Schizophrenia Bulletin vol. 43 no. 2 pp. 263–272, 2017
doi:10.1093/schbul/sbw194
Advance Access publication February 8, 2017

COMMENTARY

Explaining Delusions: Reducing Uncertainty Through Basic and Computational Neuroscience

Erin J. Feeney[1,2], Stephanie M. Groman[1], Jane R. Taylor[1] and Philip R. Corlett*,[1]

[1]Department of Psychiatry, Ribicoff Research Facilities, Connecticut Mental Health Center, Yale University, 34 Park Street, New Haven, CT 06511, USA; [2]Interdepartmental Neuroscience Program, Yale University, New Haven, CT 06511, USA

*To whom correspondence should be addressed; tel: 203-974-7866, e-mail: Philip.corlett@yale.edu

Delusions, the fixed false beliefs characteristic of psychotic illness, have long defied understanding despite their response to pharmacological treatments (e.g., D_2 receptor antagonists). However, it can be challenging to discern what makes beliefs delusional compared with other unusual or erroneous beliefs. We suggest mapping the putative biology to clinical phenomenology with a cognitive psychology of belief, culminating in a teleological approach to beliefs and brain function supported by animal and computational models. We argue that organisms strive to minimize uncertainty about their future states by forming and maintaining a set of beliefs (about the organism and the world) that are robust, but flexible. If uncertainty is generated endogenously, beliefs begin to depart from consensual reality and can manifest into delusions. Central to this scheme is the notion that formal associative learning theory can provide an explanation for the development and persistence of delusions. Beliefs, in animals and humans, may be associations between representations (e.g., of cause and effect) that are formed by minimizing uncertainty via new learning and attentional allocation. Animal research has equipped us with a deep mechanistic basis of these processes, which is now being applied to delusions. This work offers the exciting possibility of completing revolutions of translation, from the bedside to the bench and back again. The more we learn about animal beliefs, the more we may be able to apply to human beliefs and their aberrations, enabling a deeper mechanistic understanding.

Key words: delusions/behavioral neuroscience/cognitive neuroscience/computational psychiatry/associative learning

The End of the World and the Beginning of a Theory

In December 1954, the Chicago Tribune reported that Dorothy Martin was relaying a prophecy from extra-terrestrials that the world was about to end. A number of followers flocked to her but her prophecy did not manifest. She ultimately settled in Sedona, Arizona where she lived until she was 92, continuing to proselytize about aliens but evading interaction with psychiatric services. Did Martin have delusions? What about her acolytes? Defining, explaining, and ultimately understanding delusions has proven challenging (see Freeman and Bebbington, this issue). In this article, we describe how neuroscientists have tried to meet that challenge. The approach demands some simplifying assumptions. Basic neuroscience in preclinical models will not recapitulate all of the features of delusions. However, simple models can be useful.[1]

After David Marr, we suggest a multilevel analysis might be helpful[2]: The Computational: *What is the information being manipulated and to what end?* The Algorithmic: *What are the manipulations?* And The Implementational: *How are those manipulations manifest by neural signaling?* Marr believed that one could bridge these levels of analysis. We contend that the same applies with computational psychiatry: one can map from phenomenology to neural activity,[3] perhaps by building a model of subjects' behavior, estimating model parameters, and relating them to neural activity and symptom severity. If one knows the neural/behavioral instantiation of the aberrant parameters, it may be possible to intervene and mollify the symptoms.[3]

Associative learning is our preferred framework for spanning levels of enquiry. Pavlov[4] and Skinner[5] defined associative paradigms that emphasized the importance of predictions and prediction errors (PEs) in how we and other animals learn about the world (see below). Learning theories[6] are the algorithms through which associative learning might be realized. PE signals, such as those observed in midbrain dopamine cells, are implementational aspects of the account.[7] Crucially, each level pertains to both animals and humans and we believe perturbations of those levels may underlie delusions.[8]

From Acolytes to Animals

Unbeknownst to Martin, some of her followers were imposters: social psychologists, led by Leon Festinger. The academics infiltrated the group as the end-times loomed. The result was a book; "When Prophecy Fails: A social psychological study of a modern group that predicted the destruction of the world."[9] They developed the theory of cognitive dissonance, the internal discord felt from holding conflicting beliefs simultaneously[10]—in this case, between the prophecy and real world events. People in the cult responded in a variety of ways to reduce their dissonance. Many relinquished their beliefs. In some cases, however, a dissonant experience actually increased conviction. For example, failed predictions were re-contextualized as actually having come to fruition ("the aliens did come for us, but they were scared off by the crowds of press"). These deft sleights of mind[11] will be familiar to those who have spoken to patients with delusions.[12,13]

Rodents may experience cognitive dissonance. When they learn that pressing a lever leads to reward, rats might be said to believe their action leads to the reward.[14] If we degrade the contingency between action and reward, eventually pressing decreases.[5] However, immediately following the contingency change, pressing can increase dramatically.[15] Such extinction bursts[15] have been equated with the increased conviction displayed by The Seekers when their prophecy failed.[16] These superficial similarities are encouraging, but we are far from a mature understanding of human and animal belief and delusion.

Why so Little Progress?

Animal experimentation has led to remarkable progress in the treatment of human suffering. However, early drug models of hallucinations and delusions entailed anthropomorphism on the part of the experimenter, who had to infer that an animal was believing and perceiving things incorrectly. Early experimenters were discouraged. The absence of a robust psychological theory of belief formation led Jerry Fodor to assert that, whilst beliefs are among the most interesting phenomena, they are not ready to be explained in the same cognitive and neural terms as more accessible processes, such as vision.[17,18] However, there are now cognitive and neural theories of belief[19] amenable to quantitative analysis[20] across species,[19] and in the clinic.[21]

Drug Models

Delusions are challenging to study—the sufferer often denies any problem.[22] Experimental models provide a unique window onto an otherwise inaccessible disease process.[21] Ketamine, the N-methyl-D-aspartate (NMDA) glutamate receptor antagonist, transiently and reversibly engenders delusion-like ideas in healthy people[23]and other animals.[24] These delusions might be manifestations of aberrant PE,[8] the mismatch between our expectancy in a given situation and what we experience.[25] Derived from formal learning theory to explain mechanisms of animal conditioning, PE[25] is signaled by dopamine and glutamate activity in the brain.[26] It has also become a key process in theoretical models of human causal learning and belief formation.[19] By minimizing PE, we improve our ability to anticipate the causal structure of our environment and we form causal beliefs.[19] Inappropriate PE garners aberrant association and delusional belief.

PE may be excessively large or it may occur chaotically in people forming delusions.[8,27] Describing his own delusions, Peter Chadwick[28] notes that when he sits on the bus he notices that all of the doors on the left hand side of the street are painted red. He moves to the right and notices all the doors on the right are painted green. This strikes him as puzzling. He concludes that things must have been arranged that way by a nefarious Organization. PE coincident with the sensation of a green door and a red door drove an association between the doors' representation and that of The Organization, manifesting clinically as a delusion.[8,27]

Our account resonates with notions of aberrant evolutionary threat detection: a cognitive model of paranoia.[29] Threat detection is essentially a signal detection task.[30] Associative learning has been described in similar terms.[31] In paranoid individuals, the bias toward detecting threat where there is none[30] may generalize beyond threat to perception and cognition more broadly. By examining the neural and behavioral mechanisms of associative contingency detection in animals, we may well gain new insights into situations where that detection has gone awry. We turn briefly to the history of associative ideas.

From Associationism to Computational Psychiatry

Belief is typically defined as assent to the truth value of some proposition. This is a fine definition at the computational level, but we are also interested in algorithm and implementation. Thus, we align belief with learning and memory.[19] We learn beliefs from our own experiences and from others. And we store that learning as a network of associations between representations.[32] This began with Plato.[33] Aristotle outlined the first laws of association.[34] John Locke described the role of improper association of ideas in mental illness.[35] Pavlov explored the mechanisms of association empirically.[4] His conditioning paradigms have highlighted that mere contiguity is not sufficient for learning. For example, Leon Kamin discovered blocking, which involves the retardation of learning about a novel cue-outcome association when that cue is paired with a stimulus that already predicts the outcome—the pretrained cue "blocks" learning about the novel cue.[36] Blocking demands that the association of ideas is sensitive to surprise.[37]

Widrow and Hoff created a simple connectionist neural network of nodes, representing inputs and outputs as links between nodes.[38] Those links were strengthened by reducing an error signal, the mismatch between the desired output from a given input and the output that actually occurred. A similar algorithm was proposed for animal conditioning by Rescorla and Wagner[25]; environmental stimuli induce expectations about subsequent states of the world, exciting representations of those states. Any mismatch between the expectancies and actual experience is a PE. PEs are used as teaching signals to update future expectancies about stimuli and states. Under this scheme, blocking occurs because the outcome of the compound of pretrained and novel cues is completely predicted by the pretrained cue, which precludes the generation of a PE signal and, subsequently, learning about the association between the novel cue and the outcome. Consequently, a chaotic PE that just happens to co-occur with the blocked cue should weaken blocking. This has been demonstrated with amphetamine administration,[39] chemogenetic manipulations of cingulate cortex in[40] and optogenetic manipulation of dopamine neurons in experimental animals.[41] In humans, weaker blocking has been observed in patients with schizophrenia,[42] and the extent to which the neural PE signal is inappropriately engaged correlates with delusion-like beliefs.[43]

The Role of Attention and Action

Attention is also critical for associative learning. We recognize the important impact of Kapur's perspicuous "incentive salience theory" of psychosis,[44] that delusions form as a consequence of aberrant incentive salience. Incentive salience is a quality from the behavioral neuroscience of addiction.[45] Events with incentive salience grab attention and drive goal directed action.[45] Kapur argued that excess dopamine release in the ventral striatum might cause individuals to inappropriately imbue certain events with salience and to form delusions as a way of explaining those experiences. We note that aberrant salience theory was presaged by more mechanistic theories grounded in associative learning theory[46,47] and that the data on dopamine release capacity[48] have implicated the associative striatum (not the ventral striatum) in the genesis of psychosis. Nevertheless, there do seem to be phenomenological and empirical data linking the broad category of aberrant salience to delusions.[49]

There are, of course, complexities. Some rational theories of belief demand that it drives action, that people can only be said to be believers if they act consistently with their beliefs. Some people with delusions do not act on them; they may claim they are being poisoned but nevertheless they eat. This double-bookkeeping may be explained by concurrent negative symptoms sapping the motivation to act.[50] However, even people without psychotic disorders do not act consistently with their beliefs; economists find a distinction between people's expressed and revealed preferences. Future work will need to explore this phenomenon. There is a need for a statistical theory that allows interactions between perception, action, and belief. We believe Bayes theorem may fill that role.

Bayes, Predictive Coding, and Associations

Thomas Bayes' doctrine of probabilities (published posthumously in 1873)[6] has had a striking impact upon science in recent years.[51] The theorem embodies a formal approach to reasoning about data using hypotheses and captures the probabilistic nature of many of the tasks faced by organisms: to predict the environment and respond appropriately, by minimizing uncertainty about subsequent inputs. Stimuli enter through sensory organs and their neural representations are sculpted through hierarchical processing in the brain. Top-down expectations are communicated downwards.[52] Any mismatch between expectation and current input (PE) is detected and must be resolved. Depending on the relative precision of priors and PE, the error is either ignored or used to update subsequent expectations with new learning.[53,54] Pavlov believed that Helmholtz' unconscious perceptual inferences were aligned with his conditioned responses.[4] Bayesian formalisms can explain blocking,[55] perceptual learning,[56] and visual perception.[57,58] Indeed, color-after effects in the visions[59] appear to be subject to blocking.[60–62] Taken together, these findings suggest a unified model of perception, action, and belief driven by predictions and PEs.

We do not think that there is one central PE, but rather multiple hierarchies of inference that converge on an a-modal model of the self. The hierarchies can influence one another—PE in one can alter inference in another, via this a-modal model.[63] There are also lots of ways that each PE may be perturbed—it may be too precise, it may not be precise enough, these effects could occur bottom-up (pathologies of the error signal) or top-down (problems with priors).[64] Thus far, in human work on delusions, we have not discerned which of these pertains and whether the effects are consistent within levels of a particular hierarchy or across hierarchies. For example, low-level sensory perturbations in a visual module could have effects on belief higher in the hierarchy—that is, weak sensory priors (and increased low-level PE) may render cognitive priors (higher in the hierarchy) more rigid.[65]

Different hierarchies (perception, action, and belief) may engage different neuromodulators to encode the precision of priors and PE, thus one hierarchy may be impaired while others remain relatively intact (some people have delusions, others hallucinations). On the other hand, the hierarchies do interact in nonlinear ways; the Capilano Bridge Experiment gives a useful illustration. Subjects who traversed a tall suspension bridge confused their fear for romantic attraction to a person they met on the other side of the bridge (compared with a group that traversed a lower, more stable bridge).[66] The point here is that we are not accurate at inferring our inner states and that one hierarchy's unresolved PE can influence that of another hierarchy. Allen and colleagues recently reported a psychophysical example: subliminally presented disgust faces change participants' heart rate and skin conductance. These peripheral bodily changes alter the impact of priors on visual decision making.[67]

Each hierarchy may have its own neuromodulators; dopamine for action beliefs, noradrenaline for interoceptive beliefs, acetylcholine for exteroceptive beliefs,[68] but of course these systems are inextricably linked. Boosting dopamine function in the ventral tegmental area (VTA) can change cholinergic function and sensory representation in primary cortices,[69] for example. Perhaps psychosis involves the penetration of one hierarchy by processing from another. This would lead to an inordinate influence of belief on perception (hallucinations) and vice versa of perception on belief (delusions). Perhaps certain types of delusions and hallucinations coincide more readily—for example, delusions of passivity and auditory verbal hallucinations. Such co-occurrence would be a key test of our theory. If some of that work is to take place in preclinical animal models, we must decide whether they have beliefs.

Do Animals Have Beliefs?

Whether we can reasonably impute beliefs in experimental animals is an intriguing philosophical question. The answer, of course, depends on how we understand belief. We, like Dan Dennett, think a permissive definition is helpful.[70] Some philosophers deny that animals have beliefs since they don't have language and can't express the distinction between their beliefs and what pertains.[71] There is a behavioral mark that is coextensive with having such a concept of what one believes and how it differs from reality: surprise (or PE).[72] PE does not require language. Wynn (1992) showed human infants a toy and then placed it behind a screen. Next she showed them another toy and also placed it behind the screen. The screen was then lowered, revealing either two toys (the expected outcome), or only one toy (surprise). Infants looked longer at the unexpected outcome. This surprise—evidence of prior belief—precedes the development of language.[73] Thus, because rodents and primates have PE signals, they too have beliefs.

Reasoning Rats?

If Davidson is correct and only organisms with propositional representations can have beliefs, do rats reason with propositions? Some say they do.[74] Rats can learn that on some occasions a light is followed by a tone and on others, light is followed by food. A group of observing rats (who saw these animals receiving mixed presentations of light-tone, and light-food) were subsequently given simple presentations of the tone (obscuring the light). If these animals are engaging in propositional inference, the tone should indicate that the common cause of food and tone has occurred and they should expect to receive food. The animals had this expectation and approached the food cup,[75] which is hard to reconcile with simple associative theory. These findings have been simulated with an associative model that activates a representation of the food through a complex chain of associations.[76] This seems similar to imagining what may arise. Because psychosis, delusions, and hallucinations may entail a confusion of perception and imagination,[77] paradigms that require retrieved representations for learning (imagination) may be particularly useful for clinical and preclinical investigations of psychosis.

Imagining Rats?

Dating back to Jerzy Konorski (1967), associatively retrieved internal representations have been implicated in the genesis of hallucinations.[78] For example, a hungry rat is presented with a tone, and, subsequently, a sweet, sugar solution. The rat learns after only a few trials that the tone predicts sugar. The tone evokes a highly realistic, sensory representation of the sugar, which the rat has trouble distinguishing from reality.[79] If one devalues sugar solution, by pairing it with a nausea-inducing agent, rats will reduce consumption of sugar thereafter. However, with more experience of tone-sugar pairing, the tone evokes a more abstract representation of the expected reward, which the rat readily distinguishes from reality and the association becomes more resistant to this mediated devaluation.[80] Neonatal ventral hippocampal lesions (NVHLs) disrupt the development of inputs to the prefrontal cortex.[81] These animals have construct validity for psychosis (e.g., perturbed prefrontal function[82]). Following mediated devaluation, NVHL animals (compared with controls) decrease food consumption both early and later in training[83]—suggesting that they retain the rich and realistic sensory representations of the sugar evoked by the presence of the tone.[83] Their reality monitoring does not mature with more experience of tone-sugar pairings.[83]

This mediated learning may have been shown in humans,[84] however, it would bear replication in patients with psychosis. Future preclinical work might employ some of the tools that have been brought to bear on blocking—like optogenetics[41] and chemogenetics[40]—to elucidate the neural and behavioral mechanisms of mediated devaluation and its failure. For example, midbrain dopamine neurons may signal the inferred value of the sugar mediated by the tone representation.[85] Ketamine induces aberrant firing of dopamine neurons.[86] Perhaps that firing underpins the preserved sensitivity to devaluation in the NHVL rats.

Delusion Persistence

In addition to being bizarre departures from reality, delusions are remarkably fixed. In the Chicago cult we began by describing, the end of the world never came, yet some individuals increased their credulity. Likewise, when delusions are questioned, bringing them to mind may actually serve to reinforce rather than to disrupt the memory.[8,27,87] The idea here is that re-evocation of an association may strengthen a memory even when it is not formally reinforced.[27] We have modeled this process in humans and rats with ketamine.[24] By creating new associations, reactivating them under ketamine, and then testing their strength, we observed that ketamine-reactivated associations were strengthened.[87] In humans, this effect correlated with ketamine-induced psychosis and PE brain signal.[87]

Reversal Learning

Reversal learning tasks also assay belief updating. They have been translated from preclinical models to patients. Subjects are challenged to select from two or more options (e.g., nose-poke terminals, or levers). The “correct” choice yields a rewarding outcome (for review, see Izquierdo et al., 2016[88]). Reward contingencies may be deterministic.[89] Here, adaptive responding requires recognition of the reversal—a switch from positive to negative feedback, followed by inhibition of responding to the previously rewarding stimulus and ultimately learning of the new rewarding stimulus association. Probabilistic reversal learning tasks, on the other hand, place an additional demand on the subject: he or she must maintain and update a mental representation of the task structure to differentiate between probabilistic losses (correct choices that are not rewarded) and true reversals. This uncertainty provides an ideal setting to study belief and delusions.[90]

Given that delusions are fixed beliefs, one might expect patients to perseverate on previously rewarding stimuli (i.e., fail to switch) after reversals. In reality, patients with psychosis show quite the opposite response. They exhibit increased switching and achieve fewer performance-dependent reversals, even when learning of initial stimulus–outcome associations appears intact.[91–95] This switching correlates with positive symptoms in both deterministic and probabilistic tasks.[93,96,97] In fact, patients with the highest positive symptom ratings exhibit the highest rates of switching, and they require fewer trials (i.e., less evidence) to alter responding after a reversal occurs.[96] It appears these patients may exhibit hyperflexibility, perhaps analogous to the incorporation of seemingly dissonant views in an updated version of one’s world model.[98] This hyperflexibility is observed too in the classic beads task or urn problem.[99,100] Patients with delusions tend to jump to conclusions[99,100] (although meta-analyses conflict as to the specificity of this effect to delusions[101,102]). But they also flip-flop back and forth in their beliefs after jumping. This may reflect an aversion to uncertainty.[103]

Intriguingly, recent analyses of reversal learning point to patient subgroups among both unmedicated, drug-naive patients (in the earliest stages of the illness) and medicated patients in the chronic phase.[91–95] Unlike healthy controls and patients who successfuly navigate the task, a subgroup with higher positive symptoms exhibited reduced prefrontal cortex activation and choice behavior inconsistent with a Hidden Markov Model (a computational estimation of beliefs about reversals and contingency stability updated trial-by-trial[93]). Patients who switch excessively also have more positive symptoms.[96] Furthermore, in a study of schizotypal individuals, those with more unusual experiences and beliefs were more likely to show behavioral switching.[104] While these findings are not immediately explicable in terms of a worsening reversal performance with illness chronicity, it is indeed the case that positive symptoms do improve with treatment and become a less salient feature of the clinical picture in more chronic patients, and in those whose positive symptoms are treatment resistant, the switching enhancement is sustained. Thus, there seems to be a relationship between delusions, hallucinations, and reversal-learning performance. We suggest that excessive switching could be similar to the flip-flopping behavior observed in the beads task and may, through reversal learning tasks, be accessible with animal work.

One model—the methylazoxymethanol acetate (MAM) model in young rats[105]—may be consistent with the patient subgroup reported by Schlagenhauf et al. These rats, exposed prenatally to MAM to induce abnormal prefontal cortex development and dopamine activity, exhibit marked hypofrontality and hyperflexible switching behavior in probabilistic learning tasks.[105] A recent study of reversal learning as Bayesian belief updating in rhesus monkeys[106] found that haloperidol (D_2 receptor antagonist) increased reliance on priors over new information. Taken together, the preclinical data recapitulate reversal-learning dysfunction and offer new insights into how we might intervene and optimally tune reversal learning in service of treating positive symptoms.

Delusion Contents

Nearly all delusions are socially relevant; they are ideas about oneself and one’s relationships with others. Their content is crucially related to the individual’s specific fears, needs, or security.[107] How can we model such complex mutations of human social behavior in creatures like laboratory rats? Simple associative experiments in relatively asocial animals (like rodents) can have implications for social learning in primates and in particular humans.[108] The same volatility-driven processes guide both human social learning and animal associative learning, in the same neural circuits. There may not be dedicated "social modules" but rather, a learning module that deals with hard inferences, social inferences being the hardest.[108] Thus, when errors encroach on our inferential mechanisms, social inferences are the most susceptible and therefore most frequently the concern of our aberrant beliefs.

Predictive Coding of Self and Other

In our theory, the brain models incoming data and minimizes PE.[109] However, it also actively samples those data, by performing actions on the world (e.g., moving through it).[110] By predicting (and ignoring) the sensory consequences of our actions, we also model ourselves as agents that exist. And, by identifying with the top layers of the hierarchy, the conscious experience of being that self emerges.[111] Passivity experiences—the sense that one's actions are under external control—may arise when the predictive modeling of one's actions fails and the active sampling of sensory data becomes noisy.[112]

Furthermore, Ketamine augments experience of the rubber hand illusion, the spurious sense of ownership of a prop-hand if the hand is stroked at the same time as one's own hand.[113] People on ketamine get the illusion more strongly and they experience it even in a control condition when the real and rubber hands are stroked asynchronously.[113] Patients with schizophrenia[114] and chronic ketamine abusers show the same excessive experience of the illusion in the synchronous and asynchronous conditions.[115] Surprisingly, mice have been reported to be susceptible to such an illusion.[116] Stroking a rubber tail in view of the mouse, at the same time the real tail is stroked elicits a threat response when the rubber tail is approached.[116] This paradigm could provide a key test of whether a manipulation in rodents recapitulates psychosis, if mice subject to the putative psychosis model (say ketamine) perceive the illusion in the asynchronous condition.[114]

We use our model of our self to make predictions about others.[63] To the extent that we do not share generative models with interlocutors, we will have social problems in reciprocal interactions. In psychosis, there may have been poor learning through development too (the oft noted neurological soft signs present in childhood home movies[117]), however, the impairment does not manifest until young adulthood, when the model needs to be used more extensively to make social predictions.

Delusions may represent attempts by the individual to garner some social capital[118]; the evolutionary biologist Ed Hagen has argued that, by knowing important information and trying to share it with others, people with delusional disorder may increase their sense of self-worth.[119] Social defeat is an animal paradigm in which defeat to a dominant conspecific engenders a sensitized dopamine system in experimental animals. This has been extrapolated to humans in an attempt to explain the increased propensity to psychosis in immigrants that has been noted in epidemiology studies.[120] The social defeat model is useful in illustrating the impact of social deficits on the predictive learning system and how that impact may set the scene for psychotic symptoms. Returning to our initial example, the social support that is present in the Chicago cult, and absent in many (but not all) patients—it seems that sharing your unusual beliefs with a social network of like-minded believers may render them less toxic (although *Folie a Deux* and *Folie a Familie* do complicate things[121]). Having our predictions go awry constantly is a very distressing, othering, experience. On the other hand, finding a group of people with whom one feels kinship and understanding can greatly reduce uncertainty. We assume that people will behave like they have in the past and that they are like us. It is hard for people to synchronize with and understand the intentions of someone who is unpredictable, hence patients are rarely successful in convincing others to share their ideas. When they do, the people who endorse the patients' beliefs are often close family members or friends, suggesting a propensity or susceptibility and perhaps more overlapping world models.[121] We believe insights into observational learning from relatively asocial animals like rodents might well furnish a deeper understanding of the complex phenomenology of delusions.

Summary and Conclusion

We have argued that a better understanding of delusions may be achieved by taking a reductionist approach to beliefs, conceiving of them as learned associations between representations that govern perception (both internal and external) and action. Central to the process of associative belief formation is PE, the mismatch between prior expectation and current circumstances. Depending on the precision (or inverse variance) of the PE relative to the prior belief, it may drive new learning (i.e., updating of the belief), or it may be disregarded. We have argued that this process of PE signaling and accommodation/assimilation may be awry in people with psychotic illnesses. In particular, we believe delusions form when PE is signaled inappropriately with high precision, such that it garners new and aberrant learning. We have described animal research that has furnished a mechanistic understanding of PE signaling in terms of underlying neurobiology; glutamatergic mechanisms underlie the specification of PE (NMDA receptors signal top-down expectancies, α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) the feedforward error signal), and, depending on the specific hierarchy, slower neuromodulators (like dopamine, acetylcholine, serotonin, noradrenaline, and oxytocin) signal precision of priors and PE. There are thus many routes through which PE can be aberrantly signaled and many heterogeneous consequences of aberrant PE. The inferences that are perturbed give rise to the specific contents of delusions (they are about other people and one's relationships to them, because these are the hardest inferences to make). We have described how such error correcting inferential mechanisms might give rise to the sense of bodily agency (the sense of being a self) and to a sense of reality more broadly. Disrupting these senses is profoundly distressing and results in psychosis. We made suggestions for how these processes could be examined in preclinical models. Some of these data have been gathered. We believe it is time to complete the patchwork; to gather data in rodent models and human patients on the complete suite of PE-driven associative learning tasks we have outlined (blocking, reversal learning, representation-mediated devaluation, and ownership illusions). Armed with an understanding of exactly how people with delusions fare on these tasks and exactly which neural mechanisms underpin them, we will be much better placed to determine the pathophysiology underpinning delusions and to tailor treatment approaches aimed at that pathophysiology.

Funding

Connecticut Mental Health Center (CMHC) and Connecticut State Department of Mental Health and Addiction Services (DMHAS).

NIH Medical Scientist Training Program Training Grant (T32GM007205) and NINDS T32NS007224 (to E.J.F.).

NARSAD Young Investigator Award (to S.M.G.).

NARSAD Distinguished Investigator Award (to J.R.T.).

IMHRO / Janssen Rising Star Translational Research Award and CTSA (UL1 TR000142) from the National Center for Research Resources (NCRR) and the National Center for Advancing Translational Science (NCATS), components of the National Institutes of Health (NIH), and NIH roadmap for Medical Research, R01MH067073, and the NCPTSD (to P.R.C.).

The contents of this work are solely the responsibility of the authors and do not necessarily represent the official view of NIH or the CMHC/DMHAS.

Conflict of Interest

The authors declare no conflicts of interest.

References

1. Box GE, Draper, N. *Empirical Model-Building and Response Surfaces*. New York, NY: John Wiley & Sons; 1987.
2. Marr D, Poggio, T. From understanding computation to understanding neural circuitry. *Neurosciences Res Prog Bull*. 1977;204:301–328.
3. Corlett PR, Fletcher PC. Computational psychiatry: a Rosetta Stone linking the brain to mental illness. *Lancet Psychiatry*. 2014;1:399–402.
4. Pavlov IP. *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*; 1927.
5. Skinner BF. *The Shaping of a Behaviorist: Part Two of an Autobiography*. Knopf Doubleday Publishing Group; 1979.
6. Bayes T. An essay towards solving a problem in the doctrine of chances. *Biometrika*. 1958;45:296–315.
7. Waelti P, Dickinson A, Schultz W. Dopamine responses comply with basic assumptions of formal learning theory. *Nature*. 2001;412:43–48.
8. Corlett PR, Taylor JR, Wang XJ, Fletcher PC, Krystal JH. Toward a neurobiology of delusions. *Prog Neurobiol*. 2010.
9. Festinger L, Riecken HW, Schachter S. *When Prophecy Fails*. Minneapolis, MN: University of Minnesota; 1956.
10. Festinger L. Cognitive dissonance. *Sci Am*. 1962;207:93–102.
11. McKay R, Langdon R, Coltheart M. "Sleights of mind": delusions, defences, and self-deception. *Cogn Neuropsychiatry*. 2005;10:305–326.
12. Garety P. Reasoning and delusions. *Br J Psychiatry Suppl*. 1991;14:14–18.
13. Garety PA. Making sense of delusions. *Psychiatry*. 1992;55:282–291.
14. Dickinson A, Shanks DR. Instrumental action and causal representation. In: Sperber D, Premack D, Premack AJ, eds. *Causal Cognition*. Oxford, UK: Clarendon Press; 1995:5–25.
15. Keller FS, Schoenfeld WN. *Principles of Psychology*. New York: Appleton-Century-Crofts; 1950.
16. Sutherland NS, Mackintosh, NJ. *Mechanisms of Animal Discrimination Learning*. New York: Academic Press; 1971.
17. Fodor JA. *The Language of Thought*. New York: Crowell; 1975.
18. Fodor JA. *The Mind Doesn't Work That Way*. Cambridge, MA: MIT; 2000.
19. Dickinson A. The 28th Bartlett Memorial Lecture. Causal learning: an associative analysis. *Q J Exp Psychol B Comp Physiol Psychol*. 2001;54:3–25.
20. Coltheart M, Menzies P, Sutton J. Abductive inference and delusional belief. *Cogn Neuropsychiatry*. 2010;15:261–287.
21. Corlett PR, Honey GD, Fletcher PC. From prediction error to psychosis: ketamine as a pharmacological model of delusions. *J Psychopharmacol*. 2007;21:238–252.
22. Gibbs AA, David AS. Delusion formation and insight in the context of affective disturbance. *Epidemiol Psichiatr Soc*. 2003;12:167–174.
23. Pomarol-Clotet E, Honey GD, Murray GK, et al. Psychological effects of ketamine in healthy volunteers. Phenomenological study. *Br J Psychiatry*. 2006;189:173–179.
24. Honsberger MJ, Taylor JR, Corlett PR. Memories reactivated under ketamine are subsequently stronger: a potential pre-clinical behavioral model of psychosis. *Schizophr Res*. 2015.
25. Rescorla RA, Wagner, AR. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. In: Black AH, Prokasy, WF, ed. *Classical Conditioning II: Current Research and Theory*. New York: Appleton-Century-Crofts; 1972:64–99.
26. Lavin A, Nogueira L, Lapish CC, Wightman RM, Phillips PE, Seamans JK. Mesocortical dopamine neurons operate in distinct temporal domains using multimodal signaling. *J Neurosci*. 2005;25:5013–5023.

27. Corlett PR, Krystal JH, Taylor JR, Fletcher PC. Why do delusions persist? *Front Hum Neurosci*. 2009;3:12.
28. Chadwick PK. Peer-professional first-person account: schizophrenia from the inside–phenomenology and the integration of causes and meanings. *Schizophr Bull*. 2007;33:166–173.
29. Green MJ, Phillips ML. Social threat perception and the evolution of paranoia. *Neurosci Biobehav Rev*. 2004;28:333–342.
30. Green DM, Swets, JA. *Signal Detection Theory and Psychophysics*. New York: Wiley; 1966.
31. Siegel SJ, Allan, LG, Hannah, SD, Crump, MJC. Applying signal detection theory to contingency assessment. *Comp Cogn Behav Rev.* 2009;4:116–134.
32. Warren HC. *A History of the Association Psychology*. New York: Charles Scribner's Sons; 1921.
33. Plato. *Phaedo*. Oxford, UK: Oxford University Press. 1999.
34. Aristotle, ed. *On Memory and Reminiscence*. Vol 3. Oxford: Clarendon Press; 1930.
35. Locke J. *An Essay Concerning Human Unerstanding*. London: Dent; 1690/1976.
36. Kamin L. Predictability, surprise, attention, and conditioning. In: Campbell BA, Church RM, eds. *Punishment and Aversive Behavior*. New York: Appleton-Century-Crofts; 1969.
37. McLaren IP, Dickinson A. The conditioning connection. *Phil Trans R Soc Lond B Biol Sci*. 1990;329:179–186.
38. Widrow B, Hoff ME Jr. Adaptive switching circuits. *IRE WESCON Convention Rec*. 1960:96–104.
39. O'Tuathaigh CM, Salum C, Young AM, Pickering AD, Joseph MH, Moran PM. The effect of amphetamine on Kamin blocking and overshadowing. *Behav Pharmacol*. 2003;14:315–322.
40. Yau JO, McNally GP. Pharmacogenetic excitation of dorsomedial prefrontal cortex restores fear prediction error. *J Neurosci*. 2015;35:74–83.
41. Steinberg EE, Keiflin R, Boivin JR, Witten IB, Deisseroth K, Janak PH. A causal link between prediction errors, dopamine neurons and learning. *Nat Neurosci*. 2013;16:966–973.
42. Moran PM, Al-Uzri MM, Watson J, Reveley MA. Reduced Kamin blocking in non paranoid schizophrenia: associations with schizotypy. *J Psychiatr Res*. 2003;37:155–163.
43. Corlett PR, Fletcher PC. The neurobiology of schizotypy: fronto-striatal prediction error signal correlates with delusion-like beliefs in healthy people. *Neuropsychologia*. 2012;50:3612–3620.
44. Kapur S. Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. *Am J Psychiatry*. 2003;160:13–23.
45. Robinson TE, Berridge KC. Incentive-sensitization and addiction. *Addiction*. 2001;96:103–114.
46. Gray JA, Feldon J, Rawlins JNP, Hemsley D, Smith AD. The neuropsychology of schizophrenia. *Behav Brain Sci*. 1991;14:1–84.
47. Miller R. Schizophrenic psychology, associative learning and the role of forebrain dopamine. *Med Hypotheses*. 1976;2:203–211.
48. Howes OD, Montgomery AJ, Asselin MC, et al. Elevated striatal dopamine function linked to prodromal signs of schizophrenia. *Arch Gen Psychiatry*. 2009;66:13–20.
49. Anticevic A, Corlett PR. Cognition-emotion dysinteraction in schizophrenia. *Front Psychol*. 2012;3:392.
50. Bortolotti L, Broome MR. Affective dimensions of the phenomenon of double bookkeeping in delusions. *Emot Rev*. 2012;4:187–191.
51. Shanks DR. Bayesian associative learning. *Trends Cogn Sci*. 2006;10:477–478.
52. Mesulam M. Representation, inference, and transcendent encoding in neurocognitive networks of the human brain. *Ann Neurol*. 2008;64:367–378.
53. Friston K. A theory of cortical responses. *Phil Trans R Soc Lond B Biol Sci*. 2005;360:815–836.
54. Friston K. The free-energy principle: a rough guide to the brain? *Trends Cogn Sci*. 2009;13:293–301.
55. Courville AC, Daw ND, Touretzky DS. Bayesian theories of conditioning in a changing world. *Trends Cogn Sci*. 2006;10:294–300.
56. Fiser J, Berkes P, Orbán G, Lengyel M. Statistically optimal perception and learning: from behavior to neural representations. *Trends Cogn Sci*. 2010;14:119–130.
57. Itti L, Baldi P. Bayesian surprise attracts human attention. *Vision Res*. 2009;49:1295–1306.
58. Rao RP, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci*. 1999;2:79–87.
59. McCollough C. Color adaptation of edge-detectors in the human visual system. *Science*. 1965;149:1115–1116.
60. Westbrook RF, Harrison W. Associative blocking of the McCollough effect. *Q J Exp Psychol A Hum Exp Psychol*. 1984;36:309–318.
61. Sloane ME, Ost JW, Etheriedge DB, Henderlite SE. Overprediction and blocking in the McCollough aftereffect. *Percept Psychophys*. 1989;45:110–120.
62. Brand JL, Holding DH, Jones PD. Conditioning and blocking of the McCollough effect. *Percept Psychophys*. 1987;41:313–317.
63. Friston K, Frith C. A duet for one. *Conscious Cogn*. 2015;36:390–405.
64. Adams RA, Stephan KE, Brown HR, Frith CD, Friston KJ. The computational anatomy of psychosis. *Front Psychiatry*. 2013;4:47.
65. Teufel C, Subramaniam N, Dobler V, et al. Shift toward prior knowledge confers a perceptual advantage in early psychosis and psychosis-prone healthy individuals. *Proc Natl Acad Sci USA*. 2015;112:13401–13406.
66. Dutton DG, Aron AP. Some evidence for heightened sexual attraction under conditions of high anxiety. *J Pers Soc Psychol*. 1974;30:510–517.
67. Allen M, Frank D, Schwarzkopf DS, et al. Unexpected arousal modulates the influence of sensory noise on confidence. *eLife*. 2016;5.
68. Marshall L, Mathys C, Ruge D, et al. Pharmacological fingerprints of contextual uncertainty. *PLoS Biol*. 2016;14:e1002575.
69. Bao S, Chan VT, Merzenich MM. Cortical remodelling induced by activity of ventral tegmental dopamine neurons. *Nature*. 2001;412:79–83.
70. Dennett D. Do animals have beliefs. In: Roitblat HL, Meyer JA, eds. *Comparative Approaches to Cognitive Science*. Cambridge, MA: MIT Press; 1995.
71. Davidson D. *Thought and Talk*. Oxford, UK: Clarendon Press; 1984.
72. Roskies A. Davidson on believers: can non-linguistic creatures have propositional attitudes. *Open MIND*. 2015;33.
73. Wynn K. Addition and subtraction by human infants. *Nature*. 1992;358:749–750.
74. Mitchell CJ, De Houwer J, Lovibond PF. The propositional nature of human associative learning. *Behav Brain Sci*. 2009;32:183–198.

75. Blaisdell AP, Sawa K, Leising KJ, Waldmann MR. Causal reasoning in rats. *Science*. 2006;311:1020–1022.

76. Kutlu MG, Schmajuk NA. Classical conditioning mechanisms can differentiate between seeing and doing in rats. *J Exp Psychol Anim Behav Process*. 2012;38:84–101.

77. Currie G. Imagination, delusion and hallucinations. In: Coltheart M, Davies M, eds. *Pathologies of Belief*. Oxford: Blackwell; 2000:167–182.

78. Konorski J. *Integrative Activity of the Brain: An Interdisciplinary Approach*. Chicago: University of Chicago Press; 1967.

79. McDannald M, Schoenbaum G. Toward a model of impaired reality testing in rats. *Schizophr Bull*. 2009;35:664–667.

80. Holland PC. Acquisition of representation mediated conditioned food aversions. *Learn Motiv.* 1981;12:1–18.

81. Lipska BK, Jaskiw GE, Weinberger DR. Postpubertal emergence of hyperresponsiveness to stress and to amphetamine after neonatal excitotoxic hippocampal damage: a potential animal model of schizophrenia. *Neuropsychopharmacology*. 1993;9:67–75.

82. Meng Y, Hu X, Bachevalier J, Zhang X. Decreased functional connectivity in dorsolateral prefrontal cortical networks in adult macaques with neonatal hippocampal lesions: relations to visual working memory deficits. *Neurobiol Learn Mem*. 2016;134:31–37.

83. McDannald MA, Whitt JP, Calhoon GG, et al. Impaired reality testing in an animal model of schizophrenia. *Biol Psychiatry*. 2011;70:1122–1126.

84. Bernstein DM, Laney C, Morris EK, Loftus EF. False beliefs about fattening foods can have healthy consequences. *Proc Natl Acad Sci USA*. 2005;102:13724–13731.

85. Sadacca BF, Jones JL, Schoenbaum G. Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *eLife*. 2016;5.

86. Sotty F, Damgaard T, Montezinho LP, et al. Antipsychotic-like effect of retigabine [N-(2-Amino-4-(fluorobenzylamino)-phenyl)carbamic acid ester], a KCNQ potassium channel opener, via modulation of mesolimbic dopaminergic neurotransmission. *J Pharmacol Exp Ther*. 2009;328:951–962.

87. Corlett PR, Cambridge V, Gardner JM, et al. Ketamine effects on memory reconsolidation favor a learning model of delusions. *PLoS One*. 2013;8:e65088.

88. Izquierdo A, Brigman JL, Radke AK, Rudebeck PH, Holmes A. The neural basis of reversal learning: an updated perspective. *Neuroscience*. 2016.

89. Jazbec S, Pantelis C, Robbins T, Weickert T, Weinberger DR, Goldberg TE. Intra-dimensional/extra-dimensional set-shifting performance in schizophrenia: impact of distractors. *Schizophr Res*. 2007;89:339–349.

90. Fletcher PC, Frith CD. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nat Rev Neurosci*. 2009;10:48–58.

91. Waltz JA, Gold JM. Probabilistic reversal learning impairments in schizophrenia: further evidence of orbitofrontal dysfunction. *Schizophr Res*. 2007;93:296–303.

92. Waltz JA, Kasanova Z, Ross TJ, et al. The roles of reward, default, and executive control networks in set-shifting impairments in schizophrenia. *PloS One*. 2013;8:e57257.

93. Schlagenhauf F, Huys QJ, Deserno L, et al. Striatal dysfunction during reversal learning in unmedicated schizophrenia patients. *Neuroimage*. 2014;89:171–180.

94. Culbreth AJ, Gold JM, Cools R, Barch DM. Impaired activation in cognitive control regions predicts reversal learning in schizophrenia. *Schizophr Bull*. 2016;42:484–493.

95. Reddy LF, Waltz JA, Green MF, Wynn JK, Horan WP. Probabilistic reversal learning in schizophrenia: stability of deficits and potential causal mechanisms. *Schizophr Bull*. 2016;42:942–951.

96. Waltz JA. The neural underpinnings of cognitive flexibility and their disruption in psychotic illness. *Neuroscience*. 2016.

97. Laws KR, Kondel TK, Clarke R, Nillo AM. Delusion-prone individuals: stuck in their ways? *Psychiatry Res*. 2011;186:219–224.

98. Moritz S, Woodward, T. Plausibility judgment in schizophrenic patients: evidence for a liberal acceptance bias. *German J Psychiatry.* 2004;7.

99. Garety PA, Hemsley DR, Wessely S. Reasoning in deluded schizophrenic and paranoid patients. Biases in performance on a probabilistic inference task. *J Nerv Ment Dis*. 1991;179:194–201.

100. Huq SF, Garety PA, Hemsley DR. Probabilistic judgements in deluded and non-deluded subjects. *Q J Exp Psychol A*. 1988;40:801–812.

101. Ross RM, McKay R, Coltheart M, Langdon R. Jumping to conclusions about the beads task? A meta-analysis of delusional ideation and data-gathering. *Schizophr Bull*. 2015;41:1183–1191.

102. Dudley R, Taylor P, Wickham S, Hutton P. Psychosis, delusions and the "jumping to conclusions" reasoning bias: a systematic review and meta-analysis. *Schizophr Bull*. 2016;42:652–665.

103. Moutoussis M, Bentall RP, El-Deredy W, Dayan P. Bayesian modelling of jumping-to-conclusions bias in delusional patients. *Cogn Neuropsychiatry*. 2011;16:422–447.

104. Bowman CH, Turnbull, OH. Schizotypy and flexible learning: a prerequisite for creativity. *Philoctetes*. 2.

105. Kaneko G, Sanganahalli BG, Groman SM, et al. Hypofrontality and posterior hyperactivity in early schizophrenia: imaging and behavior in a preclinical model. *Biol Psychiatry*. 2016.

106. Costa VD, Tran VL, Turchi J, Averbeck BB. Reversal learning and dopamine: a bayesian perspective. *J Neurosci*. 2015;35:2407–2416.

107. Reed GF. *The Psychology of Anomalous Experience: A Cognitive Approach*. London: Hutchinson; 1972.

108. Heyes C, Pearce JM. Not-so-social learning strategies. *Proc Biol Sci*. 2015;282.

109. Friston K, Kiebel S. Predictive coding under the free-energy principle. *Philos Trans R Soc Lond B Biol Sci*. 2009;364:1211–1221.

110. Friston KJ, Daunizeau J, Kilner J, Kiebel SJ. Action and behavior: a free-energy formulation. *Biol Cybern*. 2010;102:227–260.

111. Blanke O, Metzinger T. Full-body illusions and minimal phenomenal selfhood. *Trends Cogn Sci*. 2009;13:7–13.

112. Stephan KE, Friston KJ, Frith CD. Dysconnection in schizophrenia: from abnormal synaptic plasticity to failures of self-monitoring. *Schizophr Bull*. 2009;35:509–527.

113. Morgan HL, Turner DC, Corlett PR, et al. Exploring the impact of ketamine on the experience of illusory body ownership. *Biol Psychiatry*. 2011;69:35–41.

114. Peled A, Pressman A, Geva AB, Modai I. Somatosensory evoked potentials during a rubber-hand illusion in schizophrenia. *Schizophr Res*. 2003;64:157–163.

115. Tang J, Morgan HL, Liao Y, et al. Chronic administration of ketamine mimics the perturbed sense of body ownership associated with schizophrenia. *Psychopharmacology*. 2015;232:1515–1526.

116. Wada M, Takano K, Ora H, Ide M, Kansaku K. The rubber tail illusion as evidence of body ownership in mice. *J Neurosci*. 2016;36:11133–11137.

117. Walker EF, Savoie T, Davis D. Neuromotor precursors of schizophrenia. *Schizophr Bull*. 1994;20:441–451.

118. Fineberg SK, Corlett PR. The doxastic shear pin: delusions as errors of learning and memory. *Cogn Neuropsychiatry*. 2016;21:73–89.

119. Hagen E. Non-bizarre delusions as strategic deception. In: Elton S, O'Higgins P, eds. *Medicine and Evolution: Current Applications, Future Prospect*. New York: Taylor & Francis; 2008.

120. Selten JP, Cantor-Graae E. Hypothesis: social defeat is a risk factor for schizophrenia? *Br J Psychiatry Suppl*. 2007;51:S9–S12.

121. Nielssen O, Langdon R, Large M. Folie à deux homicide and the two-factor model of delusions. *Cogn Neuropsychiatry*. 2013;18:390–408.