likely the dorsolateral prefrontal cortex. However, each delusion has damage elsewhere that confers the particular content of the delusion. Thus, Cotard patients should lack affective responses to any stimuli. Fregoli patients should have excessive familiarity responses to unfamiliar individuals. People with alien control delusions should, according to the theory, lack proprioceptive feedback from their own actions.

The 2-factor account of Mirrored-Self Misidentification suggests that mirror agnosia (a loss of knowledge about how mirrors work) renders mirror experiences unusual. As with the Capgras delusion, there are people with mirror agnosia but who do not have mirrored-self misidentification, and indeed individuals with sensory polyneuropathy who can move, but who lack proprioceptive feedback and who do not delusionally disown their limbs. In these cases, the 2-factor theorist again demands a second factor. However, unlike Capgras, Cotard and Fregoli, no objective measures have been identified that could be used to confirm that these pairs of conditions share common phenomenology. Furthermore, despite the clear, empirically verifiable, prediction, no one has reported a lack of affective responsivity in Cotard patients. Coltheart points to pure autonomic failure patients, who, as a result of diffuse and non-specific peripheral nervous system degeneration, lack affective responses to salient stimuli, but also delusions. These cases serve the purported function of the vmPFC controls for a two-factor account of Cotard delusion, in that they have a perceptual deficit but no delusions, and so they demand a second factor. However, in the absence of SCR data

from Cotard patients, it remains premature to declare Cotard as having been explained by 2-factor theory. Similarly, there has been no evidence of enhanced responding to unfamiliar faces in Fregoli, nor has the control condition (the vmPFC case equivalent) for Fregoli been identified. Again, without these two key aspects, we should not consider Fregoli delusion explained from the viewpoint of two-factor theory.

It seems then that Capgras is the flagship of the two-factor explanatory fleet. In summary, there are individuals who lack a SCR to familiar faces but who do not have delusions (the four vmPFC cases), and individuals with Capgras who lack the expected familiarity responses, have apparent damage to their right hemisphere, in particular right DLPFC and parietal cortex (the purported locus of belief evaluation). But as Churchill and/or Ian Gilmour admonish us, when a theory is perspicuous, we ought to take the time to examine its evidence base. What do we know about the four vmPFC cases?

Problems with the four Capgras vmPFC control cases?

In reviewing the literature, I recently discovered a paper by Damasio and Tranel (Tranel & Damasio, 1994) that reports more about the four vmPFC cases. Helpfully, Tranel and Damasio publish the case identifiers across their papers on different topics. This makes it possible to get a more complete picture of their brain damage, their SCR and their psychological performance. I learned that the cases (EVR-318, MR-429, FL-1164, HS-1065) who lack SCR to familiar faces and have vmPFC damage have:

1. More extensive damage –beyond vmPFC (Tranel & Damasio, 1994). EVR-318, MR-429, FL-1164, and HS1065 all had damage to anterior cingulate cortex. *Critically*, EVR-318 and HS1065 had damage to right dorsolateral prefrontal cortex.
2. More extensive skin conductance changes beyond familiar faces – they lack a response to any psychologically salient visual stimuli (Tranel & Damasio, 1994), but not to salient physical stimuli (a deep inhalation and a loud hand-clap).

This means that:

1. rDLPFC damage is not sufficient for belief evaluation dysfunction as claimed by two-factor theorists
2. The four vmPFC cases on which Factor 1 is predicated do not have a specific enough deficit to determine the Capgras delusion content, since the individuals also lack arousal responses to other kinds of psychologically salient stimuli.

Furthermore, the four vmPFC individuals also have challenges in using autonomic interoceptive signals to make decisions about rewards and social situations. Whilst they do not have Capgras delusion, they are described as sociopaths, and they make odd moral decisions – they find murder more morally permissible than control participants (Young et al., 2010). This is consistent with a belief evaluation deficit, and these cases are thus not doxastically normal. Their broader SCR dysfunction may actually be more consistent with that hypothesized for Cotard (though it is possible that Cotard cases would also lack responses to physical as well as psychologically salient stimuli too – in the absence of clear

theorizing and empirical data from the two-factor perspective, it is difficult to be sure). Taken together, these data call for a re-appraisal of two-factor theory and a rejection of some of its major tenets. At the very least, the right-dorsolateral prefrontal cortex cannot be the locus of the belief evaluation deficit, the vmPFC cases ought to be renamed, and they should not be considered appropriate controls for Capgras.

Coltheart proclaims, *'I'm still an ultra-cognitive-neuropsychologist after all these years'* (2004, p. 21), and proposes that findings from neuroimaging have no evidential value for confirming or refuting cognitive psychological theories (p. 22): *[N]o facts about the activity of the brain could be used to confirm or refute some information-processing model of cognition. This is why the ultra-cognitive neuropsychologist's answer to the question 'Should there be any "neuro" in cognitive neuropsychology?' is 'Certainly not; what would be the point?'* Whilst the neuro in question, which elicited that comment, was functional neuroimaging, it seems that the structural neuroimaging of these four cases may be an aspect of neuroscience that ought to concern even the ultra-cognitive neuropsychologist. In the following section I consider whether two independent factors are necessarily implied by the available data.

Logic of Double Dissociation

A double dissociation is the golden-road to inferring, neuropsychologically, that two cognitive processes are independent. I now weigh whether the data upon which two-factor theory is predicated involve a double dissociation or not. Double dissociation is cognitive neuropsychology's;

'single most powerful tool, which was and is currently used to reveal the functional independence of discrete components of the cognitive system' (Vallar, 2004, p. 45).

Its evidence base is typically as follows: Two patients, A and B. Patient A has difficulty performing cognitive tests for, say auditory memory but has no problem with visual memory. Patient B has the opposite problem. Based on A and B, we would conclude that auditory memory and visual memory are doubly-dissociable cognitive processes. Parkin gives the following example:

"If your TV set suddenly loses the color you can conclude that picture transmission and color information must be separate processes (single dissociation: they cannot be independent because you cannot lose the picture and still have the color). If on the other hand you have two TV sets, one without sound and one without a picture you can conclude that these must be two independent functions (double dissociation)." (Parkin, 1996).

Coltheart offers another, more instructive, metaphor: that of a chocolate factory. What can we learn of the operation of a chocolate factory by merely observing its inputs and outputs over sufficient to observe the of ways that it can break down? He assumes that all chocolate factories employ the same manufacturing process. Their inputs, among other things, include, whole cocoa beans, and yield, among other things, cocoa powder and chocolate bars. A breakdown in the chain of manufacturing operations that halts production of both cocoa powder and chocolate bars is less informative than a breakdown which halts one but not the

other product, a double dissociation will have occurred if chocolate bar production continues when cocoa powder production falters, **and**, a failure of bar production has been observed in the face of sustained powder.

If Coltheart wishes, by his own chocolate factory rationale, to be more confident of two dissociable, independent, factors in delusions, I believe he needs to find a patient with Capgras delusions who doesn't have an SCR disruption, which, has not been reported (see Figure 1), and, paradoxically, I believe would obviate the two-factor explanation. Even if there were no issue with the vmPFC cases, the data upon which two-factor theory is predicated can only be a single, not a double dissociation. Damage to the perception/belief machinery (we might even call it a hierarchy) could simply be less extensive in the vmPFC cases. If it extended further, delusions would result (See Figure 2). The true double dissociation in the Capgras case, demonstrative of two independent processes, is that between Capgras and Prosopagnosia. Prosopagnosics are unable to recognize familiar people but sometimes has a feeling of knowing, Ellis & Michael, 2001), in this case, recognition and familiarity are dissociated, **not** perception and belief.

Discrete Modules?

The notion of discrete processing units emanates from a debate around the structure of the mind (and brain). In *The Modularity of Mind* (1983), Jerry Fodor argued that the mind contains modules – informational systems that take in a specific kind of information and no other kind (Fodor, 1983). For example, an early vision module takes ambient light as an input, and outputs color representations. The modularity project developed in two stages. It began as a sketch of a blueprint for an architecture of human cognition, which was seen as organized more or less into informationally-encapsulated modules. The project developed into a program for elaborating the types of things humans should have evolved to do: what modules should we have? What modules would have been beneficial to our hunter-gather ancestors? Unfortunately, both aspects were grossly underspecified. Fodor never specifically defines what he believes a module to be and proponents of massive modularity are seemingly comfortable inventing more and more modules (like adding apps to a smart phone). This has led to a curious persistence of the idea of modularity and a damning lack of progress in defining the cognitive architecture. Twyman and Newcombe write:

“Given this lack of agreed-upon definition, the modularity position becomes analogous to the Hydra, the many-headed monster that Heracles found difficult to combat because there were too many heads to take on simultaneously, and, worse, because other heads grew while he addressed a specific one” (Twyman & Newcombe, 2010).

Coltheart acknowledges that Fodor's was a sketch of modularity not a set of rules. Ironically, in his earlier writings on cognitive neuropsychology, prior to two-factor theory, the one feature of modules upon which Coltheart insists is domain specificity – “*A cognitive system is domain-specific if it only responds to stimuli of a particular class: thus, to say that there is a domain-specific face-recognition module is to say that there is a cognitive system that responds when its input is a face, but does not respond when its input is, say, a written word, or a visually-presented object, or someone's voice*” (Coltheart, 1999, p. 118).

Factor-2, belief evaluation, which Coltheart insists is a module, seems far from domain-specific – since beliefs can be about almost any topic – therefore it would not count as a module for Fodor. Puzzlingly, Factor two does not meet Coltheart's own criterion for modularity despite two-factor theory being predicated on neural and mental modularity (Coltheart, 1999).

Cognitive neuropsychologists who favor modularity focus on functional segregation. They conceive of functionally specialized segregated brain regions that are responsible for discrete mental functions that can be damaged in isolation, leaving other functions intact. The notion of double dissociation used in cognitive neuropsychology omits the requirement of different lesion locations that is present in Teuber's (1955) definition. However, in Coltheart's own words:

'Because cognitive neuropsychology is possible, it would seem that both the mind and the brain are modular in structure.' (Coltheart, 2002).

Thus, modules must be neurally circumscribed (in order for them to be damaged independently and yield independent mental deficits).

Plunkett and Bandelow remark (2006, p. 208):

'When accompanied by evidence of non-overlapping lesions, the inference of separable mechanisms from behavioural double dissociations carries added weight.'

If this is correct, then a neuroscientific finding—that two patients have lesions in different locations—may enhance the value of behavioral double dissociation in shifting the balance of support between competing cognitive theories. By this logic, that the vmPFC cases also have DLPFC damage must weaken the support for 2-factor theory. It must shift the balance away from it. More broadly, any evidence against modularity (of mind and brain) represents a challenge to the two-factor project.

Permissible Directions of Processing

Recent work in cognitive science has argued (like Pylyshyn (Pylyshyn, 1999) and Coltheart (Coltheart, 1999) before), that cognition does not influence (or penetrate) perception (Firestone & Scholl, 2016). However, work in computational cognitive neuroscience and neural connectivity in the healthy brain as well as certain psychiatric and neurological states call this claim into question. I have proposed that top-down influences of cognition on perception are exemplified by hallucinations—perceptions without stimulus (Esquirol, 1814)—both in healthy people and in those who suffer serious mental illnesses like schizophrenia (Powers et al, 2016). In recent work, my colleagues and I demonstrated that these mental symptoms can be understood in terms of a model of mind and brain function that rejects strict, encapsulated, neuropsychological modularity and embraces some penetrability of perception by cognitive processes, namely predictive coding (A. R. Powers, Mathys, & Corlett, 2017). The observation that, even in healthy people, visual expectations can engender auditory hallucinations under the influence of precise prior beliefs, is consistent

with the penetration of perception by cognition and inconsistent with encapsulated modularity (A. R. Powers, III, Kelley, & Corlett, 2016).

On this issue, Coltheart and colleagues say the following:

“Importantly, all Two Factor accounts are predicated on a conceptual distinction (and empirical dissociation) between perception and cognition: abnormal perception as the first factor and a cognitive belief evaluation deficit as the second factor. Furthermore, Two Factor accounts are not committed to perception being cognitively penetrable, meaning that the Two Factor theory is consistent with the hypothesis presented by Firestone and Scholl.” (Ross, McKay, Coltheart, & Langdon, 2016).

To the extent that Firestone and Scholl must make an exception for conditioned hallucinations and perhaps hallucinations more broadly, these data count against modularity and 2-factor theory; two-factor theory is not explicit about cognition penetrating perception, but two factor theorists claim that perception and belief are dissociated modules, with a uni-directional flow of information from perception to belief (else only one factor – damage to top-down belief effects – would suffice – see Figure 1 – again, regardless of issues with the vmPFC cases, perception and belief are not completely dissociated), so the rejection of cognitive penetration is implicit and I believe essential to two-factor theory as espoused by Coltheart and colleagues.

Modern network neuroscience embraces a mathematical definition of modularity that, I argue, is different from psychological modularity. Instead, network modularity shares more with hodotopic ideas that combine functional specialization (within brain region - topological) and integration (across brain regions - hodological), rather than rigidly ascribing functions to particular regions and denying inter-module communication (Fotopoulou, 2014). This communication proffers a pathway forward, outlined below, but first I will summarize my points.

Summary

I believe that the evidence from the four vmPFC patients does not support a 2-factor explanation for Capgras delusion. I am concerned about the temporal demands that would need to be satisfied for two-factor theory to work. Finally, the encapsulated modularity of perception from belief on which two-factor theory is predicated does not seem to obtain.

A colleague and collaborator of Coltheart, and co-creator of two-factor theory, Martin Davies’ remarks of cognitive neuropsychology (Davies, 2010), apply to the strange status that 2-factor theory has attained (above reproach, immune to criticism, unmolested by direct empirical examination):

“...it is a familiar point about science in general that there is no logically valid deductive inference from evidence to explanatory theory. To be explanatory, a theory must go beyond a summary of the evidence and so cannot be entailed by the evidence. Within the narrower domain of psychology, it is well understood that there is no logically valid deductive inference from data, such as reaction time data,

to an explanatory theory about cognitive structures and processes. It is very unlikely that cognitive neuropsychologists regard their research program as being different from all the rest of empirical science and take themselves to have access to evidence with magical properties.” (Davies, 2010)

I suggest that acolytes of two-factor theory attribute it the magical properties Davies outlines. It appears to me to be a deduction that is unwarranted by the data. While Coltheart would likely respond that argumentation in cognitive neuropsychology is abductive inference (to the best explanation), I would contend that two-factor theory, as he and his colleagues espouse, is an erroneous summary of rather scant evidence, and so not the best available explanation of those data, nor of delusions. More broadly, it is troubling that, despite having been proposed almost 20 years ago, the SCR predictions for delusions other than Capgras have never been tested, factor two remains undefined, and most concerning, the DLPFC lesions and SCR deficits of the four vmPFC cases which undermine the theory have not been acknowledged. two-factor theory is not the best explanation for the extant data.

Pathway forward

When one criticizes a model, it is appropriate to offer steps forward that might help settle the areas of contention between the critique and the model. I shall suggest three, but no doubt there are more.

1. The predictions about SCR responses in delusions other than Capgras should be tested. Do Cotard cases lack any SCR to salient stimuli? Do Fregoli cases evince familiarity responses to unfamiliar faces? Most importantly, do all of these deficits, including those in Capgras, have the requisite precision to confer delusion contents? In Capgras, a dearth of responding to famous people is not enough, unless those famous people are the subjects of the delusion.
2. Some commitments to the features, mechanisms, and boundaries of Factor-Two - the belief evaluation deficit - must be made. The one empirical study using TMS, outlined above, related right DLPFC function to hypnotizability not delusions per se, and the DLPFC damage in the vmPFC cases represents a challenge to this neurological place-holder of belief-evaluation in right DLPFC.
3. A crux of the disagreements between 2-factor and other models of delusions involves the directionality of information flow between perception and belief. A key question then is whether or not beliefs influence perception top-down and whether that mechanism relates to delusions. There is nothing in the tenets of cognitive neuropsychology that forbids two-way arrows between modules (Ellis, 2014). Top-down effects of belief on perception have been identified, in perceptual inference, for example, beliefs impact perceptual inferences via a directional connection from orbitofrontal cortex to area V5 in a manner that correlates with delusion severity (Schmack, Rothkirch, Priller, & Sterzer, 2017). In more abstract inferences, top-down connections from DLPFC correlate with delusion severity in patients with schizophrenia (Kaplan et al., 2016). These specific behavioral tasks should be examined in neurological patients with

delusions, in order to examine the directional question. To further examine the hodological version of the model, one could elaborate upon the findings of Darby, Fox and others. In brief they used functional connectivity to identify hubs of dysconnectivity in DLPFC and retrosplenial cortex that were necessary for the presence of delusions in neurological cases (Darby et al., 2017). What is the relationship between these hubs? Even resting data can be subject to effective (i.e. directional) connectivity analyses (Park, Friston, Pae, Park, & Razi, 2018). If the DLPFC hub exerts a temporally prior (i.e. top-down) influence over the retrosplenial hub then one would favor theories that permit bi-directional information flow between perception and belief, and one would reject encapsulated modularity. If there is no top-down influence of prefrontal cortex on retrosplenial cortex, then encapsulated modularity and a uni-directional flow from perception to belief would be more likely to obtain. Furthermore, these connectivity metrics should be acquired in people with delusions, in addition to the healthy control sample that Darby and colleagues employed.

Conclusion

I have summarized the evidence base and logic of the two-factor theory of delusions and have found these to be lacking. Specifically, evidence from the cases upon which the theory is predicated suggests that they are importantly different from the claims of two-factor theorists. Furthermore, the empirical predictions of two-factor theory have gone untested for almost 20 years. Finally, the separation of perception from belief upon which two-factor theory rests may not obtain. These latter criticisms ought to be addressed with empirical data. The commitments of two-factor theory to neural and mental organization have been detailed and some predictions, based on those commitments, have been offered. Whilst resting on laurels, waiting for a competitor, is something Coltheart has advocated (Coltheart, 2004), it would appear that declaring laureate status in the case of two-factor theory may have been premature.

Acknowledgments

I am supported by NIMH R01MH112887, NIMH R21MH116258 as well as a Rising Star award from the International Mental Health Research Organization, the Clinical Neurosciences Division, U.S. Department of Veterans Affairs, and the National Center for Post-Traumatic Stress Disorders, VACHS, West Haven, CT, USA. and by the Yale University School of Medicine and Department of Psychiatry. The contents of this work are solely my responsibility and do not necessarily represent the official view of NIH or the CMHC/DMHAS. I have no relevant biomedical conflicts of interest. I am grateful to Katherine Thakkar, Albert Powers III and Jane Garrison for their comments on an earlier version of this manuscript. Their acknowledgments does not imply agreement with this piece and any errors that remain are mine.

References

- Coltheart M (1999). Modularity and cognition. *Trends Cogn Sci*, 3(3), 115–120 [PubMed: 10322463]
- Coltheart M (2002). Cognitive neuropsychology. In Pashler H, Wixted J (Ed.), *Stevens' handbook of experimental psychology: Methodology in experimental psychology* Hoboken, NJ: John Wiley & Sons.
- Coltheart M (2004). Brain imaging, connectionism, and cognitive neuropsychology. *Cogn Neuropsychol*, 21(1), 21–25. doi: 10.1080/02643290342000159 [PubMed: 21038184]

- Coltheart M (2010). The neuropsychology of delusions. *Ann N Y Acad Sci*, 1191, 16–26. doi: 10.1111/j.1749-6632.2010.05496.x [PubMed: 20392273]
- Coltheart M, Cox R, Sowman P, Morgan H, Barnier A, Langdon R, ... Polito V. (2018). Belief, delusion, hypnosis, and the right dorsolateral prefrontal cortex: A transcranial magnetic stimulation study. *Cortex* doi: 10.1016/j.cortex.2018.01.001
- Coltheart M, Langdon R, & McKay R (2011). Delusional belief. *Annu Rev Psychol*, 62, 271–298. doi: 10.1146/annurev.psych.121208.131622 [PubMed: 20731601]
- Corlett PR, & Fletcher PC (2015). Delusions and prediction error: clarifying the roles of behavioural and brain responses. *Cogn Neuropsychiatry*, 20(2), 95–105. doi: 10.1080/13546805.2014.990625 [PubMed: 25559871]
- Corlett PR, Murray GK, Honey GD, Aitken MR, Shanks DR, Robbins TW, ... Fletcher PC (2007). Disrupted prediction-error signal in psychosis: evidence for an associative account of delusions. *Brain*, 130(Pt 9), 2387–2400. doi: 10.1093/brain/awm173 [PubMed: 17690132]
- Darby RR, Laganier S, Pascual-Leone A, Prasad S, & Fox MD (2017). Finding the imposter: brain connectivity of lesions causing delusional misidentifications. *Brain*, 140(2), 497–507. doi: 10.1093/brain/aww288 [PubMed: 28082298]
- Davies M (2010). Double dissociation: Understanding its role in cognitive neuropsychology. *Mind & Language*, 25(5), pp. doi: 10.1111/j.1468-0017.2010.01399.x
- Davies M, Coltheart M, Langdon R, & Breen N (2001). Monothematic delusions: Towards a two-factor account. *Philosophy, Psychiatry, & Psychology*, 8(2–3), pp. doi: 10.1353/ppp.2001.0007
- Ellis AW, Young AW (2014). *Human Cognitive Neuropsychology (Classic Edition)* London & New York: Psychology Press.
- Esquirol J (1814). *Hallucinations. Dictionnaire des sciences médicales* Paris: CLF Panckouche, 1822
- Fodor JA (1983). *The modularity of mind : an essay on faculty psychology* Cambridge, Mass.: MIT Press.
- Fotopoulou A (2014). Time to get rid of the ‘Modular’ in neuropsychology: a unified theory of anosognosia as aberrant predictive coding. *J Neuropsychol*, 8(1), 1–19. doi: 10.1111/jnp.12010 [PubMed: 23469983]
- Griffiths O, Langdon R, Le Pelley ME, & Coltheart M (2014). Delusions and prediction error: re-examining the behavioural evidence for disrupted error signalling in delusion formation. *Cogn Neuropsychiatry*, 19(5), 439–467. doi: 10.1080/13546805.2014.897601 [PubMed: 24702287]
- Kaplan CM, Saha D, Molina JL, Hockeimer WD, Postell EM, Apud JA, ... Tan HY. (2016). Estimating changing contexts in schizophrenia. *Brain*, 139(Pt 7), 2082–2095. doi: 10.1093/brain/aww095 [PubMed: 27217338]
- McKay R, & Furl N (2017). Doppelgangers and dissociations: lesion network mapping illuminates misidentification delusions. *Brain*, 140(2), 262–265. doi: 10.1093/brain/aww323 [PubMed: 28137956]
- Park HJ, Friston KJ, Pae C, Park B, & Razi A (2018). Dynamic effective connectivity in resting state fMRI. *Neuroimage*, 180(Pt B), 594–608. doi: 10.1016/j.neuroimage.2017.11.033 [PubMed: 29158202]
- Parkin AJ (1996). *Explorations in Cognitive Neuropsychology* Oxford: Blackwell.
- Powers AR III, Kelley M, & Corlett PR (2016). Hallucinations as top-down effects on perception. *Biol Psychiatry Cogn Neurosci Neuroimaging*, 1(5), 393–400. doi: 10.1016/j.bpsc.2016.04.003 [PubMed: 28626813]
- Powers AR, Mathys C, & Corlett PR (2017). Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors. *Science*, 357(6351), 596–600. doi: 10.1126/science.aan3458 [PubMed: 28798131]
- Pylyshyn Z (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behav Brain Sci*, 22(3), 341–365; discussion 366–423 [PubMed: 11301517]
- Romaniuk L, Honey GD, King JR, Whalley HC, McIntosh AM, Levita L, ... Hall J. (2010). Midbrain activation during Pavlovian conditioning and delusional symptoms in schizophrenia. *Arch Gen Psychiatry*, 67(12), 1246–1254. doi: 10.1001/archgenpsychiatry.2010.169 [PubMed: 21135324]
- Ross RM, McKay R, Coltheart M, & Langdon R (2016). Perception, cognition, and delusion. *Behav Brain Sci*, 39, e258. doi: 10.1017/S0140525X15002691 [PubMed: 28355861]

- Schlagenhauf F, Sterzer P, Schmack K, Ballmaier M, Rapp M, Wrase J, ... Heinz A. (2009). Reward feedback alterations in unmedicated schizophrenia patients: relevance for delusions. *Biol Psychiatry*, 65(12), 1032–1039. doi: 10.1016/j.biopsych.2008.12.016 [PubMed: 19195646]
- Schmack K, Rothkirch M, Priller J, & Sterzer P (2017). Enhanced predictive signalling in schizophrenia. *Hum Brain Mapp*, 38(4), 1767–1779. doi: 10.1002/hbm.23480 [PubMed: 28097738]
- Tappin BM, & Gadsby S (2019). Biased belief in the Bayesian brain: A deeper look at the evidence. *Conscious Cogn*, 68, 107–114. doi: 10.1016/j.concog.2019.01.006 [PubMed: 30665186]
- Tranel D, & Damasio H (1994). Neuroanatomical correlates of electrodermal skin conductance responses. *Psychophysiology*, 31(5), 427–438 [PubMed: 7972597]
- Tranel D, Damasio H, & Damasio AR (1995). Double Dissociation between Overt and Covert Face Recognition. *J Cogn Neurosci*, 7(4), 425–432. doi: 10.1162/jocn.1995.7.4.425 [PubMed: 23961902]
- Twyman AD, & Newcombe NS (2010). Five reasons to doubt the existence of a geometric module. *Cogn Sci*, 34(7), 1315–1356. doi: 10.1111/j.1551-6709.2009.01081.x [PubMed: 21564249]
- Williams D (2018). Hierarchical Bayesian models of delusion. *Conscious Cogn*, 61, 129–147. doi: 10.1016/j.concog.2018.03.003 [PubMed: 29602712]
- Young L, Bechara A, Tranel D, Damasio H, Hauser M, & Damasio A (2010). Damage to ventromedial prefrontal cortex impairs judgment of harmful intent. *Neuron*, 65(6), 845–851. doi: 10.1016/j.neuron.2010.03.003 [PubMed: 20346759]

Figure 1. Perception, Belief and Double Dissociation

A. The proposed architecture of belief formation and updating: From perception to belief

B. Factor 1: The perceptual deficit evinced by vmPFC patients

C. The Missing Dissociation: Belief damage in the absence of perception

Figure 2. A Hierarchical Alternative, with Top-Down Effects

A. A layered hierarchy. Predictions are conveyed top-down, prediction errors bottom-up

B. Lower Level Damage. Nearer to sensory input. Aberrant prediction errors would be conveyed up the hierarchy. They would be reconciled as delusional beliefs. This may be the damage sustained by vmPFC patients, but the prediction errors are not precise or persistent enough to augur delusions.

C. Higher Level Damage. It is possible, in a hierarchical model, for dysfunction at the apex of the hierarchy to alter the precision of top-down priors such that they alter predictions (and prediction errors) lower in the hierarchy. This could obtain in the Capgras cases.