Contents lists available at ScienceDirect

Schizophrenia Research

journal homepage: www.elsevier.com/locate/schres

A generative framework for the study of delusions

Tore Erdmann [a], Christoph Mathys [b,c,*]

[a] Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Bonomea 265, 34136 Trieste, Italy
[b] Interacting Minds Centre, Aarhus University, Jens Chr. Skous Vej 4, 8000 Aarhus C, Denmark
[c] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Wilfriedstrasse 6, 8032 Zurich, Switzerland

ARTICLE INFO

Article history:
Received 31 July 2020
Received in revised form 28 November 2020
Accepted 30 November 2020
Available online xxxx

Keywords:
Delusion
Dirichlet process
Hierarchical predictive coding
Auxiliary hypothesis
Epistemic trust

ABSTRACT

Despite the ubiquity of delusional information processing in psychopathology and everyday life, formal characterizations of such inferences are lacking. In this article, we propose a generative framework that entails a computational mechanism which, when implemented in a virtual agent and given new information, generates belief updates (i.e., inferences about the hidden causes of the information) that resemble those seen in individuals with delusions. We introduce a particular form of Dirichlet process mixture model with a sampling-based Bayesian inference algorithm. This procedure, depending on the setting of a single parameter, preferentially generates highly precise (i.e. over-fitting) explanations, which are compartmentalized and thus can co-exist despite being inconsistent with each other. Especially in ambiguous situations, this can provide the seed for delusional ideation. Further, we show by simulation how the excessive generation of such over-precise explanations leads to new information being integrated in a way that does not lead to a revision of established beliefs. In all configurations, whether delusional or not, the inference generated by our algorithm corresponds to Bayesian inference. Furthermore, the algorithm is fully compatible with hierarchical predictive coding. By virtue of these properties, the proposed model provides a basis for the empirical study and a step toward the characterization of the aberrant inferential processes underlying delusions.

1. Introduction

Delusions are implausible beliefs which are held with absolute conviction and cannot be changed by countervailing evidence (Jaspers, 1913; American Psychiatric Association, 2013). They are among the core symptoms of psychosis and a majority of individuals with schizophrenia experience delusional beliefs during the course of their illness (Harrow et al., 1995).

Despite the importance of delusions in psychiatric nosology and their debilitating effect on patients, their underlying mental and biological mechanisms are still poorly understood. In particular, a generative computational framework for the study of delusions is still lacking. Such a framework, situated in the context of *Computational Psychiatry* (Montague et al., 2012; Stephan and Mathys, 2014; Wang and Krystal, 2014; Mathys, 2016; Adams et al., 2016; Huys et al., 2016), would allow for the systematic testing of mechanistic hypotheses regarding regarding the emergence and maintenance of delusions. This framework should be *computational* in the sense that it conceptualizes delusions in terms of formal mathematical computations imputed to the mind. Beyond that, it should be *generative* in the sense that it allows for building models of minds which can be configured so that they generate delusional beliefs (where both *belief* and *delusional* are well-defined mathematically while also reflecting the clinical usage of these terms).

In this article we make an initial suggestion for such a generative computational framework. We introduce a model that combines three strands of thinking about mind-building and delusion formation. This model is based on Dirichlet process mixture models of concept learning (Tenenbaum and Griffiths, 2001), hierarchical predictive coding (Rao and Ballard, 1999; Friston, 2005a; Sterzer et al., 2018), and the use and abuse of auxiliary hypotheses in hypothesis testing and Bayesian inference (Duhem, 1906; Quine, 1951; Strevens, 2001; Jaynes, 2003; Gershman, 2019). Based on our suggested model, we simulate agents who update their beliefs in response to new information. We show that by manipulating the single decisive parameter of our model, we can generate belief patterns which can be characterized on a spectrum from delusional to appropriate, given the agent's input. We interpret the agent's behaviour in terms of previous conceptualizations of delusions, and we point out possible empirical ways to quantify our model's delusion-generating parameter in experimentally or naturally observed behaviour.

* Corresponding author at: Interacting Minds Centre, Aarhus University, Jens Chr. Skous Vej 4, 8000 Aarhus C, Denmark.
E-mail addresses: terdmann@sissa.it (T. Erdmann), chmathys@ethz.ch (C. Mathys).

https://doi.org/10.1016/j.schres.2020.11.048
0920-9964/

2. Theory

2.1. Delusions as a consequence of aberrant inference

Our approach builds on the three conceptual foundations mentioned above. Turning first to hierarchical predictive coding, the idea that *inferential* mechanisms support the formation and maintenance of delusions has led to an influential characterization in terms of deviations from Bayesian inference (Hemsley and Garety, 1986; Coltheart et al., 2010). Similarly, biases of probabilistic reasoning have been invoked to understand the process of delusion formation, such as limited data-gathering ("jumping to conclusions", see Speechley et al., 2010; Dudley et al., 2016) or a bias against disconfirmatory evidence (Woodward et al., 2006). Furthermore, a failure to think of alternative accounts of the delusion (a lack of belief flexibility) was found to be related to how strongly a delusion was held ("delusional conviction"; e.g., Freeman et al., 2004; Garety et al., 2005), and a number of recent reviews have underlined the importance of cognitive biases and delusional ideation (McLean et al., 2017; Broyd et al., 2017; Bronstein et al., 2019).

Predictive coding (PC) is a general account of brain function (Rao and Ballard, 1999; Friston, 2005b) which assumes that the brain infers the causes of its sensations using a hierarchical model of its environment. Applied to psychosis, the account emphasizes the balance between top-down predictions and bottom-up prediction error (PE) signals (Fletcher and Frith, 2009; Corlett et al., 2010, 2016; Sterzer et al., 2018). In this framework, prior beliefs are encoded in predictions about sensory inputs. Discrepancies between these predictions and the actual sensory stimulation lead to changes in beliefs whose magnitude depends on the precision of the predictions. Delusion formation then reflects a compensatory response to imbalances of the hierarchical inference scheme (Adams et al., 2013; Corlett et al., 2016; Fletcher and Frith, 2009). Specifically, delusions might result from the attempts to explain highly precise low-level PEs. The resulting explanations are epistemically inappropriate beliefs at higher levels in the processing hierarchy (Adams et al., 2013; Schmack et al., 2013).

2.2. Central and auxiliary hypotheses

A second foundation for our approach is the notion of "explaining-away". This phenomenon occurs in Bayesian belief networks and denotes the case, when, given two potential causes for an effect, the presence of one cause makes another less likely.

In Bayesian terms, the maintenance of delusions (and beliefs in general) is usually attributed to strong prior beliefs. However, inductive inferences critically depend on the beliefs about the structural dependencies between the relevant variables. For example, what one person takes to be evidence for a hypothesis, another person interprets as contradictory evidence. This can happen without contradicting the rules of logic because the direction of belief updating depends on other beliefs (Jern et al., 2014). A ubiquitous example of this phenomenon is the "explaining-away" of evidence. This describes the case in common-effect networks in which the presence of one cause in a common effect network makes another less likely. This implies that the interpretation of an observation depends on the ability of the observer to generate additional assumptions, called *auxiliary hypotheses*, which can "explain away" the evidence or even turn it into its contrary.

The idea goes back to Duhem's (1906) and Quine's (1951) insight that evidence from an experiment cannot refute a single scientific hypothesis, but only a conjunction of hypotheses (cf. Strevens, 2001; Jaynes, 2003). Gershman (2019) presented an analysis showing that in a Bayesian model, hypotheses with weaker prior probability can act as a "protective belt" and, in the face of dis-confirmatory evidence, take the blame instead of a central hypothesis (i.e., one with a stronger prior). This represents an effective strategy of belief preservation that depends on the creation of auxiliary hypotheses.

While these demonstrations of the explaining-away effect assume the existence of auxiliary hypotheses as given, the framework we introduce here allows for the generation of new auxiliary hypotheses which serve to explain observations that, under a different configuration, could have been explained by nuancing an existing explanation.

2.3. Dirichlet process mixture models

Human reasoning processes have a characteristic ability to deal with uncertainties due to incomplete or noisy information and build open-ended models of adaptive complexity. Much of this uncertainty is due to unobserved variables and the relation between these. When reasoning about a particular course of events, we compare hypotheses about the statistical structure of the world. A common problem is to detect when observations can be partitioned into separate groups, where each group is explained by a distinct cause. A solution to this are Dirichlet process mixture models (DPMMs) (Teh et al., 2006; Doshi-velez, 2009). These allow for inferring, for each data point, the group it most likely belongs to. A version of the Dirichlet process was independently proposed by Anderson (1991) for a theory of human category learning. Fig. 1 illustrates the behaviour of the model. Notably, it allows to model the classification into anomalies that require novel categories. The inference of a separate category has a strong influence on the subsequent belief updates, since data that belong to one category are assumed to be independent of all other categories. Crucially, the Dirichlet process prior assumes the existence of a potentially infinite number of groups and is this a model for open-ended learning, adapting to increasing amounts of data by increasing model complexity. This means that it provides a solution for the problem of *model-selection*, a best model is to be chosen in terms of accuracy and complexity. The Dirichlet process represents a suitable prior for such inferences and DPMMs are a Bayesian solutions to the problem of *structure learning* Gershman and Blei (2012). For this reason, DPMMs have found broad application in the modelling of higher-order human cognition (e.g., Kemp et al., 2010; Collins and Koechlin, 2012).

2.4. Model description

We harness the power of this approach in proposing a generic DPMM that describes delusion formation and maintenance. We do this in the context of a learner performing online inference about the latent structure of the environment based on a set of observed events. This constitutes a *structure learning* problem in statistics, and the learner is assumed to solve it (in a manner consistent with Bayesian inference) by iterating two steps. First, the learner has to partition the data into separate groups based on whether they are explainable by the same underlying cause. Second, given the grouping of the data, the learner can then infer a specific model for each group. We define the act of explaining an event or observation as inference of a single cause. Causes thus provide explanations for events. That is, they are models of the learner's environment (i.e., they define a probability distribution over current and future observations). The learner is equipped with a set of prior beliefs which are encoded in a hierarchical generative model for the events. Further, the learner has a set of existing models derived from prior experience of the world, which can be used to explain new observations. However, the existing explanations stand in competition with a mechanism for generating new explanations constructed from higher levels of the model, that is, from the prior over explanations. The structure of the prior belief of the learner allows for a potentially infinite number of causes. This means that, depending on their priors, learners can consider any new observation an anomaly, i.e. as belonging to a hitherto unobserved cause. A formal description of our model is given in the Appendix. We implement Bayesian inference for this model using Algorithm 8 from Neal (2000).

The assumption of an infinite collection of causes allows learners continually to discover new ones, building new theories, as they make

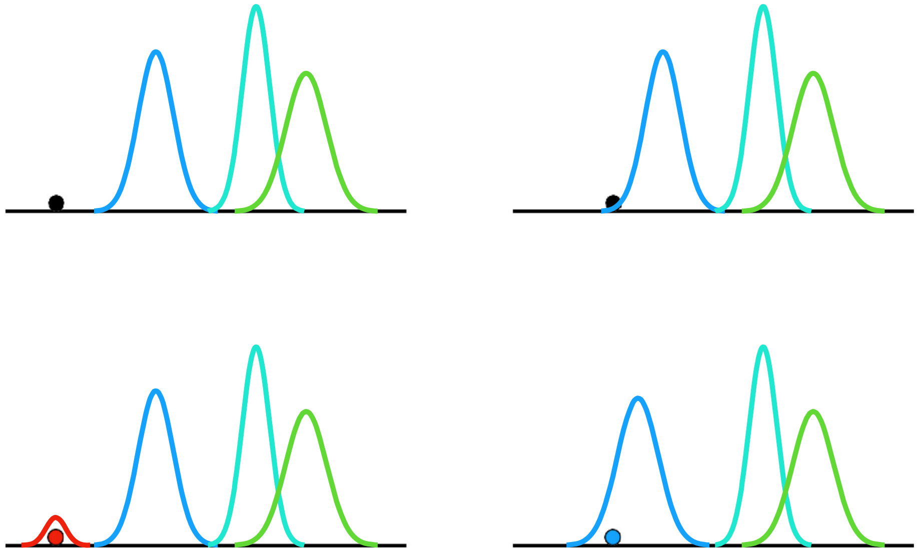

Fig. 1. Categorization and explanation in our framework (schematic). In the top panels, the same initial belief is depicted on the left and right, with separate explanations (causes) represented by Gaussians. On the left, the new observation (black dot) has a larger deviation from the existing causes than on the right. Here, the model infers a new cause and fits a corresponding cause to explain the observation (red Gaussian, bottom left). On the right, a less extreme observation is integrated into an existing cause (blue Gaussian, bottom right), which leads to a change in the structure of the corresponding explanation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

more and more observations. Still, at any point there is only a finite number of causes (at most one per individual observation) and the ease with which new causes are assumed is affected by priors and by the concentration parameter α. Low values of α favour a small number of causes that each account for many observations, while high values favour many small uniformly sized clusters of observations.

Inference about the underlying cause of an observation proceeds in two steps. In a first step, m potential explanations are drawn from the generative model M. For Gaussian models, the explanations correspond to parameter values (μ, τ), which are drawn from the prior. In a second step, these candidates are compared with the set of already known explanations in terms of their plausibility (i.e. likelihoods). The plausibility judgments are modulated by the respective prior probabilities. These are proportional to the number of previous observations accounted for by an existing explanation. The prior probability for previously unobserved causes depends only on the α parameter, which encodes a general expectation of new causes. The assignment to a cause is chosen according to these factors. The proposals for new causes drawn from the prior that were not selected are discarded after this step and new proposals are drawn for the next inference. Following the assignment of an observation to a cause, the next inference step is to integrate the information into the model associated with that cause. The specific form of this belief update depends on the form of the cause-specific models. After updating the separate hypotheses, the higher-level beliefs are updated. These may include hyper-priors over the parameters of the prior distribution for the cause-specific models and the belief about α. Intuitively, after inferring many new causes, the belief about α will change so that this becomes what is expected in the following. Iterating over these belief updates constitutes a Markov chain that leads to an approximation of the correct posterior belief (Neal, 2000).

3. Results

3.1. Simulation of the emergence of a delusion

As an illustration of our model's basic belief dynamics, we demonstrate an inference process that can be characterized as appropriate or delusional depending on the setting of a single parameter, the expected precision of explanations μ_τ. In what follows, we explain data $y \in \mathbb{R}$ based on simple Gaussian assumptions. That is, the cause-specific models are Gaussians characterized by mean and precision parameters $F(\mathrm{y}, \phi_\mathrm{k}) = \mathcal{N}(\mathrm{y}|\mu_\mathrm{k}, \tau_\mathrm{k}^{-1})$. The prior distributions for the cause-specific parameters μ_k and τ_k are independent normal $\left(\mathcal{N}\left(\mu_\mu, \tau_\mu\right)\right)$ and half-normal $(H\mathcal{N}(\mu_\tau, \tau_\tau))$, respectively. These priors influence the generation of candidates for new explanations. They also play a role in the process of updating the internal structure of existing explanations (through Bayes' rule, as in all Bayesian accounts of inference).

Of special interest is μ_τ, the *expected precision* of explanations. Under Gaussian assumptions, it is the mean of the prior on the precision parameter τ_k for explanation k. In other words, it specifies the prior belief about the expected inverse variance of observations under any of the currently held models. Generalizing beyond Gaussian assumptions, the expected precision can be cast as the negative entropy of explanations generated by the prior. In this view, high expected precision implies a prior criterion for generating explanations: it favours those explanations that, conditional on being true, assign a high likelihood value to observations.

Such strong priors about the expected precision lead to an “overfitting” of explanations, that is, generating hypotheses that over-accommodate the current data. This is related to a suggestion made in previous accounts of delusional thinking (Stone and Young, 1997; Mckay, 2012) that a bias toward “explanatory adequacy,” whereby the likelihood is over-weighted at the expense of the prior, plays a role in delusions. For example, Coltheart et al. (2010) develop their account with reference to Capgras' delusion, which involves the belief that a close friend or relative has been replaced by a physically identical impostor. Mckay (2012) explain Capgras' as arising from brain damage or disruption, which causes the face recognition system to become disconnected from the autonomic nervous system, generating anomalous data (Factor One). This disconnection occurs in conjunction with a bias toward explanatory adequacy (Factor Two), such that the affected individual updates beliefs as if ignoring the relevant prior probabilities of candidate hypotheses.

Our DPMM account provides a different perspective. The possibility to assign observations to different explanations allows for deviations from the ideal of a single coherent belief system. In this account, delusional belief updating results from an exaggerated preference for high-precision explanations. Observations are assigned to highly precise explanations, which, once generated, are evaluated only by their likelihood, which will be high by construction. In this manner, our framework allows for the co-existence of many high-precision explanations, which corresponds to a compartmentalization of an individual's worldview into many — possibly contradictory — models.

Fig. 2 illustrates this in the context of delusional misidentification as described in a case study of Capgras' delusion (Hirstein and Ramachandran, 1997). Instead of attributing small variations (whatever their origin) to randomness or coincidence, patient DS infers additional explanatory structure. Hirstein and Ramachandran (1997) proposed that Capgras' might be part of a more general memory management problem:

When you or I meet a new person, our brains open a new file, as it were, into which go all of our memories of interactions with this person. When DS meets a person who is genuinely new to him, his brain creates a file for this person and the associated experiences, as it should. But if the person leaves the room for 30 min and returns, DS's brain, instead of retrieving the old file and continuing to add to it, sometimes creates a completely new one. Why this should happen is unclear, but it may be that the limbic emotional activation from familiar faces is missing and the absence of this 'glow' is a signal for the brain to create a separate file for this face (or else the presence of the 'glow' is needed for developing links between successive episodes involving a person).

Here, instead of memory files, we suggest that observations are filed away in separate explanations. A delusion results because the expectation of high precision leads to over-precise explanations that do not generalize and therefore lead to large prediction errors in the face of additional data. At the same time, the compartmentalization of separate explanations prevents belief change and elaboration in spite of these large prediction errors since it prevents "joining the dots". These elements combined lead to the phenomenon of *aberrant salience* as proposed in predictive coding accounts of psychosis (Kapur, 2003). Our framework explains this aberrant (increased) salience as prediction errors resulting from overly precise explanations. The emergence of central delusional beliefs is all but inevitable under these circumstances: anything confirming an existing explanation will (simply by the mechanics of the Bayesian inference mechanism associated with our DPMM) increase this explanation's "pull", but not its reach, while anything contradicting it is explained away with high precision.

While our framework is silent on the content of the central beliefs that are likely to emerge, it allows for models where candidate explanations generated are predominantly self-related, derogatory, grandiose, etc. Specific models of this kind within the proposed framework will be the focus of future work.

3.2. Simulation of delusion maintenance

In order to show delusion maintenance, we again make Gaussian assumptions, but this time with an established central belief. We simulate two learners differing only in expected precision μ_τ, with identical initial belief and presented with identical observations. Fig. 3 shows the main result. Two belief systems differing only in their priors on μ_τ change in a radically different manner when presented with observations that are either integrated (low μ_τ) into the existing explanations (i.e. clusters), or mostly require new explanations (high μ_τ) to be accounted for. Observations are created by sampling from a uniform distribution and the initial belief is represented by a cluster (n_1=200) constituting an initial central hypothesis. After generation of 50 new observations, we compute the predicted labels for them. Next, we compute the posterior for the labels z_i and the cause-specific parameters $\phi_k = (\mu_k, \tau_k), k=1,...$ by running a Gibbs sampler for 10 iterations, which is sufficient for convergence of the (now updated) central hypothesis. In each iteration the labels are re-sampled according to their full-conditional probabilities and the cause-specific are parameters re-estimated accordingly. This corresponds to Algorithm 8 in Neal (2000).

Fig. 3 shows the change in the belief regarding the "central hypothesis". The bottom left panel shows the updated belief of an agent with a

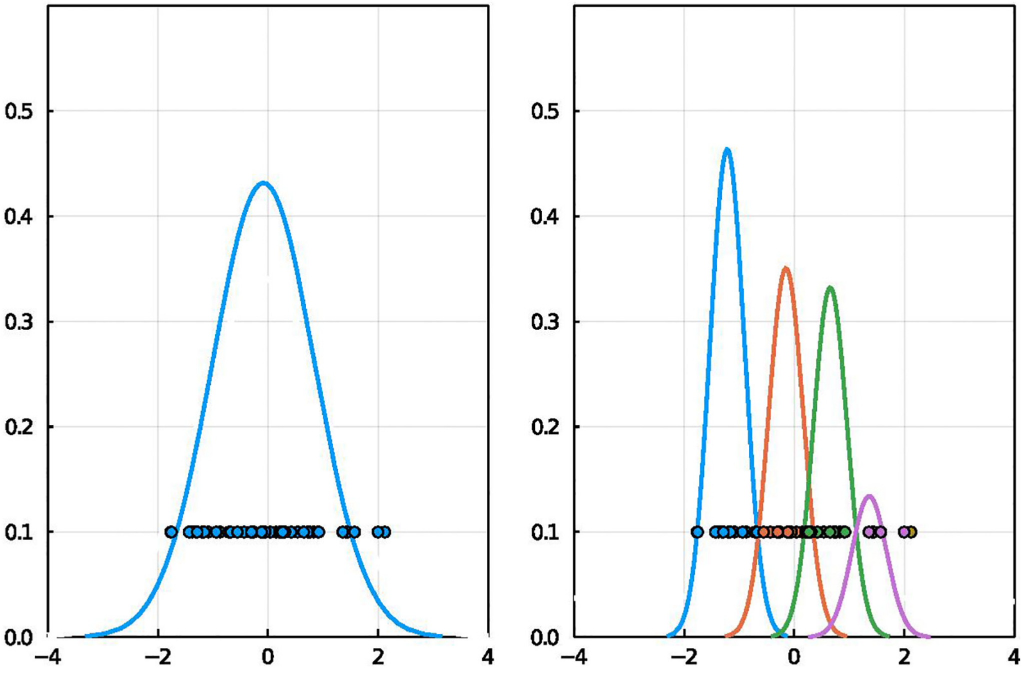

Fig. 2. A simulation of delusional mis-identification. In a case study, Hirstein and Ramachandran (1997) presented Capgras' patient DS with a sequence of photographs of the same model's face looking in different directions (here, we represent the photographs as points on a line; observations that are perceptually similar fall close on this abstract dimension). The left panel shows a simulation of inference in healthy observers: a single underlying cause ("the same person, photographed multiple times"; represented as a single Gaussian) is inferred. On the right, the inference observed in patient DS simulated ("different women who looked just like each other"; represented by multiple Gaussians). The two simulations from our model differed only in the expected precision (left: $\mu_\tau = \frac{1}{100}$, right: μ_τ=100). Inputs and all other parameters were equal.

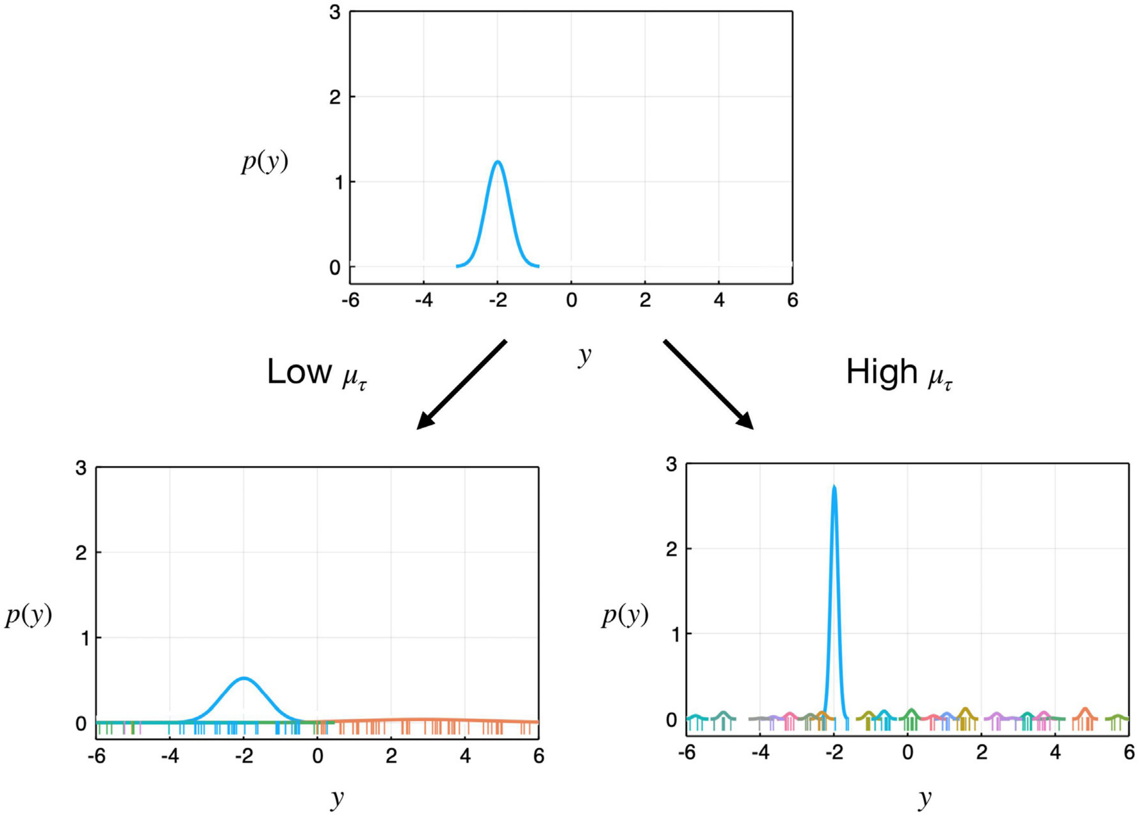

Fig. 3. Belief preserving evidence integration. Initial belief (upper row) and final belief (lower row) after inference given new observations. The difference in final beliefs is a function of the expected precision μ_τ alone. All other settings and inputs are the same. Bottom left: $\mu_\tau \sim HN(100, 10)$. The existing explanation (blue Gaussian) is elaborated (i.e., broadened) in response to new observations, which are to a considerable extent integrated into the already existing, but now elaborated, model. Bottom right: $\mu_\tau \sim HN(1/100, 10)$. The existing explanation is narrowed, but its dominance remains unaffected. New observations which do not fit it exactly are explained away (i.e., assigned to their own little ad hoc explanations). While both of these ways of processing the same information correspond to Bayesian inference (albeit under different values for μ_τ), the inference process on the right can be characterized as delusional. Further details and the code for reproducing this simulation can be found here: https://tinyurl.com/y3m79qdw. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

relatively low value of μ_τ, i.e. a value encoding the expectation of rather imprecise observations, corresponding to wide cause distributions. For this learner, the updated belief given the presented observations is more imprecise. In other words, it has become capable of integrating observations that where somewhat outside its initial distribution, leading to a widening of the density. This can be seen as signalling a reduction of certainty regarding the initial explanation for the observations. The right column shows the updated belief of an agent with a relatively high value of the expected precision parameter μ_τ. Given this prior, the agent ends up with a belief that is not changed much in terms of "content" (i.e. the expected observations under the model k, namely μ_k) and is more precise than before. Inference with such a prior exhibits a confirmatory arbitration of evidence which leads to the reinforcement of current beliefs. Even slight deviations are treated as outliers so as to maintain the parameters and meaning of the central hypothesis. Note the simple Gaussians we used here serve to make a general point. It is in principle straightforward to replace them with more complex Bayesian networks representing nontrivial causal structures.

Under conditions of delusional belief updating (i.e., aberrant μ_τ), the separation of explanatory categories prevents making connections between observations that challenge current beliefs and which could lead to very different beliefs altogether. Applying the simulation in 3 to the example by Coltheart et al. (2010) (p. 279), we may take the input to represent the various observations of their Capgras' patient:

For example, the subject might learn that trusted friends and family believe the person is his wife, that this person wears a wedding ring that has his wife's initials engraved in it, that this person knows things about the subject's past life that only his wife could know, and so on.

Each of these observations would normally lead to a change in the central belief. However, the generation of ad-hoc explanations as in our simulation could explain how the subject maintains the impostor belief.

4. Discussion

We have introduced a framework allowing for the description and generative construction of delusional inference. This is based on approximate Bayesian inference using Dirichlet process mixture models applied to structure learning problems. We have shown how an optimal inference algorithm can, endowed with particular higher-order beliefs, exhibit behaviour resembling delusional inference. Importantly, the outcome of the inference process was influenced by the prior beliefs about the expected precision of explanations. A strong belief in precise observations leads to the plentiful generation of over-fitting explanations, some of which are bound to coincide with an observation, leading to their acceptance over an a priori more plausible explanation.

4.1. Relation to previous work

Hierarchical predictive coding is one of the most promising computational frameworks for the description of delusions, and a misalignment in the hierarchical signalling of precision has often been invoked as the underlying reason for the emergence of delusions (Corlett et al., 2007, 2009; Fletcher and Frith, 2009; Sterzer et al., 2018). Our framework is fully consistent with these ideas. Indeed, it is exactly (not to say precisely…) an exaggerated expected precision μ_τ which is sufficient to explain the formation and maintenance of delusional information processing. However, the approach we introduce goes beyond previous predictive coding accounts of delusions in that it comes with a fully specified generative algorithm. Furthermore, the large prediction errors entailed by an over-fitting structure learning process provide the basis for the phenomenon of aberrant salience, which in our framework can explain the emergence of central beliefs with high “pull” surrounded by ad-hoc explanations shielding them from elaboration.

Our model builds on and extends *latent cause models* in reinforcement learning (Courville et al., 2006; Redish and Johnson, 2007). Gershman et al. (2010) showed how state classification can be derived as rational inference in a Dirichlet process mixture model. While these authors focus on the role of the concentration parameter α, we investigate the role of prior beliefs on the inference of new causes and belief change. Another important difference is that in their model, inputs consist of features which include the context that needs to be inferred, while in our model the agent receives no additional cue about context but has to infer this from the observations alone. Furthermore, our model has an additional hierarchical layer which allows for varying prior beliefs about the precision of observations.

4.2. Single-factor versus dual-factor explanations of delusions

There is a debate about whether delusions can be explained by a single factor or whether there need to be at least two. Hierarchical predictive coding is the classic example of a single-factor framework (Fletcher and Frith, 2009), while two factors are required according to Coltheart et al. (2010). Our model speaks to this question in that it provides a generative process where changing a single parameter is enough to get from appropriate to delusional thinking. While this indicates that one-factor explanations of delusion formation and maintenance are possible, the framework does not preclude the presence of additional factors. For example, the process of hypothesis generation could be disordered in addition to the expected precision μ_τ. Furthermore, the framework allows for quantitative comparisons of single-factor and *k*-factor hypotheses.

Our framework takes the perspective that belief states are never per se delusional, but rather *the way information is processed* can be delusional. From this perspective, it is the combination of the largely immutable central belief and the disconnected auxiliary hypotheses proliferating around it which together constitute the delusion. The delusionality does not lie in any one belief but in the way a belief (i.e., a model of the world) is prevented from being deepened and broadened. Instead, all the information that could drive such a deepening and broadening is explained away. While the models in our simulations were simply clusters of observations explained by Gaussians, Dirichlet process mixture models are not restricted to such simple examples. In principle, such Gaussian clusters can be replaced with elaborate causal models as in Tenenbaum et al. (2011). From the perspective of our framework, delusions are initially adequate causal models in need of elaboration. They are formed by arresting the development of a particular causal model and are maintained by the same mechanism — keeping the model insulated from new evidence.

4.3. Limitations and extensions

Our model does not by itself speak to the question how maladaptive expected precision μ_τ could evolve developmentally. However, it fits closely with the concept of *epistemic trust*. This is “an individual's willingness to consider new knowledge from another person as trustworthy, generalizable, and relevant to the self” (Fonagy and Allison, 2014) and is of great clinical importance in the conceptualization and treatment of borderline personality disorder. Our framework allows us to interpret μ_τ as an inverse quantification of epistemic trust (i.e., as a quantification of epistemic mistrust): low μ_τ leads to the integration of new information and to a corresponding broadening and enrichment of existing models of the world, while high μ_τ leads new information to be explained away when it doesn't fit an existing model exactly, accompanied by a narrowing of explanations. This provides a mechanistic computational account of epistemic (mis)trust, and it will be interesting to study the relation between empirical measures of expected precision μ_τ and epistemic trust in future work.

An important limitation is that we have not estimated μ_τ from observed behaviour. Not least, this is due to the difficulty of devising behavioural experiments where participants are given scope to behave in a sufficiently open-ended manner for ecologically valid forms of delusional behaviour to emerge while still keeping to a controlled experimental setting. For the study of delusional belief dynamics, popular experiments in computational psychiatry such as reversal learning tasks (Schlagenhauf et al., 2014; Waltz, 2017) or the beads task (Adams et al., 2018; Baker et al., 2019) are too restricted in the range of behaviour they allow. We therefore face the challenge of coming up with tasks that enable us to apply our framework to experimental data.

Examples of applications of DPMMs to experimental data are Collins and Koechlin (2012) and Donoso et al. (2014), where the authors model inferential computations underlying reasoning processes in the prefrontal cortex (PFC). Specifically, they showed that the PFC is involved in the monitoring of the reliability of the current and a number of counterfactual behavioural strategies in a learning paradigm. While in their tasks the reasoning processes were about behavioural strategies, similar *metacognitive* processes may be used in the inferential domain, for example in model selection. In this domain, it is challenging to infer metacognitive processes from behavioural data because the mapping from reasoning to actions is hard to constrain adequately — not too simple (e.g., tasks involving binary choices, not requiring higher-order reasoning) and not too open-ended (defying formal analysis and modelling). It is therefore important to ground the design of such tasks in formal accounts such as the one we propose here. Furthermore, functional imaging combined with formal modelling can reveal differences in inference processes that may not be expressed in directly observable behaviour. Taken together, behavioural tasks calibrated for meta-inference, neuroimaging, and hierarchical modelling frameworks like the one proposed here hold promise for the understanding of delusions, which play out mostly within the unobservable realm of thought and only rarely relate to behaviours in predictable ways.

5. Conclusion

Our proposed framework is an initial attempt at a formal conceptualization of delusional thinking. While previous computational descriptions stopped short of proposing a fully generative process, our framework provides this. It covers the spectrum from delusional to appropriate treatment of new information with adjustments to only a single parameter, and it can describe the emergence and maintenance of a delusion as a one-factor process. Furthermore, our framework is consistent with Bayesian inference and hierarchical predictive coding. While this is only a first step which without doubt will be improved upon and empirical applications are still missing, it sets a benchmark by combining the properties just mentioned: generativity, simplicity, single-factor sufficiency, and consistency with Bayesian inference.

Declaration of competing interest

The authors declare no conflicts of interest.

Acknowledgements

This work was supported by AUFF Starting Grant no. AUFF-E-2019-7-10 (CM).

Contributors

Both authors developed the theory and wrote the manuscript. Tore Erdmann wrote the software and created the figures.

Role of funding sources

Funding sources had no influence in study design; nor in the collection, analysis and interpretation of data; nor in the writing of the report; nor in the decision to submit the article for publication.

Appendix A*Details of model and inference algorithm*

Formally, our model performs inference for a mixture model with a Dirichlet process (DP) prior. We assume a data set $y = (y_1,\ldots,y_n)$ and a corresponding set of latent labels $z = (z_1,\ldots,z_n)$. The generative model can be written as follows:

$$\phi_k \sim G_0 \tag{1}$$

$$(z_1,\ldots z_n) \sim CRP(\alpha) \tag{2}$$

$$y_i \sim F\left(y_i, \phi_{z_i}\right), i = 1,\ldots,n \tag{3}$$

CRP denotes the *Chinese restaurant process*, a particular representation of the DP that provides a probability distribution over the space of data partitions. For the choices we make in our simulation, this becomes

$$\mu_k \mid \mu_\mu, \tau_\mu \sim N\left(\mu_\mu, \tau_\mu\right) \tag{4}$$

$$\tau_k \mid \mu_\tau, \tau_\tau \sim HN(\mu_\tau, \tau_\tau) \tag{5}$$

$$(z_1,\ldots,z_n) \sim CRP(\alpha) \tag{6}$$

$$y_i \sim N\left(\mu_{z_i}, \tau_{z_i}\right), \quad i = 1,\ldots,n. \tag{7}$$

Based on the partition structure in the generative model we can write the joint probability as

$$p(y,z,\phi). \quad = p(z|\alpha) \prod_{k\in 1,\ldots,K} \left(\prod_{i:z_i=k} p_N(y_i|\mu_k,\tau_k) p_N\left(\mu_k|\mu_\mu,\tau_\mu\right) p_{HN}(\tau_k|\mu_\tau,\tau_\tau) \right), \tag{8}$$

where $p_G(y|\theta)$ denotes the density of distribution $G(\theta)$ evaluated at y. Due to exchangeability of the DP, we can compute the full-conditional distributions by assuming the current observation has index n, where the full-conditional has a simple form that we use to perform Gibbs sampling:

$$P\left(z_n = k \mid y_n, \{(\mu_k,\tau_k)\}_{k=1}^{K+m}, \{n_k\}_{k=1}^{K}, \alpha, m\right) = p(z_n = k \mid z_1,\ldots,z_{t-1}) \cdot p_{\mathcal{N}}(y_i \mid \mu_k,\tau_k) \tag{9}$$

with the prior probability for that assignment, $p(z_n = k|z_1,\ldots,z_{n\text{-}1})$, given by

$$\frac{n_k}{n-1+\alpha}, \quad \text{if } k \text{ is an existing cause, i.e. } k \le K \tag{10}$$

$$\frac{\alpha/m}{n-1+\alpha}, \quad \text{if } k \text{ is a new cause, i.e. } K < k < K+m \tag{11}$$

and temporary candidate parameters for the m new components drawn their respective priors μ_k~$N(\mu_\mu,\tau_\mu)$ and τ_k~$HN(\mu_\tau,\tau_\tau)$, $k = K < k < K + m$.

The parameters $\{z_1,\ldots,z_n,\phi_1,\ldots,\phi_K\}$ represent the state of a Markov chain that is iteratively updated and can be used to estimate functions of the posterior over the parameters. Specifically, we iterate draws from the full-conditionals of the z and the cluster parameters ϕ according to Algorithm 8 in Neal (2000).

Simulation details

For the simulations for Fig. 3, we first initialize a single the cluster with an initial dataset $D_{init} = \{(y_i, z_i)\}_{i=1}^{200}$. This means computing the posterior for cluster k given all data with $z_i = k$. We simulated Random-Walk-Metropolis-Hastings single chains to obtain J=1000 samples from the posterior $\phi_j^* \sim \pi(\mu_k, \tau_k \mid \mu_\mu, \tau_\mu, \mu_\tau, \tau_\tau)$ and setting $\phi_k = \frac{1}{J}\sum_j^J \phi_j^*$.

Given this initial belief state (a mixture with a single cluster), which was kept identical for the simulations with different priors, we perform Bayesian inference using Markov chain Monte Carlo sampling according to Algorithm 8 in Neal (2000). Specifically, we scan through new batch of data $D_{new} = \{y_i^*\}_{i=1}^{50}$ and sample the labels initial values for the z_i^*, i= 1,…, 50 according to the predictive probabilities. For each change in the partition implied by the z_i, we update the affected cluster parameters by performing 10 MCMC steps toward the posterior (as described for the initialization), starting from an initialization at the previous estimate. After the initialization pass, we perform additional iterations where we iterate 20 times over all observations, both D_{init} and D_{new} and resample the cluster labels according to the algorithm detailed above. The simulation was performed with the following hyperparameter settings: μ_μ=0.0, τ_μ=1/10, $\tau_\tau = 10$ and with the prior only differing for $HN(\mu_\tau^{(j)},\tau_\tau)$, where, $\mu_\tau^{(1)}$=1/100 and $\mu_\tau^{(2)}$=100 for the two models. The simulation was implemented in Julia (https://julialang.org) and our code is freely available at: https://tinyurl.com/y3m79qdw.

References

Adams, R.A., Stephan, K.E., Brown, H.R., Frith, C.D., Friston, K.J., 2013. The computational anatomy of psychosis. Front. Psych. 4. https://doi.org/10.3389/fpsyt.2013.00047.

Adams, R.A., Huys, Q.J.M., Roiser, J.P., 2016. Computational psychiatry: towards a mathematically informed understanding of mental illness. J. Neurol. Neurosurg. Psychiatry 87, 53–63. https://doi.org/10.1136/jnnp-2015-310737.

Adams, R.A., Napier, G., Roiser, J.P., Mathys, C., Gilleen, J., 2018. Attractor-like dynamics in belief updating in schizophrenia. J. Neurosci., 3163–17 https://doi.org/10.1523/JNEUROSCI.3163-17.2018.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). American Psychiatric Pub.

Anderson, J.R., 1991. The adaptive nature of human categorization. Psychol. Rev. 98, 409.

Baker, S.C., Konova, A.B., Daw, N.D., Horga, G., 2019. A distinct inferential mechanism for delusions in schizophrenia. Brain 142, 1797–1812. https://doi.org/10.1093/brain/awz051.

Bronstein, M.V., Pennycook, G., Joormann, J., Corlett, P.R., Cannon, T.D., 2019. Dual-process theory, conflict processing, and delusional belief. Clin. Psychol. Rev. 72, 101748. https://doi.org/10.1016/j.cpr.2019.101748.

Broyd, A., Balzan, R.P., Woodward, T.S., Allen, P., 2017. Dopamine, cognitive biases and assessment of certainty: a neurocognitive model of delusions. Clin. Psychol. Rev. 54, 96–106. https://doi.org/10.1016/j.cpr.2017.04.006.

Collins, A., Koechlin, E., 2012. Reasoning, learning, and creativity: frontal lobe function and human decision-making. PLoS Biol. 10, e1001293. https://doi.org/10.1371/journal.pbio.1001293.

Coltheart, M., Menzies, P., Sutton, J., 2010. Abductive inference and delusional belief. Cogn. Neuropsychiat. 15, 261–287. https://doi.org/10.1080/13546800903439120.

Corlett, P., Honey, G., Fletcher, P., 2007. From prediction error to psychosis: ketamine as a pharmacological model of delusions. J. Psychopharmacol. 21, 238–252. https://doi.org/10.1177/0269881107077716.

Corlett, P.R., Krystal, J.H., Taylor, J.R., Fletcher, P.C., 2009. Why do delusions persist? Front. Hum. Neurosci. 3. https://doi.org/10.3389/neuro.09.012.2009.

Corlett, P., Taylor, J., Wang, X.J., Fletcher, P., Krystal, J., 2010. Toward a neurobiology of delusions. Prog. Neurobiol. 92, 345–369. https://doi.org/10.1016/j.pneurobio.2010.06.007.

Corlett, P.R., Honey, G.D., Fletcher, P.C., 2016. Prediction error, ketamine and psychosis: an updated model. J. Psychopharmacol. (Oxford, Engl.) 30, 1145–1155. https://doi.org/10.1177/0269881116650087.

Courville, A.C., Daw, N.D., Touretzky, D.S., 2006. Bayesian theories of conditioning in a changing world. Trends Cogn. Sci. 10, 294–300. https://doi.org/10.1016/j.tics.2006.05.004.

Donoso, M., Collins, A.G.E., Koechlin, E., 2014. Foundations of human reasoning in the prefrontal cortex. Science 344, 1481–1486. https://doi.org/10.1126/science.1252254.

Doshi-velez, F., 2009. The infinite partially observable Markov decision process. In: Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 22. Curran Associates, Inc., pp. 477–485.

Dudley, R., Taylor, P., Wickham, S., Hutton, P., 2016. Psychosis, delusions and the “jumping to conclusions” reasoning Bias: a systematic review and meta-analysis. Schizophr. Bull. 42, 652–665. https://doi.org/10.1093/schbul/sbv150.

Duhem, P.M.M., 1906. La théorie physique: son objet, et sa structure. Chevalier & Rivière.

Fletcher, P.C., Frith, C.D., 2009. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. Nat. Rev. Neurosci. 10, 48. https://doi.org/10.1038/nrn2536.

Fonagy, P., Allison, E., 2014. The role of mentalizing and epistemic trust in the therapeutic relationship. Psychotherapy 51, 372–380. https://doi.org/10.1037/a0036505.

Freeman, D., Garety, P.A., Fowler, D., Kuipers, E., Bebbington, P.E., Dunn, G., 2004. Why do people with delusions fail to choose more realistic explanations for their experiences? An empirical investigation. J. Consult. Clin. Psychol. 72, 671–680. https://doi.org/10.1037/0022-006X.72.4.671.

Friston, K., 2005a. A theory of cortical responses. Phil. Trans. R. Soc. B: Biol. Sci. 360, 815–836. https://doi.org/10.1098/rstb.2005.1622.

Friston, K., 2005b. A theory of cortical responses. Phil. Trans. R. Soc. B: Biol. Sci. 360, 815–836. https://doi.org/10.1098/rstb.2005.1622.

Garety, P.A., Freeman, D., Jolley, S., Dunn, G., Bebbington, P.E., Fowler, D.G., Kuipers, E., Dudley, R., 2005. Reasoning, emotions, and delusional conviction in psychosis. J. Abnorm. Psychol. 114, 373–384. https://doi.org/10.1037/0021-843X.114.3.373.

Gershman, S.J., 2019. How to never be wrong. Psychon. Bull. Rev. 26, 13–28. https://doi.org/10.3758/s13423-018-1488-8.

Gershman, S.J., Blei, D.M., 2012. A tutorial on Bayesian nonparametric models. J. Math. Psychol. 56, 1–12. https://doi.org/10.1016/j.jmp.2011.08.004.

Gershman, S.J., Blei, D.M., Niv, Y., 2010. Context, learning, and extinction. Psychol. Rev. 117, 197–209. https://doi.org/10.1037/a0017808.

Harrow, M., MacDonald, A.W., Sands, J.R., Silverstein, M.L., 1995. Vulnerability to delusions over time in schizophrenia and affective disorders. Schizophr. Bull. 21, 95–109. https://doi.org/10.1093/schbul/21.1.95.

Hemsley, D.R., Garety, P.A., 1986. The formation of maintenance of delusions: a Bayesian analysis. Br. J. Psychiatry 149, 51–56. https://doi.org/10.1192/bjp.149.1.51.

Hirstein, W., Ramachandran, V.S., 1997. Capgras syndrome: a novel probe for understanding the neural representation of the identity and familiarity of persons. Proc. R. Soc. B Biol. Sci. 264, 437–444.

Huys, Q.J.M., Maia, T.V., Frank, M.J., 2016. Computational psychiatry as a bridge from neuroscience to clinical applications. Nat. Neurosci. 19, 404–413. https://doi.org/10.1038/nn.4238.

Jaspers, K., 1913. Allgemeine Psychopathologie für Studierende, Ärzte und Psychologen. Springer-Verlag.

Jaynes, E.T., 2003. Probability Theory: The Logic of Science. Cambridge University Press.

Jern, A., Chang, K.m.K., Kemp, C., 2014. Belief polarization is not always irrational. Psychol. Rev. 121, 206–224. https://doi.org/10.1037/a0035941.

Kapur, S., 2003. Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. Am. J. Psychiatr. 160, 13–23. https://doi.org/10.1176/appi.ajp.160.1.13.

Kemp, C., Tenenbaum, J.B., Niyogi, S., Griffiths, T.L., 2010. A probabilistic model of theory formation. Cognition 114, 165–196. https://doi.org/10.1016/j.cognition.2009.09.003.

Mathys, C., 2016. How could we get nosology from computation? In: Redish, A.D., Gordon, J.A., Lupp, J. (Eds.), Computational Psychiatry: New Perspectives on Mental Illness. Volume 20 of Strüngmann Forum Reports. MIT Press, Cambridge, MA, pp. 121–135

Mckay, R., 2012. Delusional inference. Mind Lang. 27, 330–355. https://doi.org/10.1111/j.1468-0017.2012.01447.x.

McLean, B.F., Mattiske, J.K., Balzan, R.P., 2017. Association of the Jumping to conclusions and evidence integration biases with delusions in psychosis: a detailed meta-analysis. Schizophr. Bull. 43, 344–354. https://doi.org/10.1093/schbul/sbw056.

Montague, P.R., Dolan, R.J., Friston, K.J., Dayan, P., 2012. Computational psychiatry. Trends Cogn. Sci. 16, 72–80. https://doi.org/10.1016/j.tics.2011.11.018.

Neal, R.M., 2000. Markov chain sampling methods for dirichlet process mixture models. J. Comput. Graph. Stat. 9 (2), 249–265.

Quine, W.V., 1951. Two dogmas of empiricism. Philos. Rev. 60, 20–43. https://doi.org/10.2307/2181906.

Rao, R.P.N., Ballard, D.H., 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. 2, 79. https://doi.org/10.1038/4580.

Redish, A.D., Johnson, A., 2007. A computational model of craving and obsession. Ann. N. Y. Acad. Sci. 1104, 324–339. https://doi.org/10.1196/annals.1390.014.

Schlagenhauf, F., Huys, Q.J.M., Deserno, L., Rapp, M.A., Beck, A., Heinze, H.J., Dolan, R., Heinz, A., 2014. Striatal dysfunction during reversal learning in unmedicated schizophrenia patients. NeuroImage 89, 171–180. https://doi.org/10.1016/j.neuroimage.2013.11.034.

Schmack, K., de Castro, A.G.C., Rothkirch, M., Sekutowicz, M., Rössler, H., Haynes, J.D., Heinz, A., Petrovic, P., Sterzer, P., 2013. Delusions and the role of beliefs in perceptual inference. J. Neurosci. 33, 13701–13712. https://doi.org/10.1523/JNEUROSCI.1778-13.2013.

Speechley, W.J., Whitman, J.C., Woodward, T.S., 2010. The contribution of hypersalience to the "jumping to conclusions" bias associated with delusions in schizophrenia. J. Psychiatry Neurosci. 35, 7–17. https://doi.org/10.1503/jpn.090025.

Stephan, K.E., Mathys, C., 2014. Computational approaches to psychiatry. Curr. Opin. Neurobiol. 25, 85–92. https://doi.org/10.1016/j.conb.2013.12.007.

Sterzer, P., Adams, R.A., Fletcher, P., Frith, C., Lawrie, S.M., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M., Corlett, P.R., 2018. The predictive coding account of psychosis. Biol. Psychiatry 84, 634–643. https://doi.org/10.1016/j.biopsych.2018.05.015.

Stone, T., Young, A.W., 1997. Delusions and brain injury: the philosophy and psychology of belief. Mind Lang. 12, 327–364. https://doi.org/10.1111/j.1468-0017.1997.tb00077.x.

Strevens, M., 2001. The Bayesian treatment of auxiliary hypotheses. Br. J. Philos. Sci. 52, 515–537. https://doi.org/10.1093/bjps/52.3.515.

Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M., 2006. Hierarchical Dirichlet processes. J. Am. Stat. Assoc. 101, 1566–1581. https://doi.org/10.1198/016214506000000302.

Tenenbaum, J.B., Griffiths, T.L., 2001. Generalization, similarity, and Bayesian inference. Behav. Brain Sci. 24, 629–640. https://doi.org/10.1017/S0140525X01000061.

Tenenbaum, J.B., Kemp, C., Griffiths, T.L., Goodman, N.D., 2011. How to grow a mind: statistics, structure, and abstraction. Science 331, 1279–1285. https://doi.org/10.1126/science.1192788.

Waltz, J.A., 2017. The neural underpinnings of cognitive flexibility and their disruption in psychotic illness. Neuroscience 345, 203–217. https://doi.org/10.1016/j.neuroscience.2016.06.005.

Wang, X.J., Krystal, J.H., 2014. Computational psychiatry. Neuron 84, 638–654. https://doi.org/10.1016/j.neuron.2014.10.018.

Woodward, T.S., Moritz, S., Cuttler, C., Whitman, J.C., 2006. The contribution of a cognitive bias against disconfirmatory evidence (BADE) to delusions in schizophrenia. J. Clin. Exp. Neuropsychol. 28, 605–617. https://doi.org/10.1080/13803390590949511.