SISSA

Scuola
Internazionale
Superiore di
Studi Avanzati

Bayesian accounts of (mis-)belief

Investigations of human inference processes
and their failure modes

Tore Erdmann

A thesis presented for the degree of
Doctor of Philosophy

SISSA Neuroscience Area
PhD course in Cognitive Neuroscience
Trieste, Italy
07.01.2022

Contents

Parts of this thesis have been published as journal articles or presented as poster:

Chapter 2: A Generative Framework for the Study of Delusions. Published in: *Schizophrenia Research, 2,2021, doi:10.1016/j.schres.2020.11.048*.

Chapter 3: Rule Learning Through Active Inductive Inference. To be published in the proceedings of the International Workshop on Active Inference (IWAI) as part of ECML PKDD 2021 (European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases).

Chapter 4: Information-sampling and delusional ideation. Presented as poster at the 2019 European Conference for Schizophrenia research (ECSR; Berlin, Germany).

Chapter 1

Introduction

Reason is the capacity to use conscious deliberation and logic to create new information from existing knowledge. It has truth as its ultimate goal. The human brain, on the other hand, evolved through a process of evolution that is not aimed at truth, but rather at survival or reproduction. Nevertheless, a common assumption is that the brain's information processing system evolved to favor beliefs that are true, or at least the ones that approximate reality, as these should be best for survival. For example, an animal holding a false belief about where the water hole is should have an evolutionary disadvantage. It seems clear that the human brain's ability to reason has been a great evolutionary advantage, that has allowed our species to control its environment and position itself on the top of the food chain. However, at the same time, our reason is bounded. This is evident not only through common errors of reasoning in everyday situations, many of which are documented by psychologists [1]. More strikingly, phenomena such as confirmation bias and rationalization show how the human reasoning capacity can seemingly become hijacked. In these cases, people, independent of their level of intelligence, mis-use the remarkable abilities to draw connections between distant observations and infer latent causes that the achievements of science are built on, to defend

incoherent beliefs of world conspiracy or supernatural causation (assumed to be false here for the purpose of this discussion). At the extreme end of the spectrum of mis-belief lie delusions. Jaspers [2] considered delusions "the basic characteristic of madness". They make up the core criteria for the diagnosis of psychosis, that is, an individual's loss of contact with reality (as defined by consensus; or in the sense of a previous personal reality). The Diagnostic and Statistical Manual of Mental Disorders (DSM), which is published by the American Psychiatric Association (APA) provides a classification of mental disorders using certain standard criteria. It was undergone various editions, with the first one published in 1952 and the latest (DSM-5, [3]) in 2013. In it, delusions are defined within the description of schizophrenia as follows:

> Delusions are fixed beliefs that are not amenable to change in light of conflicting evidence. Their content may include a variety of themes (e.g., persecutory, referential, somatic, religious, grandiose). Persecutory delusions (i.e., the belief that one is going to be harmed, harassed, and so forth by an individual, organization, or other group) are most common. Referential delusions (i.e., belief that certain gestures, comments, environmental cues, and so forth are directed at oneself) are also common. Grandiose delusions (i.e., when an individual believes that he or she has exceptional abilities, wealth, or fame) and erotomanic delusions (i.e., when an individual believes falsely that another person is in love with him or her) are also seen. Nihilistic delusions involve the conviction that a major catastrophe will occur, and somatic delusions focus on preoccupations regarding health and organ function.
>
> Delusions are deemed bizarre if they are clearly implausible and and not understandable to same-culture peers and do not derive from ordinary life experiences. An example of a bizarre delusion is the belief

> that an outside force has removed his or her internal organs and replaced them with someone else's organs without leaving any wounds or scars. An example of a nonbizarre delusion is the belief that one is under surveillance by the police, despite a lack of convincing evidence. Delusions that express a loss of control over mind or body are generally considered to be bizarre; these include the belief that one's thoughts have been "removed" by some outside force (thought withdrawal), that alien thoughts have been put into one's mind (thought insertion), or that one's body or actions are being acted on or manipulated by some outside force (delusions of control). The distinction between a delusion and a strongly held idea is sometimes difficult to make and depends in part on the degree of conviction with which the belief is held despite clear or reasonable contradictory evidence regarding its veracity.

While earlier versions of this definition used "false beliefs", this has been changed to "fixed beliefs". This is because, the falsity of a belief, while possibly being correlated, is actually irrelevant for its classification as pathological. For example, beliefs about, such as being followed or marital infidelity could naturally be true, although they may be considered as delusions by someone who is unable to verify if the event in question happened. This is an important point, which is that delusional beliefs are pathological not because of falsity or their content, but rather the way in which the belief is held. The application of Wakefields harmful dysfunction analysis of disorder [4] to belief implies that delusions are pathological beliefs, because they have to a negative impact on wellbeing (they are harmful) and fail to perform the functions of beliefs (they are malfunctioning) [5].

Regarding the study of delusions, a difficulty is that the brain processes supporting belief formation and maintenance are unknown [6]. Further problems are the lack of commonly-agreed upon and precise definitions and operationalizations.

Recent developments in artificial intelligence and probabilistic inference have accelerated the development of computational models of higher cognitive functions and have enabled the formalization of belief [7, 8, 9]. Specifically, the framework of Bayesian inference holds promise as a quantitative basis for specification of hypotheses.

An inferential perspective on delusions

Historically, delusions have been regarded by some as empty speech acts. For example, this view was held by Bleuler (see [10]), who noticed that one of his patients, who thought he was the commander of an army battalion was never seen shouting orders as an actual commander might do. Previous theories on delusions have long considered the hypothesis that delusions could arise from an individuals attempts to explain an abnormal or unusual experience ([10, 11, 12]). The observation that only some patients who encounter anomalous perceptual experiences generate a bizarre explanation and accept it as belief, while others do not, has led to proposals of an additional reasoning bias that could explain this. There have been several proposals about the nature of this bias. One influential proposal was a bias for observational adequacy, i.e. a bias towards inferences that are overly accomodating of observation [13]. Another one proposal involved a deficit of global consistency checking associated with damage to the right hemisphere [14]. A third proposal was that a deficit in belief evaluation wherein direct first-person evidence is favored over more equal weighting of different sources of evidence [15].

While there is currently no agreed upon definition of belief in philosophy, for the present work we will assume this general definition: a belief is a functional state of an organism that is available for the guidance of action. Normal beliefs have certain characteristics, such as being changable by evidence, and form the basis of higher-order cognitive faculties such as reasoning and planning. A useful representation

of degrees of belief can be given in terms of probability distributions. Given a space of possible outcomes (those that have non-zero probability of occuring), the mass of probability is equal to the degree of belief. That is, absolute certainty about a given outcome would imply a point mass at that value. Beliefs about structure or causal relationships are more complex probabilities that are commonly represented as compositions or networks of probability distributions. Further, such networks may include variables that are unobserved, that is, latent variables. The change of some beliefs in response to changes in others, such as when receiving new information can be understood in terms of conditional probabilities. A change in our belief about a hypothesis h is coherent when the *posterior probability is equal to the product of the* likelihood and the *prior, normalized by the* marginal. This is Bayes' theorem, stated mathematically:

$$P(h|d) = \frac{\mathcal{L}(d|h)P(h)}{P(d)}. \tag{1.1}$$

For example, when guessing the outcome of a dice roll, and we are given the information that the outcome is even, our new belief can be derived by application of Bayes' theorem and ought to be $P(h = x) = 1/3,\ x \in \{2, 4, 6\}$, that is, the probability mass is distributed over the events that are possible given our information. Delusions have been described as deviations from normative Bayesian inference (see [16] for a review). Formally, the proposed deviations may be operationalized as prior probabilities, or even as a change in the way the posterior is computed. For example, a bias of prior of likelihood might be implemented via a modification Bayes' theorem itself. While conceptually elegant, this view remains underspecified, and only still a point of departure. The problem is that for the exact calculation, the reasoner needs to consider every possible hypothesis $h \in \mathcal{H}$, and the number of hypotheses to consider becomes unwieldly very quickly. For

example, in a variable selection problem, deciding which of of p independent variables play a role in predicting a dependent variable, one would need to consider 2^p separate hypotheses. For a simple causal relation between 5 binary variables, where every hypothesis is a directed acyclic graph, the number of hypotheses is 543 and for 6 variables it is already 29281. Further, if considering the unfolding of events in time, an exhaustive consideration of alternatives (growing as fast as the numbers of permutations) becomes even less plausible.

Regarding delusions as beliefs about the causal structure of the world, it is therefore, rather implausible that the brain is computing the exact posterior according to Bayes' theorem. Instead, it is likely that approximate computations are involved, for which there are many different possibilities. Thus, the pattern of suboptimal inferences seen in delusions may be due to a complex interaction of inaccurate assumptions or prior beliefs, possibly faulty data (i.e. hallucinations) and their interaction with the effects of approximate computations. This suggests that the advancement of our understanding of delusions will require paradigms that investigate all these possible aspects of inference, which are currently lacking.

One approach, of which we make use of numerous times in this work, is Markov Chain Monte Carlo. This is a sampling-based algorithm for approximate inference that is able to deal with large hypothesis spaces. The idea here, is to avoid having to explicitly compute the normalizing constant $P(d) = \sum_{h \in \mathcal{H}} \mathcal{L}(d|h)$. This is possible through an iterative process of proposing local changes to ones currently maintained hypothesis and stochastically accepting them, with a finely tuned acceptance threshold which ensures that the process eventually evolves into a sampling-based representation of the posterior belief. These ideas have been used much in the field of cognitive science, but are only starting to be used in computational psychiatry.

Overview

In this thesis, I aim towards a formal characterisation of delusions. To take the first steps towards this goal, I will develop further the theoretical bases of inferential alterations in delusions and apply these ideas to different inferential processes that are commonly considered to be altered with delusional ideation. The structure of the subsequent chapters is as follows. The general theme is how uncertainty about the structure of the environment induces the need to arbitrate between different alternative hypotheses. In complex environments, the process of structure learning is of central importance, because different underlying structural relationships may be lead to the same observations, while requiring opposing actions: someone who tries to convince you that you are misunderstanding the situation may either be genuinely trying to help, or may be "part of the conspiracy". Furthermore, just the possibility of different latent structures licenses explaining-away: when the information provided by another actor conflicts with your beliefs, this can be explained away by considering some hidden intentions on their side. This in turn can lead even a perfectly rational Bayesian reasoner to have beliefs that are resistant to disconfirmation. Formally, this problem may be cast as model-selection and can be described as a search through a model space.

In chapter 2, I will introduce a generative framework for modelling delusional inferences that incorporates ideas from predictive coding and structure learning. After describing the model, I will show through simulations how beliefs may form and resist disconfirmation in a simplified context of world-model building. While the framework in chapter 2 is presented using abstract models for explanations of experience, I will develop a model that deals with highly structured, rule-based, explanations in chapter 3. Together, these ideas form the basis for several empirical investigations of different inferential processes presented in the remaining chapters.

chapter 4 contains an empirical investigation in active information-sampling using a search task. In chapter 5, I probe belief-formation in two novel paradigms. In the first section, I will apply the model from chapter 3 to analyze behavioral data in a rule learning task, while in the second section, I apply similar ideas to model evidence-accumulation under structural uncertainty.

Chapter 2

A generative framework for the study of delusions

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* [17, 18, 19, 20, 21, 22], 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 chapter 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 [23], hierarchical predictive coding [24, 25, 26], and the use and abuse of auxiliary hypotheses in hypothesis testing and Bayesian inference [27, 28, 29, 30, 31]. 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.

2.1 Theory

2.1.1 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 [32, 33]. Similarly, biases of probabilistic reasoning have been invoked to understand the process of delusion formation, such as limited data-gathering ("jumping to conclusions", [34, 35]) or a bias against disconfirmatory evidence [36]. 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.,[37, 38]), and a number of recent reviews have underlined the importance of cognitive biases and delusional ideation [39, 40, 41].

Predictive coding (PC) is a general account of brain function [24, 42] 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 [43, 44, 45, 26]. 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 ([9] [45], [43]). 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 ([9] [46]).

2.1.2 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 [47]. 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 Duhems ([27]) and Quine's ([28]) insight that evidence from an experiment cannot refute a single scientific hypothesis, but only a conjunction of hypotheses [29, 30, cf.]. [31] 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.

2.1.3 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) [48, 49]. 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 [50] for a theory of human category learning. Figure 2.1 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* [51]. For this reason, DPMMs have found broad application in the modelling of higher-order human cognition [8, 52].

2.1.4 Model description

We propose 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

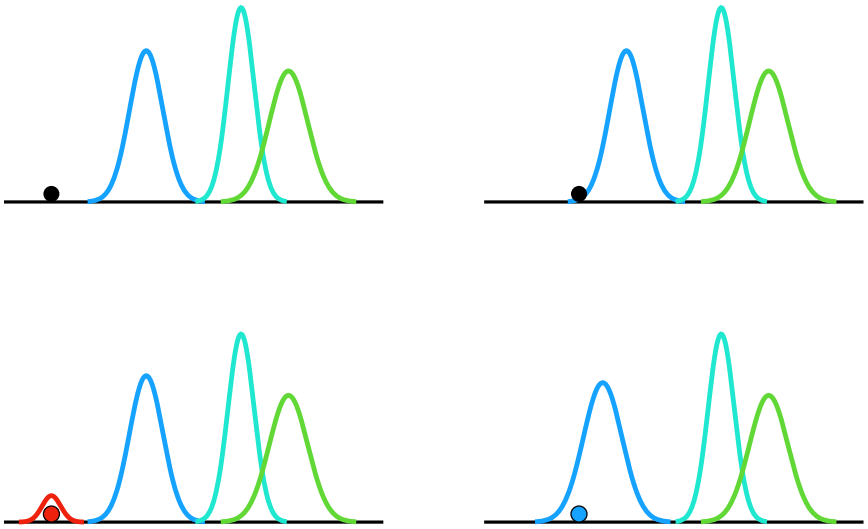

Figure 2.1: **Categorization and explanation in our framework (schematic).** In the top panels, the same initial belief is depicted on the left and right, with separate explanations (causes) represented by Gaussians. On the left, the new observation (black dot) has a larger deviation from the existing 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.

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 appendix A. We implement Bayesian inference for this model using Algorithm 8 from [53].

The assumption of an infinite collection of causes allows learners continually to

discover new ones, building new theories, as they make 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 [53].

2.2 Results

2.2.1 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 $F(y, \phi_k) = \mathcal{N}(y|\mu_k, \tau_k^{-1})$. The prior distributions for the cause-specific parameters μ_k and τ_k are independent normal ($\mathcal{N}(\mu_\mu, \tau_\mu)$) and half-normal ($\mathcal{HN}(\mu_\tau, \tau_\tau)$), 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 "over-fitting" 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 [13, 54] 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, [33] 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. [54] 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 towards 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.

Figure 2.2 illustrates this in the context of delusional mis-identification as described in a case study of Capgras' delusion [14]. Instead of attributing small variations (whatever their origin) to randomness or coincidence, patient DS infers additional explanatory structure. [14] 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 minutes 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 [55]. 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.

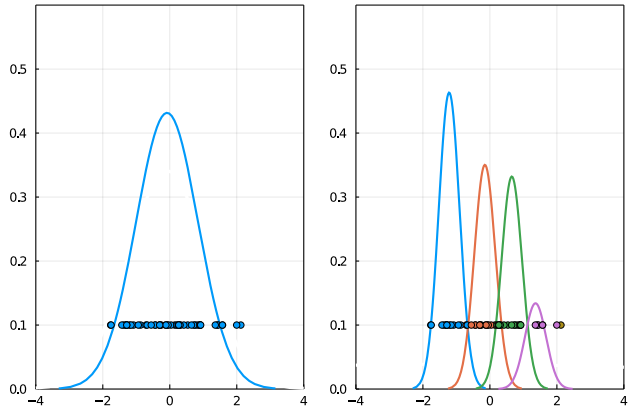

Figure 2.2: **A simulation of delusional mis-identification.** In a case study, [14] presented Capgras' patient DS with a sequence of photographs of the same models face looking in different directions (here, we represent the photographs as points on a line; observations that are perceptually similar 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 the expected precision (left: $\mu_\tau = \frac{1}{100}$, right: $\mu_\tau = 100$). Inputs and all other parameters were equal.

2.2.2 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. Figure 2.3 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 hypoth-

esis. 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 $\phi_k = (\mu_k, \tau_k),\ k = 1, \ldots$ 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 [53].

Figure 2.3 shows the change in the belief regarding the "central hypothesis". The bottom left panel shows the updated belief of an agent with a 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

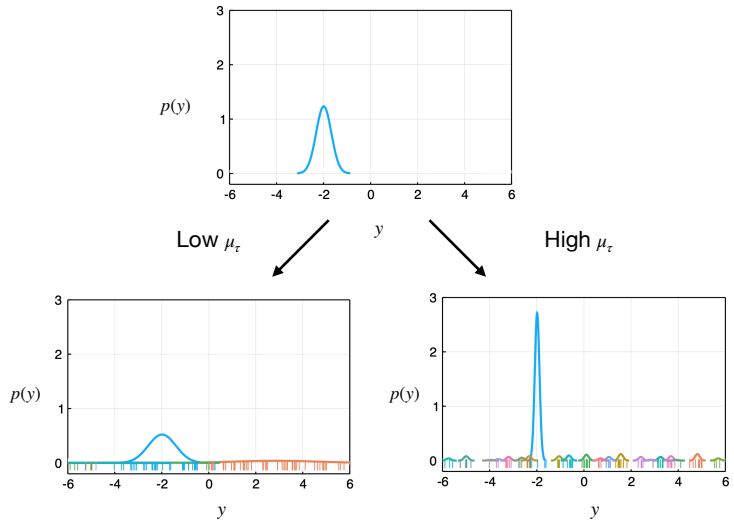

Figure 2.3: **Belief preserving evidence integration.** Initial belief (upper row) and final belief (lower row) after inference given new observations. The difference in final beliefs is a function of the expected precision μ_τ alone. All other settings and inputs are the same. Bottom left: $\mu_\tau \sim \mathcal{HN}(100, 10)$. 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: $\mu_\tau \sim \mathcal{HN}(1/100, 10)$. 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 technical details can be found in the appendix A and the code for reproducing this simulation here: https://tinyurl.com/y3m79qdw

challenge current beliefs and which could lead to very different beliefs altogether. Applying the simulation in 2.3 to the example by [33](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 wifes initials engraved in it, that this person knows things

about the subjects 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.

2.3 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.

2.3.1 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 [56, 57, 43, 26]. 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 [58, 59]. [60] 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.

2.3.2 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 [43], while two factors are required according to [33]. 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 [7]. 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.

2.3.3 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 individuals willingness to consider new knowledge from another person as trustworthy, generalizable, and relevant to the self" [61] 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 [62, 63] or the beads task [64, 65] 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 [52] and [66], 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.

2.4 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.

Chapter 3

Rule learning through active inductive inference

Abstract We propose a grammar-based approach to active inference based on hypothesis-driven rule learning where new hypotheses are generated on the fly. This contrasts with traditional approaches based on fixed hypothesis spaces and Bayesian model reduction. We apply these two contrasting approaches to an established active inference task and show that grammar-based agents' performance benefits from the explicit rule representation underpinning hypothesis generation. Our proposal is a synthesis of the active inference framework with language-of-thought models, which paves the way for computational-level descriptions of false inference based on an aberrant hypothesis-generating process.

3.1 Introduction

The account described in the chapter 2 was illustrated through examples with simplified, abstracted models of explanations. This shows an additional difficulty, inference is not a filtering process, but rather like a search. Here, we describe a

prior that allows to model the learning of highly structured hypotheses, such as rules, within the active inference framework.

Structure learning is a fundamental problem for an active inference agent. Logically structured concepts can be found in domains such as mathematics, social systems or causal processes [7]. The likelihood mapping of a POMDP with discrete state space can be represented as a matrix with elements indicating the likelihood of an observation given a state. Current approaches for learning this mapping rely on separately estimating the individual elements of the matrix [67, 68]. Here, we propose an approach for structure learning that uses a prior based on context-free grammars (CFG; [69]), which were invented in linguistics to describe the structure of sentences in natural language and are used to define programming languages in computer science. From such a grammar, the agent can, through recursive composition and substitution of terms, generate an infinite number of expressions, which represent the underlying structure of (parts of) its environment. As a proof of concept, we will illustrate our approach by applying it to a rule learning problem inspired by the task in [67].

This approach has previously been used in cognitive science, psychology [7, 8, 70] and, in particular, in "language of thought" models [71]. Previous work has shown that these models can account for various features of human concept learning ([72]). Furthermore, this approach has been used to explain surprise signals in the striatum [73].

We will work through a simplified version of the task of [67]. For ease of presentation and to place our focus on structure learning, we remove state uncertainty and all intra-trial actions except for the final choices. All remaining uncertainty is thus about the hidden rule. However, our proposal can be straightforwardly applied to the case including state uncertainty and observations corrupted by noise. In this task, see Fig. 3.1, the agent has to infer a rule, that is a deterministic

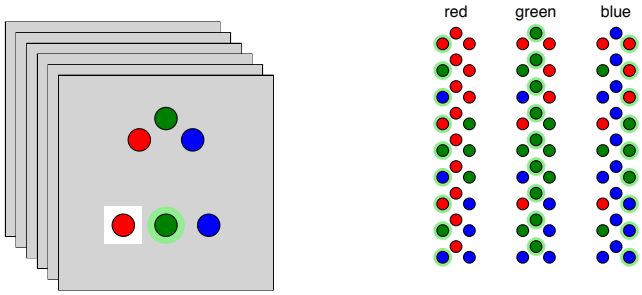

Figure 3.1: Left: shows the display during a trial. The agent sees context variables (the three circles in the upper half), makes a response (indicated by the white box around the red circle) and, having made a choice, the correct choice (highlighted in green) . Right: The possible contexts arranged according to the value of the middle circle, which implies where to look for the correct choice (highlighted in green). The correct response is equal to the color of the circle on the left, in the center or on the right when the color of the central circle is “red", “green" or “blue", respectively.

mapping from three context variables to the correct choice.

3.2 Active inference

Solving this task consists of finding a policy $p(a_t|c_t, \theta)$, that gives the probability of a choice $a_t \in \{1, 2, 3\}$ given the context variables c_t and some parameters θ. The generative model the agent holds of the task is

$$c_t^{(j)} \sim U(\{1, 2, 3\}),\ j = 1, 2, 3 \tag{3.1}$$

$$c_t = (c_t^{(1)} c_t^{(2)}, c_t^{(3)}) \tag{3.2}$$

$$o_t \sim f(c_t, \cdot) \tag{3.3}$$

$$p(r_t = 1|a_t, o_t) = \exp(\ell(a_t, o_t)) \tag{3.4}$$

$$\ell(a_t, o_t) = \begin{cases} 0 & \text{if } a_t = o_t \\ -4 & \text{else} \end{cases} \tag{3.5}$$

where $f(c, o)$ is a function representing the hidden rule. That is, it returns the probability of observing the outcome $o \in \{1, 2, 3\}$ in context c. The prior about reward observations $p(r_t|a_t, o_t)$ represents an optimistic bias, so that the agent's beliefs are biased by desirable states and not the actual task dynamics, which is $r_t = \mathbb{I}(o_t = a_t)$. This model implies a distribution over trial sequences, which we denote $\tau = (c_{1:T}, a_{1:T}, o_{1:T}, r_{1:T})$, that factorizes as

$$p(\tau|f) = \prod_{t=1}^{T} p(r_t|a_t, o_t)p(o_t|c_t, f)p(a_t)p(c_t). \tag{3.6}$$

Given the biased prior over rewards we obtain the following posterior over actions when conditioning on $r_t = 1,\ \forall t = 1, \ldots, T$ and summing out o_t, which is unknown at the time of the action,

$$p(c_{1:T}, a_{1:T}, o_{1:T}|r_{1:t} = 1, f) \propto \prod_{t=1}^{T} p(r_t = 1|a_t, o) \tag{3.7}$$

In keeping with the active inference framework, the expected log model evidence is minimized by computation of the posterior over action, which can be done at each trial t by choosing

$$p(a_t|r_t = 1, c_{1:t}, o_{1:t-1}) = \sigma(-G_{a_t}) \tag{3.8}$$

$$G_{a_t} = \sum_{o} l(a_t, o) \cdot \mathbb{E}_{p(f|c_{1:t-1}, o_{1:t-1})}\Big[p(O_t = o|c_t, f)\Big] \tag{3.9}$$

For the implementation, this means we need to be able to evaluate the agent's posterior predictive about the belief about the outcome o_t. The above constructions leads to the maximization of the following objective (see [74])

$$D_{KL}(p^*(\tau)||p(\tau)) = \mathbb{E}_{\tau\sim p^*(\tau)}\Big[\log p(\tau) - \log p^*(\tau)\Big], \tag{3.10}$$

which is the Kullback-Leibler divergence of the agent's beliefs about its future states and a desired distribution over these p^*, and which is equivalent to the free energy of the expected future, which is a lower bound on the expected log model evidence [75].

3.2.1 Evidence accumulating agent

A straightforward solution for learning the rule is available if we represent it as a stochastic vector consisting of independent Dirichlet variables, $f(c, o) = \theta_{c,o}$, with $\theta_{j,\cdot} \sim Dir(\alpha_0)$, $j = 1, \ldots, 27$, for which the posterior can be computed by accumulation of concentration parameters:

$$p(\theta_{j,o}|c_{1:t}, o_{1:t}) = Dir(n_{c,o} + \alpha_0) \tag{3.11}$$

where $n_{c,o}$ is the number times (up until time t) the agent has observed outcome o for context c. If we define a matrix α with entries $\alpha_{c,o} = n_{c,o} + \alpha_0$, the expectation in eq. 3.8 is a that of a categorical-Dirichlet distribution and the action is chosen via

$$G_{a_t} = \sum_o \ell(a_t, o)) \cdot \frac{\alpha_{c_t,o}}{\sum_j \alpha_{c_t,j}}. \tag{3.12}$$

3.2.2 Bayesian model reduction

If the agent knows that there must be a deterministic rule, it can quickly recognize the rule by comparing the evidence for each potential model in a set of hypothetical models and accept a model if its evidence exceeds a certain threshold.

The model space can be considered the set of deterministic, one-to-one mappings from each color to each response (of which there 6) which are combined with the 6 possible mappings between the central color and which location the color-to-response mapping should be applied to (see [67]). There are thus 36 hypotheses,

for which the evidence is computed on each trial. This allows us to represent the priors through sets of prior concentration parameters as derived in [67]. A condition for this agent is that the space of hypotheses is specified for the agent beforehand, which is a strong assumption in general. We will now introduce a way to model acquisition of new models. This has the advantage of being based on weaker assumptions about (and a different conception of) prior knowledge.

3.3 Grammar-based rule induction

Here, we describe how rule learning can be supported through a structured prior over an auxiliary space of symbolic rule expressions. Each such rule expression is defined by a syntax tree, consisting of logical connectives (and, or), and references to the observations in a trial. The "leaf nodes" of the tree are predicates of some part of the observation c_t, for example $color(c_t^{(1)}) = red$, which is either true or false (see appendix B for an example). An agent can learn a rule expression that accurately predicts the outcome of the unknown rule f by searching the space of rule expressions for hypotheses which are then evaluated against the available evidence. Hypotheses are represented by expressions that can be generated by iterating the following set of re-write (or production) rules:

$$\begin{aligned}
\text{(Start)}\quad & S \rightarrow f(c, o) \iff (D) \\
\text{(Disjunction)}\quad & D \rightarrow C \vee D \mid P \mid false \\
\text{(Conjunction)}\quad & C \rightarrow P \wedge C \mid P \mid true \\
\text{(Predicate)}\quad & P \rightarrow color(Loc) = Col \\
\text{(Location)}\quad & Loc \rightarrow c_1 \mid c_2 \mid c_3 \\
\text{(Color)}\quad & Col \rightarrow \text{“}red\text{”} \mid \text{“}green\text{”} \mid \text{“}blue\text{”}
\end{aligned}$$

These rules indicate how symbols on the left hand side of the $\rightarrow$ can be replaced by one of the options on the right hand side (options are separated by |). From this grammar, given certain production probabilities (which give the probability of each possible production for each line in the grammar; can be assumed uniform), we can generate rule expressions (we refer the interested reader to Wikipedia, for examples, or [76] for a comprehensive treatment). Note that we omit the trial index t in the formulas (since the rules only refers to variables in the current trial) and instead use the subscript to denote the location (1, 2 or 3) of the context variable.

Each generated expression describes some arrangement of context observations. Say, we wanted to describe the rule for when the correct color is red (as given in the caption of 3.1). This can be expressed as $color(c_2) = \text{“}red\text{”} \wedge color(c_1) = \text{“}red\text{”}) \vee (color(c_2) = \text{“}blue\text{”} \wedge color(c_3) = \text{“}red\text{”})$, which can be generated through step-wise replacement of the above rules. The prior probability of a formula (i.e. a sequence of substitutions from the grammar) is equal to the product of the probabilities of the individual substitutions. This prior naturally places higher probability on shorter and less complex expressions since they include fewer terms in the product.

For the rule learning task described above, we want to model the contexts that correspond to the three outcomes (and actions), so we will make the procedure to be learned a function of both the observed context c and the outcome o, changing the rule in the topmost line above to be a context-sensitive expression of the form

$$S \rightarrow f(c, o) \iff$$
$$((o = \text{“}red\text{”}) \wedge D) \vee ((o = \text{“}green\text{”}) \wedge D) \vee ((o = \text{“}blue\text{”}) \wedge D),$$

wherein the D terms will come to represent the parts of the rule that imply the corresponding outcome. We can then evaluate expressions with regard to each

possible outcome to determine if the context c matches the outcome o. Starting from the above expression and generating sub-expressions according to the above grammar, we can represent the true hidden rule described in Fig. 3.1 as follows:

$$
\begin{aligned}
f(c,o) \iff & ((o = \text{“}red\text{”})\wedge \\
& ((color(c_2) = \text{”}red\text{”} \wedge color(c_1) = \text{”}red\text{”})\vee \\
& (color(c_2) = \text{”}blue\text{”} \wedge color(c_3) = \text{”}red\text{”}))) \\
& \vee ((o = \text{“}green\text{”})\wedge \\
& ((color(c_2) = \text{”}green\text{”})\vee(color(c_2) = \text{”}red\text{”} \wedge color(c_1) = \text{”}green\text{”})\vee \\
& (color(c_2) = \text{”}blue\text{”} \wedge color(c_3) = \text{”}green\text{”}))) \\
& \vee ((o = \text{“}blue\text{”})\wedge \\
& ((color(c_2) = \text{”}red\text{”} \wedge color(c_1) = \text{”}blue\text{”})\vee \\
& (color(c_2) = \text{”}blue\text{”} \wedge color(c_3) = \text{”}blue\text{”})))
\end{aligned}
$$

However, we can represent this rule more succinctly by adding more abstract terms to the grammar. For example, by adding two new production rules to the grammar above:

$$
\begin{aligned}
P &\rightarrow color(Loc) = COL \;\mid\; o = color(Loc) \\
Loc &\rightarrow c_1 \mid c_2 \mid c_3 \mid c_{Loc}
\end{aligned}
$$

The last production will lead to a “subsetting”, such as c_{c_2}, which means that the value of c_2 indexes the context variables (with the colors mapped to the numbers $\{1,2,3\}$). The expression $o = color(Loc)$ evaluates to true if the outcome matches the variable Loc. With these additions, we can now represent the true

rule as a much shorter expression

$$f(c, o) \iff (o = color(c_{c_2})). \tag{3.13}$$

This shorter representation of the rule helps the agent to discover it much more quickly. This is because shorter rules have higher prior probabilities of being produced.

The above rule expression defines a function that evaluates to `true` if the action a is correct given the observation o and `false` otherwise. The likelihood of this expression is given by its match with the observed data, that is, the number of examples for which the rule f evaluates to true,

$$p(f|o_{1:t}, a_{1:t}, c_{1:t}) \propto \bigwedge_{c,o} f(c, o) \tag{3.14}$$

or, if assuming that some observations might be outliers to the rule, we have

$$p(f|o_{1:t}, a_{1:t}, c_{1:t}) \propto e^{-\gamma Q(f)} \tag{3.15}$$

where $Q(f) = |\{(c, o) \in (c_{1:t}, o_{1:t}) : f(c, o) = false\}|$ (the count of examples for which the rule expression evaluates to false) and γ is a parameter denoting the probability that a given example is an outlier. Here, the probabilities need not be normalized, since any normalization constants cancel in the MCMC acceptance probability. The truth value of the procedure $f(a, o)$ follows from the evaluation approach in mathematical logic [77] and is defined recursively:

1. $f(a, o)$ is a node.
2. If a node is a predicate, it can be evaluated directly
3. If it is a logical connective then it is evaluated by first evaluating the sub-expressions separately and then applying the logical function to the result.

For example, $a \wedge b$ is true only if both sub-expressions a and b are true.

In our implementation, we represent the agent's belief about the correct rule expression as a set of samples that are approximately distributed according to the posterior distribution implied by the above likelihood and prior. This posterior is updated on each trial by running a Markov Chain Monte Carlo (MCMC) chain for a fixed number of iterations. The set of expressions that was visited during the walk is taken to represent the posterior belief. This construction leads to the posterior predictive distribution, given a set H_t of hypotheses. Formally, if we denote the chain representing the belief update in trial t by $H^{(t)} = \left(h_1^{(t)}, \ldots, h_n^{(t)}\right)$, we can evaluate the posterior expectation in action selection in eq. 3.8 approximately as follows

$$p(a_t = a | r_t = 1, c_{1:t}, o_{1:t-1}) = \sigma(-G_{a_t}) \tag{3.16}$$

$$G_{a_t} = \sum_o l(a_t, o) \cdot \mathbb{E}_{p(f|c_{1:t-1}, o_{1:t-1})}\Big[p(O_t = o | c_t, f)\Big] \tag{3.17}$$

$$\approx \sum_o \ell(a_t, o)) \cdot \frac{\sum_i f_{h_i^{(t)}}(c_t, a)}{\sum_{j \in \{1,2,3\}} \sum_i f_{h_i^{(t)}}(c_t, j)} \tag{3.18}$$

which can be seen as a model average of all hypotheses that were visited by the Markov chain during the computation of the posterior.

The iterations of the MCMC procedure propose changes to the expression by randomly selecting a sub-expression and replacing it with a newly generated sub-expression. The Metropolis-Hastings acceptance probability for a proposal balances the probability of the proposal and the reverse proposal, the prior probabilities and the likelihood (see eq. 3.14) of the current and proposed expressions (tree-substitution MCMC; see [72] for details). The belief update can thus be performed by running n MCMC iterations, starting from the current state of the chain. For the current task, once the true rule has been found, proposal for moves away from it will have very low probability. In general, when the rule cannot be

known with certainty, the chain will move between alternatives and thereby lead to a representation of the remaining uncertainty in the posterior belief about the rule.

3.4 Experiments

We simulated learning in four agents who completed 20 trial sequences each. These sequences contained 27 trials and were generated by randomly shuffling the 27 unique combinations of context variables. The four agents differed in substantial ways and could be characterized as concentration parameter accumulating agents (Agents 1 and 2, described in section 3.2.1) with (Agent 2) and without (Agent 1) model-selection (by Bayesian model reduction, sec. 3.2.2) after each trial; and the grammar-based agents (Agents 3 and 4) with the simple grammar described in 3.3 (Agent 3) and an extended grammar described below (Agent 4).

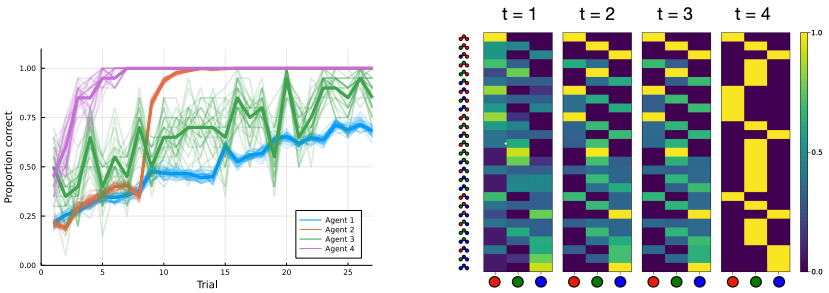

Figure 3.2: Left: Proportion of correct choices (averaged over simulations) for the four agents with uncertainty indicated via bootstrapped estimates (thin lines). Right: Belief of (purple) grammar-based agent for the first 4 examples of a particular trial sequence. Each heatmap shows the probability of an action (x-axis) to be correct in a given context (y-axis).

A comparison of the performance of the different agents are shown in Figure 3.2, where the average proportion of correct responses is shown over trials. As can be seen, the grammar-based agents show higher proportions of correct responses

already during early trials. This is due to the nature of the rule expressions, which can be extrapolated from rapidly.

The best-performing agent (purple in Fig 3.2) is Agent 4 with the extended grammar, that has two additional production rules contained in its grammar (see sec. 3.3). Figure 3.2 (right) shows the Agent 4's belief about the rule during the early trials of a particular trial sequence. The examples presented to the agent were $((\bullet,\bullet,\bullet),(\bullet,\bullet,\bullet),(\bullet,\bullet,\bullet),(\bullet,\bullet,\bullet))$, for which the correct responses were $(\bullet,\bullet,\bullet,\bullet)$. We can inspect the set of hypotheses held by the agent. At $t = 3$, the hypotheses with the highest weights (about $n/3$ occurrences) are:

1. $a = color(o_1)$, "answer equal to the left circle"

2. $a = color(o_{o_2})$ "answer equal to the color at location indicated by o_2"

3. $a = color(o_{o_3})$ "answer equal to the color at location indicated by o_3"

The agent cannot tell between these explanations until observing the outcome in the 4th trial, when the predictions of hypotheses 1 and 3 are disproved and the agent correctly infers the rule.

These results show how learning speed relates to underlying assumptions. As opposed to Agents 1 and 2, who need to be equipped with a fixed hypothesis set, the grammar-based agents can learn arbitrary rules, including such for which maintaining a fixed hypothesis set would be infeasible, as long as they can be represented within the language spanned by their grammar. For example, if we had just told the agent: "In this game, there is a deterministic mapping between the three colored circles and the correct response", the hypothesis space would have to cover a space of mappings containing 2^{81} elements. Comparing each of these candidates at the end of a trial would be infeasible (at a rate of 10^9 evaluations per second (1 evaluation per nanosecond), it would take about 77 million years to

evaluate all candidates). The code for all experiments reported here is available at https://github.com/ilabcode/IWAI2021.

3.5 Discussion

We have shown a novel way to perform structure learning in active inference agents. In particular, we demonstrate how an agent can use grammar-based structure learning to develop a model in a bottom-up fashion. This is different from the traditional approach of Bayesian model reduction, which can be considered a top-down approach. The assumption of a grammar that spans a hypothesis space is weaker and hence more generalizable than pre-defining a finite set of hypotheses. Other ways of searching for rule expressions are possible, such as genetic algorithms, but these do not represent uncertainty and are therefore not well suited as a basis for adaptive prediction and decision-making.

Our results showed differences between the two grammar-based agents that were apparent in the speed by which they learn the rule. For the task presented here, both agents converge to the same behavior, but their underlying rule representations are different. This highlights how higher-order inferences can depend on the base of concepts and abstractions they are built upon. In terms of the behavior, the agents will look the same, however, their representational vocabulary differs and so they will find separate explanations for the rule (which do describe the same contingencies), which also have different complexity (as clearly visibly in the number of terms). Given a way to update their own grammars through experience, two agents starting with different grammars but in similar environments might develop a similar conceptual toolbox. One way to enable this would be to add special "lambda expression" terms to the grammar. Such an encoding of the lambda calculus within the hypothesis language leads to the ability to define new

terms and apply or re-combine them (see [78]).

An interesting aspect of your hypothesis-generating grammar-based approach is the ways in which the assumptions underlying the generation of hypotheses of can influence what the agent finally takes to be the most promising course of action. This can become a useful tool for understanding aberrations in world modeling such as those apparent in psychiatric illnesses, which might have to do with a deficient hypothesis-generating process. For example, hypotheses generated from a grammar that is poorly attuned to a domain can seem bizarre to outside observers. Such misattunement may be the result of aberrant learning processes that update the production probabilities of a grammar, or the addition or removal of terms.

The agent described in [67] did not include model-selection considerations in its actions since they were outside of its generative model (and, in any case, the actions in the task were uninformative in that regard). By contrast, with a grammar-based approach, the structure is part of the agent's prior. Therefore its actions can subserve the testing of freshly generated hypotheses about the hidden structure of a task, which corresponds to active learning. Crucially, this could be made relevant in a version of the rule learning task where the agent can choose its next set of context variables. This would require planning, where the agent finds the optimal plan for testing its currently most promising hypotheses — an interesting avenue for future research based on the approach introduced here.

Chapter 4

Information-sampling and delusional ideation

Abstract Delusional ideation is associated with probabilistic reasoning biases both in patient and general-population samples. Here, we investigate a complementary aspect of information-gathering due to the *explore/exploit dilemma*. Using a recently developed searching task, we estimate individual-specific parameters for tendencies towards directed and undirected exploration in a general-population sample with varying tendencies for delusional ideation. We find preliminary evidence that participants with higher scores on the Peter's delusional ideation (PDI) questionnaire are less guided by model-based uncertainty, showing less directed exploration, but similar levels of random exploration compared to people with lower scores. Further, higher PDI scores were associated with less well calibrated confidence judgements about search performance. These findings illustrate the value of the combination of tasks that are computationally characterized and allow to make inferences about inference processes. Further, these results point towards a potential usefulness for predicting the outcomes of treatments of delusional ideation.

4.1 Introduction

Delusions are fixed beliefs that are resulting from incorrect inference about reality [3]. They can be severely debilitating, contribute to social isolation of patients and have among the highest estimated disease burden among mental diseases [79]. Previous work has found evidence for various probabilistic reasoning biases to be associated with delusional ideation with the most prominently reported bias the jumping-to-conclusion (JTC) bias, a tendency to stop gathering data earlier than implied by an (Bayesian) ideal observer analysis [80, 81, 35].

However, in this task, since there is only one *kind* of action, and hence conflates different styles of information-gathering, that typically come up in more realistic environments with multiple uncertain options. Such situations create a fundamental trade-off between pursuing options that we expect to give high rewards (exploitation) or sampling other options that have lower expected rewards, but may result in new information that could lead to higher rewards in the long-term. This dilemma is central to reinforcement learning, and has started to be treated in computational neuroscience [82, 83, 84].

Optimal decision-making does not only depend on the expected reward, but also on the expected information gain. Further, these expectations have to be derived form a model of the environment. Given such a model, we can make *directed* exploration, thinking through future states (and, in particular, their effects on the agent's later choices), which is intractable in all but the most simple situations and would thus have to be computed approximately. Without any model, or due to limited processing capacity, one may only resort to *random* exploration.

Previous empirical work has a number of novel tasks that allow to estimate these behavioral strategies [85, 86, 87]. The authors in [85] introduced the horizon task, where people made decisions in two contexts that differed in the number

of choices they could make (time horizon). Using a “restless-bandit task", work by [88] provided behavioral evidence for “exploration bonuses” in people. In [89], the authors investigated a new paradigm based on “function learning”. This was further developed by [87] to show that human exploration and generalization in spatial “grid” domains can be modelled by function learning via Gaussian processes. Another study probed exploration in risky environments [90] using another version of the grid-search task. They found some people to not generalize far even in the risk-free condition, which might indicate individual differences in the tendency to uncertainty reduction. Further, [91] found that children use less generalization and more directed exploration. Of particular relevance is a recent study that found evidence of reduced exploration in patients with schizophrenia [92]. This study employed the “horizon task” and found reduced directed exploration, but no difference in random exploration, and did not assess confidence in relation to amount of information gathered through exploration.

Here, we aim to investigate information-gathering in a general population sampling with varying tendencies for delusional ideation. In particular, we are interested how exploration strategies relate with metacognitive processes. To this end, we collected behavioral data on the grid-search task by [87], but additionally added asked for of confidence judgements of the participants about their own search performance (we asked them whether they thought they found the global optimum).

Figure 4.1 shows the display shown to the participants during the task. Participants do not only choose that they want to sample evidence, but they can to some extend control what kind of evidence. This allowed us to characterize searching behavior in terms of directed and un-directed (random) exploration.

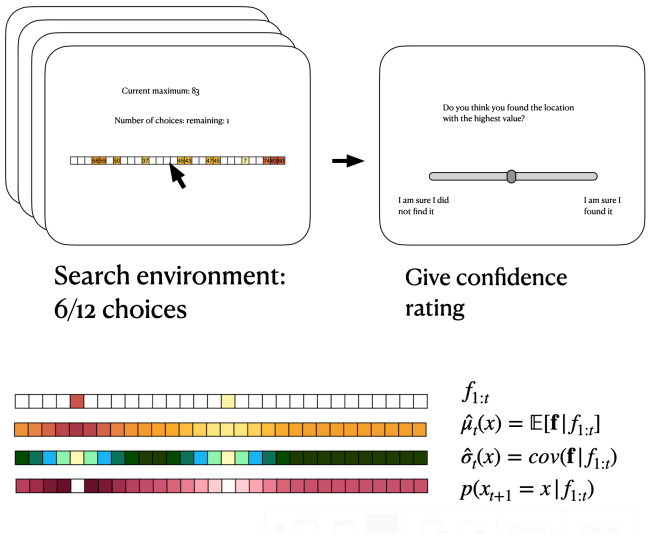

Figure 4.1: **Gridsearch-task:** Top panel: The display as seem by the participants during the experiment. The environment was represented as a row of boxes. After clicking on a box, its numeric value was shown inside it and through its fill color. After each trial, participants gave confidence ratings using a slider (shown right). Bottom panel: For a given state of the task, we modelled the expected values and uncertainties of all choices which combine to determine the action values (and choice probabilities; see methods 4.2.4).

We use computational modeling to explain differences in behavior in terms of differences in the inferential process (as described by two parameters of the belief model). Given the previous findings with regard to the JTC bias, we hypothesized that people with tendencies for delusional ideation would gather less information, and be more confident even when having less information available.

4.2 Materials and Methods

4.2.1 Participants

We recruited 62 participants online through Amazon Mechanical Turk. We did not restrict the participants with regard to the number of previously accepted assignments or other characteristics. Our measure of the participants' tendencies for delusional ideation was Peters et al. Delusions Inventory (PDI; [93]). This scale is widely used as a measure of delusion and delusion-like beliefs in both clinical and non-clinical samples.

4.2.2 Procedures

The participants completed the grid-search task followed by the original version of the beads task ([81]). In the beads task, participants were shown a sequence of beads that they were told could come from one of two jars. Each jar contained 60% of a dominant color and the participants were told to figure out which jar the sequence of beads came from. For each sequence they could request to see another bead or decide for one of the two jars. We used three particular sequences, (1) 01000010001011110111; (2) 01000100101000011001; and (3) 10111101110100001000 (where "0" and "1" to stand in for a particular color). The number of beads the participants requested before deciding, called *draws-to-decision* were recorded for each of the three sequences. After completion of both tasks, participants were asked to fill out the PDI questionnaire and several other questions (age, gender, education, prior psychiatric diagnoses).

4.2.3 Grid-search task

We adapted a version of the exploration task from [87]. Figure 4.1 shows the display during a trial. The participants could see the the values of their previous choices and selected the next box to reveal. A single trial consisted of either 6 or 12 choices ("search horizon" condition). This restricted number forced participants to use a sampling strategy to select their sampling locations that optimally explored the environments. The underlying structure was created by drawing random samples from a Gaussian Process distribution with length-scale parameter 2 and evaluating the draws on an equally spaced grid of values. The resulting values were re-scaled to lie on in range $(0, 100)$ and then represented the values of the 30 boxes. After each trial, the participants rated their confidence regarding whether they found the box with the maximum value.

4.2.4 Analysis

Model-agnostic analysis of confidence

We compare tendencies to delusional ideation (as measured by the PDI score) with behavior in the task. We analyzed the confidence ratings using a regression model that was defined as follows:

$$\begin{aligned}
\eta_{ij} &= \alpha_i + \beta_1 \cdot x_j^{\text{correct}} + \beta_2 \cdot x_j^{\text{horizon}} + \beta_3 \cdot x_i^{\text{pdi}} + \beta_4 \cdot x_j^{\text{horizon*pdi}} \qquad (4.1)\\
y_{ij} &\sim \mathcal{N}(\eta_{ij}, \sigma^2),\ i = 1, \ldots, 62,\ j = 1, \ldots, 16\\
\alpha_i &\sim \mathcal{N}(0, 1),\ i = 1, \ldots, 62\\
\sigma^2 &\sim \mathcal{HN}(0, 1)
\end{aligned}$$

We also tried different versions, adding additional predictors, but found the above to best balance data-fit and parsimony.

Model-based analyses of search behavior

In accordance with the true generative process underlying the spatial distribution of values (which was based on a smooth underlying function), and following previous work that found it to best account for behavior [87], we use a Gaussian Process (GP) prior over the unknown underlying function. Learning can then be modelled as posterior computation. A GP is defined as a collection of points where any subset of these points is distributed according to a multivariate Gaussian. For a function that maps from input space to real scalar outputs $f : \mathcal{X} \mapsto \mathbb{R}$ we have:

$$f \sim \mathcal{GP}(m, k),$$

where m and k are the mean function and covariance kernel (see below) of the Gaussian process. For this model, the participant is assumed to maintain beliefs about each of the spatial locations $x_k,\ k = 1, \ldots, 30$. At any time t, given a set of choices and revealed values at those choice locations $\{x_i, y_i\}_{i=1}^{t}$, we obtain a posterior predictive belief (see [94]) about each of the locations that is given by a multivariate Gaussian with mean $\hat{\mu}$ and variance $\hat{\sigma}^2$ as:

$$\hat{\mu}_t(x^*) = k_*^T (K + \sigma_x^2 I)^{-1} y_t \tag{4.2}$$

$$\hat{\sigma}_t^2(x^*) = k(x^*, x^*) - k_*^T (K + \sigma_x^2 I)^{-1} k_* \tag{4.3}$$

where k denotes the kernel k defines covariance function based on the spatial proximity of the boxes. We use the squared-exponential kernel ($k(x, y) = \exp(-((x-y)/l)^2)$), which leads to smooth spatial correlations of neighboring boxes. The parameter l specifies the width of the kernel with larger values leading more smoothness and smaller values to more roughness in the spatial distributions of the values.

For the fitting procedure, since the estimation of the l parameters of the kernel

were not of primary interest here, we selected three fixed values $l = 1, 2, 3$. These values can be seen as representing spatial under-generalization, correct generalization and over-generalization, respectively. To maximize the function participants need to appropriately balance exploration and exploitation. We compare different *acquisition functions*, that define the utilities of choosing any particular location. We compared 5 different models, which we combined with the choices for l such that we fit 15 models in total per subject. In the following, we will describe the different models.

Model 1 (baseline) In general we model the subjects' choice of the next location x_{t+1} using the softmax choice rule:

$$p(X_{t+1} = x | x_{1:t}, y_{1:t}) = \frac{\exp(u(x) * \tau)}{\sum_{x' in \mathcal{X}} \exp(u(x') * \tau)} \tag{4.4}$$

where τ is a parameter denoting choice stochasticity (or random exploration), and $u(x)$ denotes the *utility* of choosing the location x. The utilities for this model are given by the Upper Confidence acquisition function (UCB):

$$u_{\text{UCB}}(x) = \hat{\mu}_t(x) + \beta \cdot \hat{\sigma}_t^2(x) \tag{4.5}$$

where β denotees the *exploration bonus* parameter.

Model 2 (with inertia) For second model, following [87] we tested inclusion of an inertia parameter γ, which modelled a preference for choices close to the last choice. This is implemented via a re-weighting of the utilities according to their distance from the last choice:

$$u_{\text{UCB}^*}(x_t) = u_{\text{UCB}}(x_t) \cdot \left(1 - \left(\frac{d(x_{t-1}, x_t)}{\max_x d(x, x_t)} \cdot \gamma\right)\right) \tag{4.6}$$

with the "inertia parameter" $\gamma \in (0, 1)$, denoting the degree of preference for locations close to the previous choice.

Model 3 (with decaying exploration bonus) Additionally, we added a new parameter β_{rate} to model a tendency for reduced exploration (and increased exploitation) towards the end of the trial. We define a decaying function

$$\beta(t) = \beta_0 * \exp(-\beta_{rate} * t) \tag{4.7}$$

where t is the number of choices in the current trial and this time-varying temperature was then multiplied with the expected uncertainty:

$$u_{\text{UCB-decay}}(x) = \hat{\mu}_t(x) + \beta(t) \cdot \hat{\sigma}_t^2(x) \tag{4.8}$$

Models 4 and 5 (greedy with and without inertia) Additionally, included a simpler "greedy" acquisition function, that assigns utility to the choices based solely on their expected values:

$$u_{\text{greedy}}(x) = \hat{\mu}(x) \tag{4.9}$$

These utilities were again transform via the softmax, such that the resulting policy resembles an "ϵ-greedy" policy.

4.2.5 Model fitting and model checking

All models were implemented and fit in `Julia` (using the package `Turing.jl`), with the No-U-Turn sampler (NUTS). The 4 MCMC chains were run for 200 iterations per model and dataset and each chain was controlled for convergence visually and via the R-hat statistic. Figure 4.2 shows the PSIS-LOO values for each model.

Models 2 and 3 were best-fitting in terms of their PSIS-LOO values per participant, with model 2 having one parameter less. We thus concluded Model 2 to best explain our observed data. We compared the PSIS-LOO values (parteo-smoothed importance sampling leave-one-out) for each model across participants which confirmed Model 2 as having the best predictive validity.

The priors for the parameters were as follows:

- $\beta \sim \text{Exponential}(1)$
- $\tau \sim \text{Half-Normal}(1, 0.1)$
- $\gamma \sim \text{Beta}(0.1, 0.1)$
- $\beta_{rate} \sim \text{Exponential}(0.1)$

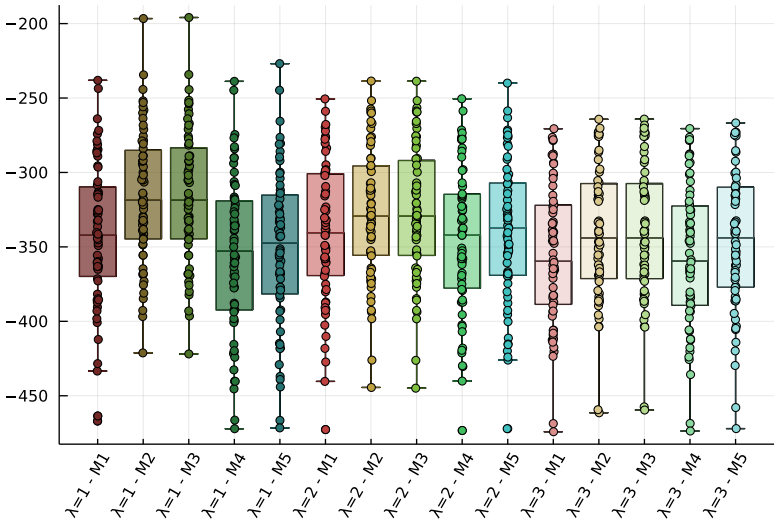

Figure 4.2: **Model comparison:** PSIS-LOO values for all 15 models that were fitted to participant data, where M1-5 denotes models 1 to 5 as defined above. Higher values indicate a better fit.

To validate our parameter estimates reported, we performed parameter recovery simulations. For this, we repeatedly simulated data (10 times per parameters

estimated for each subject) and estimated the parameters for those data. The results are shown in figure 4.3 and are indicative of good recovery.

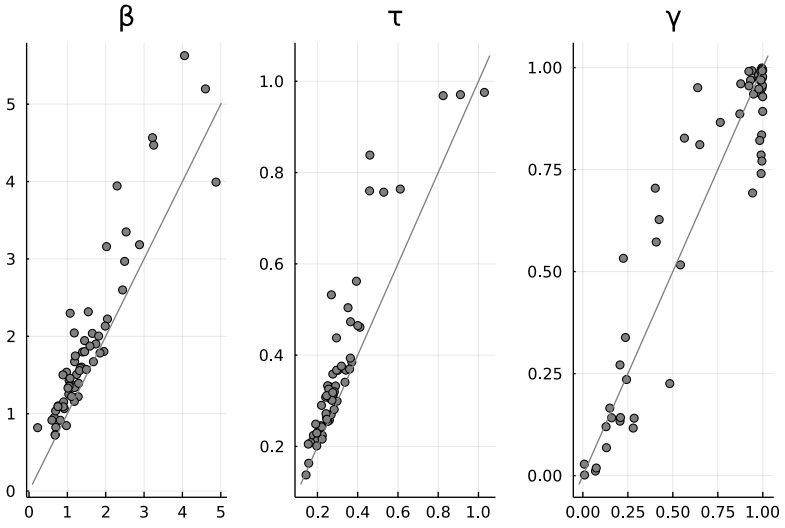

Figure 4.3: **Parameter recovery:** Simulated (x-axis) vs. Estimated (y-axis) parameter values for the model winning the model comparison (Model 3 with $\lambda = 1$).

4.3 Results

Figure 4.4 depicts two statistics of the observed data (and, for reference, a dataset generated by simulated random behavior). In the grid-search task, participants performed slightly better than chance in both short and long horizon conditions. The distribution of step-sizes is distinctly different from a random behavior. Note that the left panels of figure 4.4 shows the optimum found up to click t, which is by definition monotonically increasing.

4.3.1 Higher PDI relates to task behavior that is less directed explorative

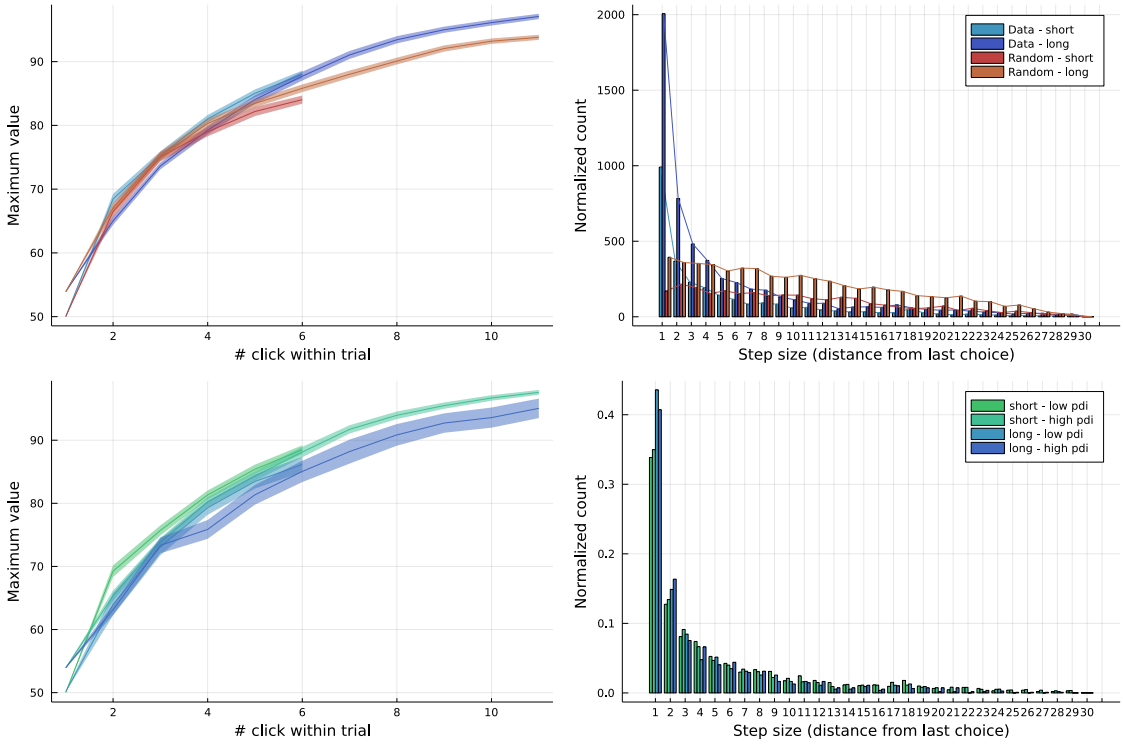

Figure 4.4: **Behavior:** Upper row: Observed data statistics: average cumulative maximum found by participants during unfolding of a trial (left) and the distribution of step sizes split by horizon condition. Lower row: Split by PDI score: Shown are the same statistics as above but here split into two groups according to the PDI score of the participants.

We find slight differences in performance (as depicted in fig 4.4 panel (d)), and in the distribution of step-sizes, which participants in the "high PDI" group performing slightly worse and showing more localized exploration (more clicks with lower distance to the previous click).

We report the estimates from the model winning in an extensive model com-

parison (see section 4.2.5 for details). Comparing the GP model with a simpler *greedy* strategy (that ignores information-gain during action selection; see section 4.2.4), which is defined by a fixed exploration bonus parameter setting $\beta = 0$, we find that participant's behavior is better explained by a model that includes a non-zero value for β and also includes γ, an "inertia" parameter that operationalizes a preference for choices close the one's previous choice.

Figure 4.5 shows the parameter estimates for the winning model of the model comparison. The data did not allow us to determine a difference in the exploration bonus parameter in the high vs. low PDI groups as hypothesized. Instead we find slightly higher stochasticity (or random exploration) and high values for the inertia parameter γ.

We compared models with regard to three possible values for the generalization parameter λ. Figure 4.2 shows the values of model goodness. The model family with $\lambda = 1$ won, indicating a tendency for undergeneralization (or underestimation of the smoothness of the underlying functions).

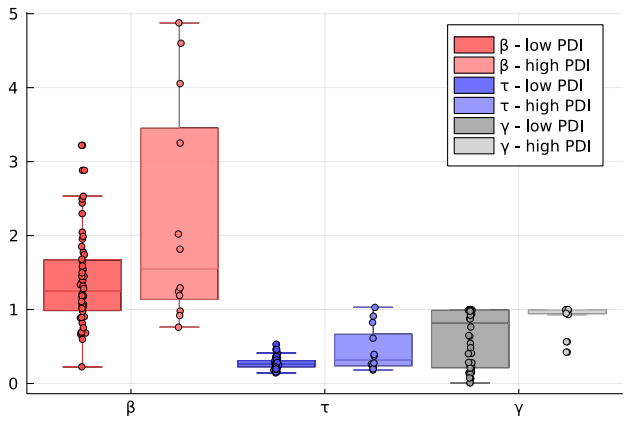

Figure 4.5: **Parameter estimates:** Shown are the parameter estimates for model 3 for all subjects, but split according to the subjects' PDI scores.

4.3.2 Confidence judgements of people with higher PDI are higher and less sensitive to the available information

Figure 4.6 shows the raw confidence ratings split by "horizon" condition (the number of choices the participants had for that trial; either 6 (short) or 12 (long)). We fit a linear mixed-model to the confidence judgments of the participants on each trial (see eq. 4.1 for definition). Although the confidence ratings are bounded, most of the ratings fell within the bounds (see fig. 4.6 which allows us to model them as continuous. The results are summarized in table 4.1. We find that PDI had a main effect (HPD : $[0.049, 0.138]$; indicative of over-confidence) and an interaction with the search-length (horizon; HPD : $[-0.105, -0.04]$). Specifically, we found higher PDI values attenuating the otherwise positive effect of the search-length on confidence (HPD : $[1.41, 1.81]$).

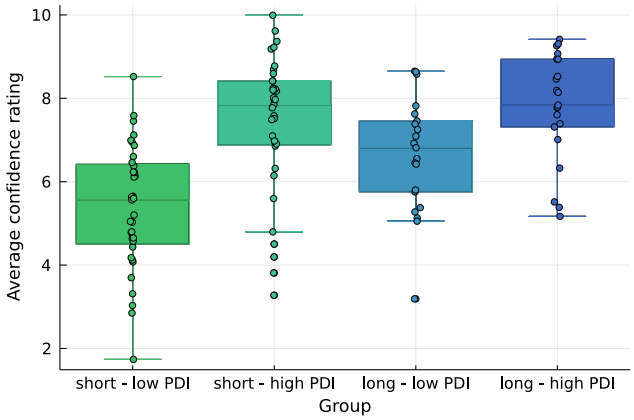

Figure 4.6: **Confidence judgements:** Shown are the average confidence judgements for all trials of all subjects, split according to the participants' PDI score and the horizon condition (long or short).

Variable	Mean	Std.Err	90% HPD Interval
Correct	1.159	0.132	[0.996, 1.340]
Horizon	1.615	0.152	[1.410, 1.807]
PDI	0.094	0.036	[0.049, 0.138]
*PDI*Horizon*	−0.073	0.025	[−0.105, −0.040]

Table 4.1: **Results for regression of confidence judgments:** Posterior means, and standard deviations and highest-posterior-density (HPD) intervals for the parameters of the regression model.

4.3.3 Relation to beads-task

We collected data from three bead sequences per subjcet and did not attempt a model-based analysis. Instead, compared the mean number of beads sampled *draws-to-decision* (DTD) to PDI scores and the parameter estimates. Figure 4.7 shows scatter plots of the DTD, the participants' PDI scores, and the parameters estimates for the best-fitting models. Further, we calculated correlations between the mean confidence judgements across trials and the draws-to-decision, however which was not significant.

4.4 Discussion

We have investigated the relation of information-sampling and tendency for delusional ideation in a general population sample using a novel task that allows to characterize specific features of information-gathering behavior. We found subtle differences in performance and confidence ratings between subjects.

We found over-confidence and weaker sensitivity to the amount of available information in high PDI subjects. Further, through model-based analyses, we found that the apparent lack of exploration in the same subjects to be due not

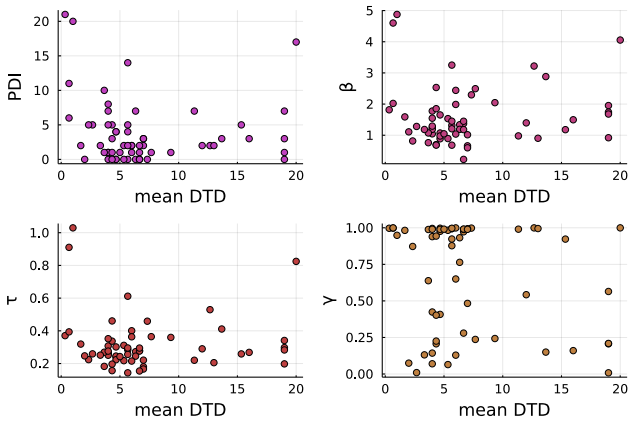

Figure 4.7: **Beads task data:** Scatter plots showing the relations (or rather lack thereof) among the observed mean draws-to-decision (DTD) values for each subject and the PDI scores and parameter estimates from the search-task of the same.

to less directed or random exploration, but to rather stem from greater "inertia", that manifested in a tendency to not move as much from choice to choice within a trial.

Our results go some way towards a convergence of evidence regarding the relation of information-gathering and delusional ideation. We show how modelling can help explain observed differences in behavior in terms of computational processes. While our data did not allow to detect differences in terms of sampling strategy, we did find differences in the confidence judgements, with our findings in line with previous work showing over-confidence [95, 96, 97, 98]. This fits with previous work. The confidence results (under-confidence in more and over-confidence in less informative) are consistent with *jumping-to-conclusions*. Deciding after having seen fewer samples in the beads-task is, from a computational perspective, similar to having high confidence about one's performance even though one actually has little information available.

In our sample, we did find reduced exploration as in the schizophrenia sample in [92], however through our model-based results we found this not to be a feature

of peoples' information-processing, but rather due to inertia, which rather has a motivational or resource-rational character. This might be seen as evidence against the continuum hypotheses of schizophrenia, or that apparent information-processing differences could be simply due to differences in motivation or processing capacity. A simpler explanation might be the relatively low PDI values of thee participants in our sample (see limitations below).

Further, the authors in [99] found people with schizophrenia to fail to appropriately model their opponents play despite consistent (rather than random) patterns that can be exploited in the simulated opponents play. This is manifest as a failure to weigh existing evidence appropriately against new evidence. Further, participants with schizophrenia show a jumping to conclusions bias, reporting successful discovery of a winning strategy with insufficient evidence. These were in line with our findings, where participant's reported high confidence in their finding the maximium, even in the "short horizon" condition, where they arguably had only little evidence available. Also, we potentially found under-generalization (as shown by the model family with $\lambda = 1$ winning), however this was also found in a healthy sample by [87].

4.4.1 Limitations

While we do find several interesting trends, our analyses are plagued by weak data and we are not able to gather enough evidence.

We used a online sample, and did not obtain a sample with enough participants scoring in the higher range of the PDI scale and did not collect further information about the participants (IQ, working memory etc.). Especially we recommend future investigations to employ screening experiments, in order to select samples with sufficient variance in the scores. Further, we collected data for 16 of the search task, which amounted to 16 confidence judgements by the participants. This

provided only limited statistical power and did not allow us determine why the over-confidence happens. Finally, the version of the beads task was not informative. Specifically, there were just three trials and we did allow subjects to choose up to 19 beads within a trial. Therefore, the relation of our observed beads task with the task behavior in the search task has to be interpreted carefully. Future work should use a newer version such as that from [65].

4.4.2 Conclusions

We have studied behavior in a searching task in relation with peoples' tendencies for delusional ideation and see clearest differences in metacognition. Generalizations requiring the search of conceptual spaces will be an exciting avenue for future work. Under the assumption that belief-formation corresponds to Bayesian posterior computation, the notion of search becomes important. In all but the most simple inference problems, the computation of the posterior or optimization of the likelihood function require searching of a hypothesis or solution space. For example, previous researchers have phrased theory formation as a stochastic search [8, 70], planning [100] and causal learning [101]. We suggest that these are the exactly the kinds of high-level faculties that are most-likely affected in people with delusions. Thus, it is important to understand how people search. This study provided a first step, using a task requiring explicit search of a spatial environment.

Chapter 5

Investigations of inference processes in delusional ideation

In this section I will investigate two different sorts of relevant inference processes. For each, I have developed a novel task and collected data from general population samples of participants recruited online. The first, described below in section 5.1, is a simplification the the task described in chapter 3. The second task, see section 5.2 is an extension of the classic motion perception paradigm based on random-dot-kinematograms (RDKs).

5.1 Rule-learning from binary cues

5.1.1 Introduction

In the literature on delusions, a major theme is the weighting between prior and likelihood in models of belief-updating in probabilistic inference tasks. However, how these findings generalize to more complex inferential problems is unclear. When we assume structural uncertainty, the space of possible hypotheses is large

and rational analysis shows that the optimal belief-updating cannot be a filtering procedure, but rather is a search through a hypothesis space, which must include recurring re-consideration of different underlying structures. Indeed, in such situations (that might be called reasoning problems) the problem are long-range, non-linear dependencies from data to beliefs that are largely unconstrained and thus hard to model. An example for such problems is rule-learning. In the chapter 3, we have described a model for rule-learning that may serve as a base for our investigation. Here, I aim to test probe several findings from the literature on delusions in a more ecological task that was completed by a general population sample.

Delusions have been described in terms of *abductive inference* [33], that is, inference to the best explanation. This type of inference is often contrasted with *deductive* and *inductive* inference. Deductive inference is typically associated with logical derivation from the general to the specifics or the scientific strategy of falsification. *Inductive* inference, which is learning about the general from particulars. This is often assumed to apply to Bayesian inference, where learning proceeds by starting from a prior distribution or belief, obtaining data, and updating the prior to the posterior distribution [102]. Still, the framework of Bayesian inference can also serve as a model of abductive inferences if we assume certain structured prior beliefs, that effectively act like explanatory preferences [103]. Thus, the "theoretical values" of explanations, which are data-independent, should be taken into account. These can be purely informational, such as simplicity and unification. There has been work showing that inferences of people are indeed sensitive to explanatory values [104, 105]. Apart from that, decision-theory predicts biases that are due to valuation, with choices leading to futures that are more favorable becoming more likely.

Interestingly, that seeking to explain certain data, rather than not trying to

explain them (i.e. verbally or in writing) in itself may bias people into find overly broad patterns, which can impair learning concepts that involve exceptions [106]. Previous researchers have shown developed simulations of rule learning through model-selection in the "Active Inference" framework [67]. The trial setup of the task that behavior was simulated for has been adapted for our empirical study of human rule learning behavior. Also relevant is work by Bramley et al. [101], that has led to the development of a model for online causal learning, which was fit to data of a task where the participants could choose causal intervention to learn the conncetions of certain causal devices. In their model Bramley et al. combined several approximations that allowed for efficient inference of causal structures. Their algorithm was representing a single hypothesis that, for each new trial was updated by running a Gibbs sampler for a number of iterations. Gibbs sampling is a Markov Chain Monte Carlo (MCMC) method, that allows to update a single variable at a time based on its full-conditional distribution. Asymptotically, the states visited by the resulting Markov chain in hypothesis space come to approximate the posterior over the causal system. By inclusion of a inverse temperature parameter, the algorithm could be brought to update in a more greedy fashion, which made the algorithm become like a search in hypothesis space of directed causal graphs. Our modelling approaches in this chapter are inspired by this and constructed similarly. Further, Bramley et al. weere considering multiple interacting factors in online causal learning: Prior beliefs about the system, approximate inference and active intervention selection. Testing the interplay of all these factors will be imporant in understanding delusions.

Here, we study a rule learning, with a hypothesis that is highly structured, and will model the inference as a incremental search procedure. We will test a weighting account of a primacy and recency effect where either early or late evidence might play a greater role in the computation of the posterior. This was

inspired by previous authors that showed that delusional ideation might relate to a tendency to form beliefs on the basis of early data [65, 107].

5.1.2 Binary rule-learning task

The task consisted of four blocks, where for each of these a different rule was to be learned by the participants. Every block consisted of repeated rounds, and each round consisted of two phases. In the "training phase", participants completed a number of trials, where they study the rule by repeatedly trying to predict the correct outcome for a presented configuration of context stimuli. During this phase, at the end of each trial, the participants were given feedback (correct/incorrect) and were asked to rate their confidences (whether they thought they understood the rule). Afterwards, in the "test phase", the participants were asked to give written descriptions of the rule and "act it out" by predicting the outcome for each possible configuration without feedback.

Figure 5.1 depicts the structure of the task. If a participant correctly predicted all outcomes in the "test phase", they would advance to the next block and start learning a new rule. Otherwise, the participant would complete another set of training and test phases. If the participant again does not have perfect performance on the test phase, they complete a third an final round before advancing to the next block. The rules are mappings from configurations of context variables to $\{\text{true}, \text{false}\}$ meaning that the rule applies for a particular configuration. There were three binary context variables. Generally, the rules to be learned in this task were such that they applied to roughly half of the sample space and thus could not be confused with a simple stochastic response biased to one of the two response options.

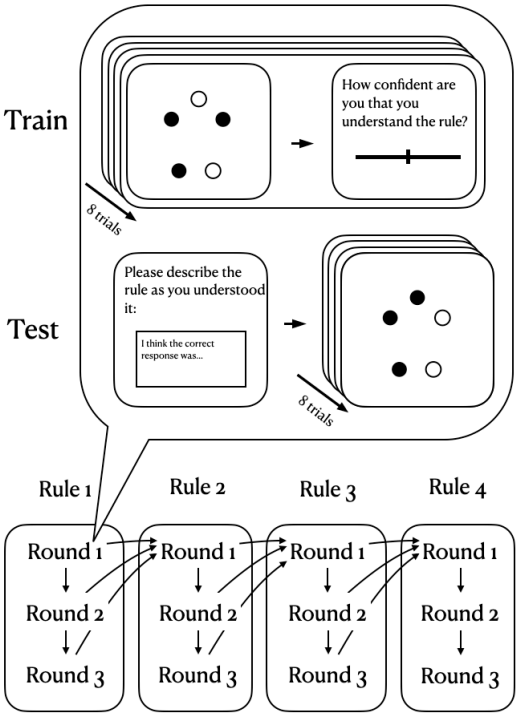

Figure 5.1: **Task and trial structure:** For the training phase, trials consisted of choice, feedback and confidence displays. After 8 trials, a test phase was completed. This procedure was then completed for 4 blocks and up to three rounds per block.

5.1.3 A model for binary rule learning

The problem formulation underlying the current task is similar to that described in 3.2. This model the approach of chapter 3, though here adapted to account for behavior in our round-based task. It is based on a Probabilistic Context-free Grammar (PCFG) prior over expressions that represent hypotheses about the hidden rule. The learner then incrementally computes a posterior belief about the hidden rule over the course of a each round (by searching the space of rule expressions).

The grammar is defined by the following *production rules*:

$$S \rightarrow \forall x\, \ell(x) \iff D \tag{5.1}$$

$$D \rightarrow D \vee D \mid D \vee P \mid C \tag{5.2}$$

$$C \rightarrow C \wedge D \mid C \wedge P \mid P \tag{5.3}$$

$$P \rightarrow F_1 \mid F_2 \mid F_3 \tag{5.4}$$

$$F_1 \rightarrow \text{false} \mid \text{true} \tag{5.5}$$

$$F_2 \rightarrow \text{false} \mid \text{true} \tag{5.6}$$

$$F_3 \rightarrow \text{false} \mid \text{true} \tag{5.7}$$

where $F_j,\ j = 1, 2, 3$ denote the three binary features that could either be "on" (white / true) or "off" (black / false). The intermediate symbols were S for start, D for disjunction, C for conjunction and P for prediate.

The expression could be evaluated with regard to the configuration of context cues in a given trial, a procedure that results in either true or false that would allow to predict the correct action (by mapping black and white to the predicted outcomes false and true). This provides a likelihood, which in turn allows us to compute a posterior over expressions. More precisely, the likelihood of a f given data $c_{1:t}, o_{1:t}$ is

$$\mathcal{L}(f|c_{1:t}, o_{1:t}) \propto \exp\{-b \cdot \sum_t \delta(o_t, f(c_t))\}, \tag{5.8}$$

which is equal to the number of falsely predicted outcomes, though allowing for outliers via scaling with the inverse temperature parameter b.

And the posterior was approximately computed with MCMC using Metropolis-Hastings proposals as described in chapter 3. The grammar defines a unique *parse tree* for every rule expression, that gives the sequence of choices for generating that expression from the grammar starting from the starting symbol S. This means,

that given a current hypothesis f_t, we can generate a "local move" in rule space by exchanging only parts of the parse tree that defines a given rule. For this, a new proposal would be generated by selecting a random node in the parse tree, and generating a random expression using the production probabilities of the grammar.

We tested the possibility that subjects learning showed an over-weighting of early information. This was operationalized by assuming a time-varying outlier probability $b(t)$, which we set as either exponentially or linearly increasing or decreasing. This has the effect of biasing the final posterior belief in the direction of early or late evidence for decreasing or increasing temperature schedules, respectively. We can represent the temperature schedule as a vector $b_{1:T}$, which leads to a simple adjustment of the likelihood:

$$\mathcal{L}(f|c_{1:t}, o_{1:t}) \propto \exp\{\sum_t -b_t \cdot \delta(o_t, f(c_t))\}, \tag{5.9}$$

The temperature sequences $b_{1:T}$ were defined as exponential functions:

$$b_t = \beta_0 \cdot \exp(-t \cdot \beta_{rate}). \tag{5.10}$$

Given a belief trajectory, we have to specify a mapping from beliefs to actions. To this end, we assume that the decision-making is making use of only a small number (5) of hypotheses from the posterior, and we thus choose the 5 hypotheses with the highest posterior probability and weight them by the same probability. Choices were modeled by taking a model-average over this weighted hypothesis set $\{w_j, h_j\}_{j=1}^5$and assuming a probability matching choice:

$$p(A_t = a|c_t, \{w_j, h_j\}_{j=1}^5) = \gamma \cdot 1/2 + (1-\gamma) \cdot \sum_j w_j \delta_{h_j(c_t)}(a) \tag{5.11}$$

where $\delta_{h_j(c_t)}(a)$ is a delta function located at the prediction of hypothesis for the

current trial evaluated at the actual choice (so 1 if they are equal and else 0).

Every belief update consisted of a run of the sampler and resulted in a posterior that was represented by a set of weighted rule expressions. We used the same order of stimuli for all participants, and thus only computed the belief trajectories once for all parameter combinations. To simulate predictions for the test phase, we would select a final policy and use this to make predictions for each of the test stimuli. The four rules that participants had to learn were representable in the "rule language" spanned by the above grammar as follows:

Rule	Formula	Meaning
1	$F_2(1)$	The second (middle) circle has value 1 (is white).
2	$(F_1(1) \wedge F_3(1)) \vee (F_1(0) \wedge F_1(0))$	The left and right circles match (both are 1 or 0).
3	$(F_2(1) \wedge F_3(0)) \vee (F_2(0) \wedge F_3(1))$	The middle and right circles do not match (are unequal).
4	$(F_2(1) \wedge F_3(1)) \vee (F_2(0) \wedge F_1(1))$	If the middle is white, the correct response is to match the right circle, while if the middle is black, the correct response is to match the left circle (i.e. when it is white, response white or else black).

5.1.4 Experiment

Participants

We recruited two sets of participants from Amazon's Mechanical Turk marketplace. One of these was including only "master workers", that had already completed a large number of tasks without rejections (indicative of their doing high quality work), and the other was selected from a larger screening study. However, due to poor performance the second sample was excluded from the analysis. This decision was based on inspection of the written descriptions of the rules of each block provided by the participants, which were deemed to not be intelligible, i.e. single

words or a repeated sentences that were reproduced from the task instructions and not referring to the stimuli. Further, anaylses of the mouse trajectories indicated that participants in that batch did not always sample the context stimuli before selecting their responses and thus were likely not following task instructions. Therefore, in the following, we report the results for the 15 participants that provided intelligible rule descriptions and consistently sampled the context stimuli of each trial. Participants additionally answered several demographic questions and the Peters et al. Delusions Inventory (PDI).

5.1.5 Results

Model-agnostic analysis

First, we looked at the performance. Figure 5.2 shows the average performance of all participants over each rule and round. For the first rule, which was only dependent on the value of one of the stimuli and therefore easier, we see clear learning with most participant's understanding the rule by the end of the third round. For rules 2-4 we can see some signs of learning on the group level, indicated by increasing performance over the rounds, but generally, performance was at chance level. This is also reflected in the proportion of correct responses during the test phases of each round, which are depicted in figure 5.3. For comparison, model predictions given different values for the inverse temperature parameter b are shown and, by visual inspection, the observed performance seems comparable to the model for relatively low inverse temperatures ($b \in \{\frac{1}{5}, \frac{1}{2}\}$).

Model-based analysis

For comparison of the observed behavior with the model, we computed the model predictions for parameter configurations chosen to lie on a grid. The belief tra-

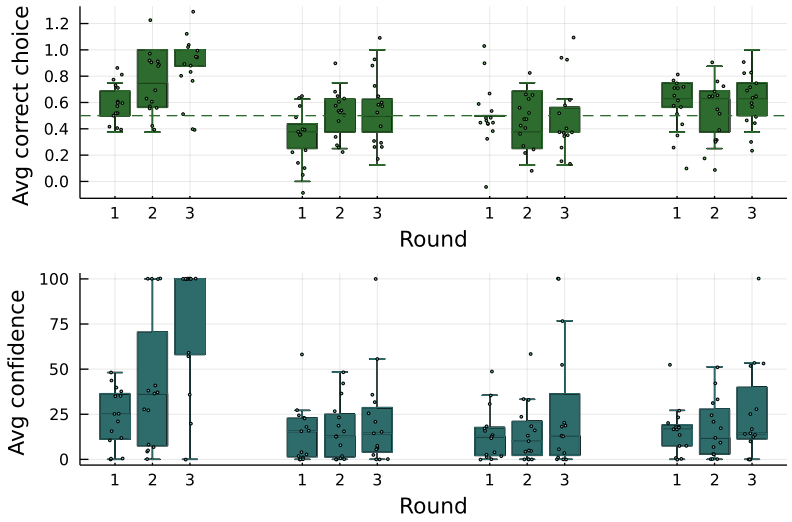

Figure 5.2: Upper row: Proportions of correct (rule-consistent) choices of all participants over the three rounds of each block. The dashed line marks 0.5, i.e. random performance. Lower row: Average confidence ratings of all participants for each of the the three rounds of each block.

jectories implied by different model versions were evaluated via the likelihood of producing the observed choices during the trials of a given round, as defined in equation 5.11.

We fit the model with different values for the parameters determining the weighting of early vs. late trials, β_0 and β_{rate}, and the lapse-rate η. Figure 5.4 shows, for each block, the differences in log-likelihood values of the best-fitting parameters and those from a purely random agent (i.e. an agent choosing uniformly at random). As can be seen, for the first rule, which was easier, most participants were best-fit by low lapse-rates. For rules 2 and 3, the bad performance of the participants is reflected in the lower likelihood values of the best-fitting parameters with higher lapse-rates (of up to $\eta = 1.0$) ocurring more often. In the fourth block, we again see higher likelihood values, but from the performance, and by inspection of the written descriptions of the rule provided by the participants, none of them

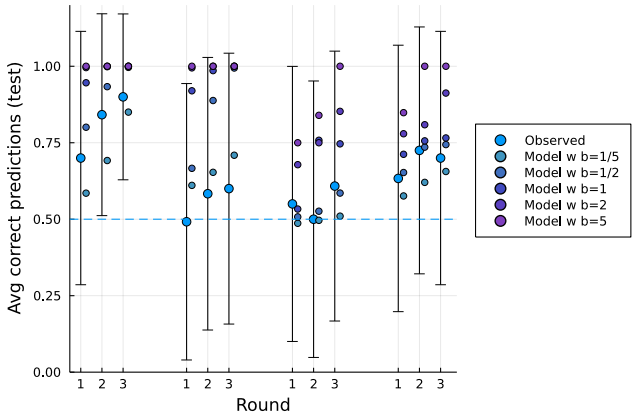

Figure 5.3: **Generalization:** Shown are the correct choices in the test phases, averaged over all participants for each round of each block.

found the true rule, but instead a heuristic that predicted the correct action on a portion of the configuration space.

Due to the indeterminate fits of model and observed responses in the other blocks, we restricted further analyses to the first block. First, we probe the relation of the estimated values of the β_0 parameter and the performance. We find a strong positive correlation ($\rho = 0.778$, 95% bootstrap C.I.: $[0.624, 0.987]$), that indicates that the higher inverse temperature b (and thus lower outlier probabilities) indeed predict better performance (at least for the first block) and that our rule-learning model can indeed account for the participants behavior in the task.

Regarding the hypothesis of over-weighting of early evidence, we compared the log-likelihood scores for the different model variants (with constant, decreasing and increasing temperatures) for all the participants. We computed correlations to quantify the relationship of the participant's PDI values with average confidence ratings and the estimates for the β_0 and β_{rate} parameters for block 1. We found no evidence for any relationships, but note that due to the low variability in PDI scores and the low number of subjects with adequate performance, our data does

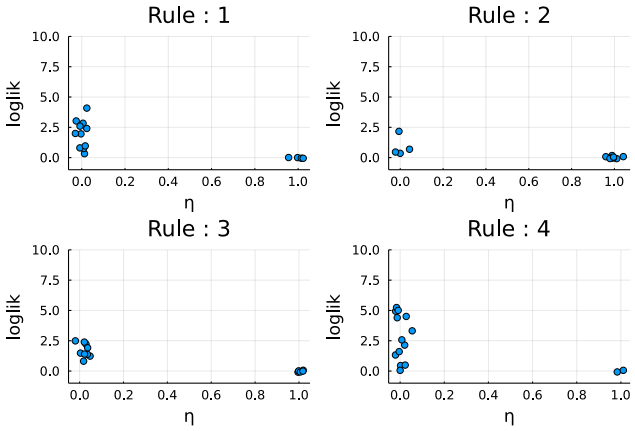

Figure 5.4: **Log-likelihood values vs. lapse-rate:** Estimated η values and log-likelihood differences (relative to random model) for best-fitting model for each participant and each rule to be learned.

not allow to draw conclusions regarding any population.

Mouse cursor tracking

During each trial, we collected the cursor position on the screening as the participants were sampling the cues and making their responses. The task was programed such that the context stimuli were hidden and had to be revealed, by hovering the mouse over them. We hypothesized that depending on their importances for a rule, participants might over the course of training pick up upon that and sample the features that are relevant for the rule more. Figure 5.5 shows the average over participants and trials within one round that a particular feature was visited. As can be seen, there were no differences in time spent over the course of a round. This might have been due to the participants not having learned the rule, although we see the same pattern even for the first rule.

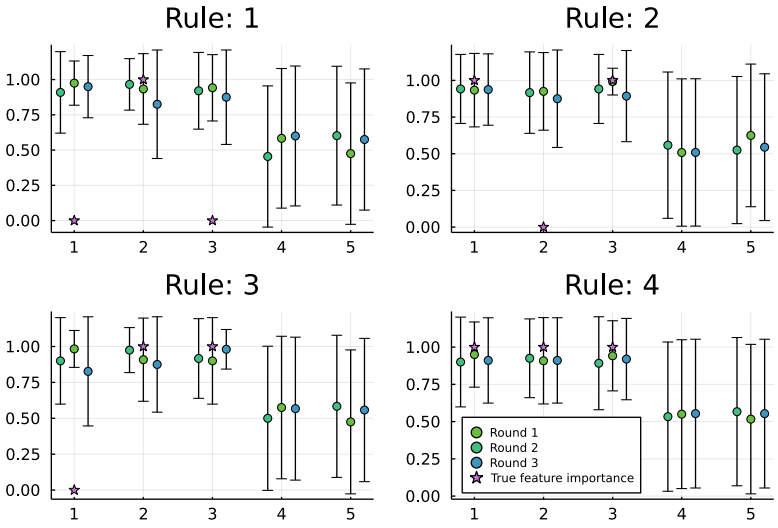

Figure 5.5: **Active sampling of context features:** The proportions of trials paritcipants sampled each of the areas of interest over each round and block: $1-3$ correspond to the three context features and 4 and 5 to the response options white and black. The star indicates the true feature importance (if it was included in the true rule expression of that block).

5.1.6 Discussion

We have tested the simplest version of a biased learning model that over-weights early experience during belief-formation. This hypothesis was taken over from prior work on the beads task and delusional ideation. However, the data from our sample did not provide enough information to test our hypothesis. What we can learn from this experiment? While I cannot conclude much about delusional ideation given the present data, I have developed a task and computational model that may provide a basis for future investigations. However, while the model could account for parts of the data – mainly those collected in the first block, where the rule to be learned was less complex – a crucial difficulty lies in the discrepancy between the model and our participants inferences. Here, due to poor performance in the blocks testing more complex rules, we could not be confident that our model

was descriptive of the inference process, and thus could not compare the hypotheses that were operationalized in the different paramterizations. Our intention was to pose an inference problem of sufficient difficulty to the subjects. This was achieved by giving the subjects a problem with a large hypothesis space, and where they were free to express their belief about the rule through their choices. The drawback of this freedom, however, was that the modelling of this decision becomes much more involved. Another major limitation here is that our sample did not show sufficient variation in PDI scores, with mostly low scores. The general point here is that recruitment of appropriate population is not trivial. The participants should vary in the tendency to delusional ideation while at the same time not be too compromised in their task performance.

For future work, there are multiple recommendations. Firstly, the sample should be balanced and contain sufficient participants scoring higher on the PDI questionnaire. Secondly, the task should be tuned to have appropriate difficulty. For this, one might need to perform online inference during the data collection and adaptively select the rules given to each participant. Another option is to include a training phase before the current task. In this phase, the participants will be shown a number of rules that are sampled from the grammar, which might increase the chance that they will have the relevant expressions available when trying to infer the rule during the task. Finally, the task itself was kept quite abstract, which might have been an impediment to learning. Future work will have to address each of these concerns to address the topic of rule learning. Studying rule learning is a challenging problem. The current work, I hope, may serve as a guide to potential pitfalls in studying this topic.

5.2 Evidence accumulation under model uncertainty

5.2.1 Introduction

In the previous section, we have investigated rule-learning as it occurred over a series of trials during an experimental task. In contrast, this section deals with evidence-accumulation and metacognition during inference on the time-scale of mere seconds, while following the general theme of inference under uncertainty about structure.

Perceptual disturbances in schizophrenia are associated with perceptual biases potentially involving stronger priors [108, 109, 110]. An important idea, the strong priors account seems to rather explain what happens "on average". For example, [65] found the decisions of schizophrenia patients to be explainable by a model with a prior over-weighting parameterization, They explained this as the result of people likely forming beliefs early on and then committing to them. This would be a different inference process altogether, that just shows prior over-weighting on average. A candidate for this, though happening on a much shorter timescale during evidence accumulation, might be premature commitments occurring. These would be of great interest to study in subjects with tendency to delusional ideation also to test the long standing notion of "jumping-to-conclusions". In work by Salvador et al. [107], the authors investigated evidence accumulation under ketamine-infusion, which is known to lead to NMDA receptor hypofunction, and while recording EEG. Under ketamine, participants displayed greater uncertainty and impaired inferences, but intact visiual processing. These effects were associated with unbalanced neural coding of the sensory evidence and premature response preparation. The authors showed through simulation that these results are consistent

with a "premature commitment" for a category, after which evidence is distorted in favor of the decision that the commitment was made for. Another work by Balsdon et al. [111] showed that people could be modelled as setting latent evidence threshold, in conditions where they could not control the amount evidence they would receive. They did not seem to set bounds for their confidence accumulation and thus showed a dissociation between the two processes. In their model, perceptual evidence and confidence are accumulated in parallel, and observers commit to perceptual decisions (when a threshold was crossed), but would continue to accumulate sensory evidence for evaluating confidence. The same authors followed up on this work with an EEG study [112]. Here, they found additional evidence for premature decision commitments, provided by spectral power patterns in motor cortex and, again, an attenuation of the coding of perceptual evidence. They also found a distinct neural representation, localised in the superior parietal and orbitofrontal cortices, that was associated with suboptimal confidence reports.

Another recent study by Rollwage et al. [113] employed a two-decision task, where the subjects made an initial decision and confidence rating and then could revise their decision after a second presentation of sensory evidence. They observed that when people had high confidence in the first decision, their neural processing was modulated with confirmatory evidence being amplified and dis-confirmatory evidence reduced. Similarly, in the field of motion perception, Stocker & Simoncelli [114] proposed an account of biased, "conditional" perception, where the evaluation of sensory evidence is conditional on certain decision-induced biases. This line of work has been followed with results indicating a bias in evidence-accumulation, where when subjects commited on a decision initially, this decision lead to a down-weighting of contradictory evidence subsequently [115, 116, 117].

Very relevant to the current work is a very recent study by [118]. The authors used a random dot kinematogram stimulus, which on half of the trials had the di-

retion of coherent motion change by 90° after the first half of the presentation time. The participants were asked to report the direction they perceived at the end of the trial, and patients with schizophrenia tended to stick with the initial direction more often in trials where there was a change. Evidence accumulation is widely studied. The standard approach is to present subjects with noisy evidence for one of two response alternatives. Models of the evidence accumulation are based on drift-diffusion models. Here, we go beyond existing paradigms, and study evidence accumulation under structural uncertainty. Intuitively, uncertainty about the underlying structure increases total uncertainty. However, similarly to the existence of explanatory preferences in explanation finding [103], there may similarly be explanatory preferences influencing early perceptual processing. Previous research has shown that even in early preceptual processing, the latent structure of the inputs play a role [119, 120].

Under structural uncertainty, priors have greater complexity beyond being stronger. They may prefer different *kinds* of structure. Therefore, we have developed a task based on Random-dot-kinematograms (RDK), where there are potentially multiple drifts underlying the movement of a could of moving dots. Perceiving the underlying drifts requires to compare different hypotheses, such has about the number of drifts present, and balance detection of weak drifts with correctly rejecting spurious drifts that are due to noise.

5.2.2 Experiment 1

Task

The task consisted of a training phase consisting of 20 trials, followed by two blocks of 50 trials each. Figure 5.6 depicts the structure of a single trial. Each trial consisted of of the presentation of a RDK with potentially multiple underlying

drifts for 4.0 seconds (i.e. 240 frames assuming a framerate of 60Hz). Afterwards, in the response phase, the participants selected the directions they detected, using a custom interface. Additionally, participants could indicate their confidence for each given response given a confidence slider. Participants received points for each correctly detected response, and additional points for appropriately setting their confidence. The mechanism to determine the bonus points was by drawing a random number and checking whether it fell to the left or to the right (and whether the direction was in fact correct) of where the participants placed the slider. During the stimulus presentation phase, from frame to frame, each dot was assigned to be either "coherent", following one of the underlying drifts (according to pre-specified proportions), or heading into a random direction. Dots that exited the stimulus area were re-inserted at random locations within the display. For the first block, the true number of drifts was given at the start of each trial, while in the second block, no such information was given and the subjects had to additionally infer the number of drifts. Unknown to the participants, the stimuli shown in the second block were created with drifts with non-uniform weights, while the first block always had uniform weights.

Model

Due to the latent structure of the stimuli, the inference problem in this task cannot be solved by simply summing up the log-likelihoods for every direction, especially when the number of present drift directions is unknown. To do so, one would still need a criterion or threshold to use to select which ones to choose. This is a model selection problem that might be solved optimally by computation of the model evidence (or integrated likelihood), but that would require computation (and representation) of all possible models which is infeasible. Here, I follow a different approach, based on a generalization of the drift diffusion model for binary

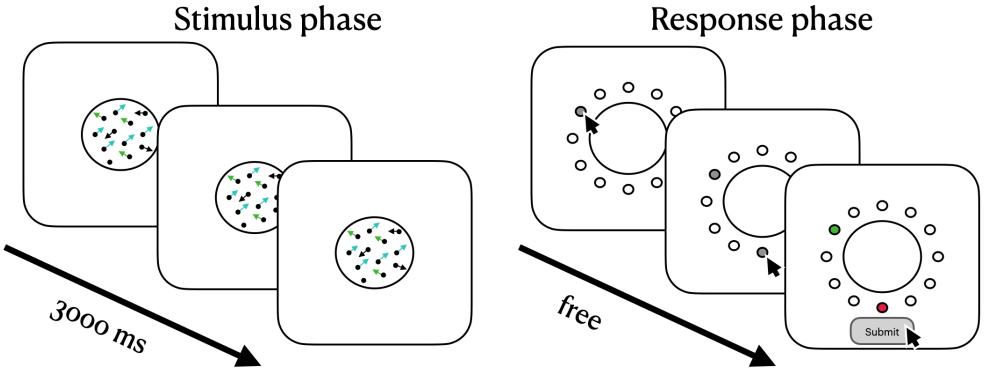

Figure 5.6: **Trial structure:** Each trial consisted of a stimulus presentation and a subsequent response specification phase. During the presentation phase, a proportion of dots moved coherently or randomly. For the coherent movement, the dots could follow multiple directtions, depending of the structure of the stimulus. In the response phase, subjects selected all directions they believed to be underlying the stimulus.

perceptual decision-making. The drift diffusion model assumes that the belief performs a random walk that is biased in the direction of the true category by the percpetual evidence. Similarly, in this work I model the evidence accumulation as a biased random walk, though not on the real numbers, but on the space of possible structures underlying the stimuli, which are represented as mixture models. Concretely, I represented the different drift directions that may be present the indicators $\mathcal{I} = (I_1, \ldots, I_{12})$ and model inference as the approximate computation of a posterior belief over these indicator variables. This approximate computation is assumed to be performed by Markov Chain Monte Carlo (MCMC), which leads to a random walk biased towards the exact posterior belief implied by the evidence.

The observations provided by the motion stimulus are the frame-to-frame displacements of the dots. These displacements can be assumed to follow a bivariate Normal distribution, with the possible drifts corresponding to the 12 directions available in the response interface. The possibility of multiple drifts results in a mixture likelihood. Together, one can write the generative model of the stimulus

as follows:

$$b_0 \sim Beta(1,3) \tag{5.12}$$

$$I_j \sim Ber(b_0),\ j = 1,\ldots,12 \tag{5.13}$$

$$\gamma \sim Beta(1,3) \tag{5.14}$$

$$p(\mathcal{O} = 1|I) = \exp(-4 * \sum_j \left[I_j \in \{1,2,3,4\}\right])) \tag{5.15}$$

$$z_{i,t} \sim Cat(\frac{(1-\gamma)\cdot I_1}{12},\ldots,\frac{(1-\gamma)\cdot I_{12}}{12},\gamma) \tag{5.16}$$

$$y_{i,t} \sim \begin{cases} MVN(\mu_{z_i},\Sigma_{z_i}) & \text{if } z_{i,t} < 13 \\ \mathcal{U}_{(-2,2)\times(-2,2)} & \text{else} \end{cases} \tag{5.17}$$

where the Iverson bracket $[\cdot] = 1$ if its argument is true and 0 otherwise, and the $\mu_k,\ k = 1,\ldots,12$ are defined as $\mu_k = (\cos(\theta_k),\sin(\theta_k)),\ \theta_k = \frac{(k-1)\cdot 2\pi}{12}$ and assumed known by the subjects. The covariance matrix is defined $\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$. The z variables are indicators for each observation indicating which underlying drift it was generated by. We denote the outlier probability by γ and include a 13th "outlier" component in the mixture, that generated displacements from a uniform distribution and thus accounted for incoherently moving dots (of which there was a fixed proportion). We also include a prior belief about their own action $p(\mathcal{O} = 1|\mathcal{I})$, that effectively penalizes beliefs that include more than four directions and that will be conditioned on for the posterior computation. Furthermore, to model the fact that subjects knew the correct number of drifts K_{true} during the first block, the generative model was modified with an additional term:

$$p(\mathcal{K} = 1|\mathcal{I}) = \exp(-4 * (\sum_j I_j - K_{true})^2) \tag{5.18}$$

that inference process was conditioned on and which lead to a penalization of deviations from the true number of drifts and thus helps steer the search towards

the correct belief.

We do not assume that our subjects are able to visually follow or even register all of the moving dots, and thus do not estimate the z for every time point. Instead, we assume that subjects compute the likelihood via the following expression:

$$\mathcal{L}(y_{\cdot,t}|\mathcal{I},\gamma) = \sum_j \frac{(1-\gamma)\cdot I_j \cdot p(y_{i,t}|\mu_j,\Sigma_j)}{K} + \gamma \tag{5.19}$$

where $K = \sum_j I_j$. The approximate computation of the posterior is implemented as a Metropolis-Hastings MCMC algorithm. We randomly draw an initial hypothesis, which represents the belief without any uncertainty. Then, we randomly propose local changes: the removal, addition, shift or splitting of any element of the current set of drifts that form the belief or the merging of two elements. When iterating these local adjustments, this the sequence forms a Markov chain that has as its limiting distribution the correct target, that is the posterior over the indicators $\mathcal{I}$. The choices are assumed to be generated by taking the last visited state of the chain and choosing the directions for which $I_j = 1$. This can also be seen as a particle filter with a single particle.

Our model of the inference process is based on a forward model and, while optimal in the sense that it solves the inference problem faced by the subjects, it does not permit the numerical evaluation of a likelihood function. One approach to deal with this is Approximate Bayesian Computation (ABC). Here, data summaries are chosen, and then parameters can be inferred by comparing the summaries from simulated data with those that were observed. Here, we avoid the use of summaries as this means a loss of information and instead we make use of the earth mover's distance (EMD) to approximately evaluate the likelihood. Intuitively, considering two distributions as two different ways of piling a fixed amount of earth, the EMD is the minimum cost of turning one pile into the other, where the cost is defined

as work: the amount the probability mass moved times the distance by which it is moved. This gives a distance measure that deals with multi-model distributions, while also taking into account the neighborhood relations between the clusters. For the current task, it is useful as it allows to distinguish a near miss from a clear false positive. In order to obtain an approximate likelihood we want to compare the simulated posterior with the observed responses on a given trial. We can represent both of these via their signatures, that is, discrete measures that assign probability mass to the twelve possible directions $\theta_k,\ k = 1, \ldots, 12$ of the underlying drift and thus summarize the structure of a stimulus (and response configuration).

More concretely, for two discrete measures $P = \{(p_1, w_{p1}), \ldots, (p_{12}, w_{p12})\}$, and $Q = \{(q_1, w_{q1}), \ldots, (q_{12}, w_{q12})\}$, the optimal flow γ^* that transforms signature P into Q, where $\gamma_{i,j}$ denotes the amount of mass transported from bin i in P to bin j in Q. It is computed by solving the following linear optimization problem:

$$\gamma* = \underset{\gamma \in \mathbb{R}_+^{m \times n}}{\operatorname{argmin}} \sum_{ij} \gamma_{i,j} d_{i,j} \tag{5.20}$$

$$\text{s.t.} \gamma 1 = 1,\ \gamma^T 1 = 1,\ \gamma \geq 0 \tag{5.21}$$

where $d_{i,j}$ is the distance between clusters p_i and q_j, here defined as $d_{i,j} = 1 - \cos(\theta_i - \theta_j)$.

Results

First, I look at the performance, this is quantified via the earth mover's distance from the participants' responses to the true underlying structure of a given trial. The panel A of figure 5.7 shows the distribution of errors in blocks 1 and 2. As can be seen, performance dropped for the second block. We fit psychometric curves to the responses for each individual. For this model, all directions are considered to be independent, an assumption that is clearly false, since the subject knows that

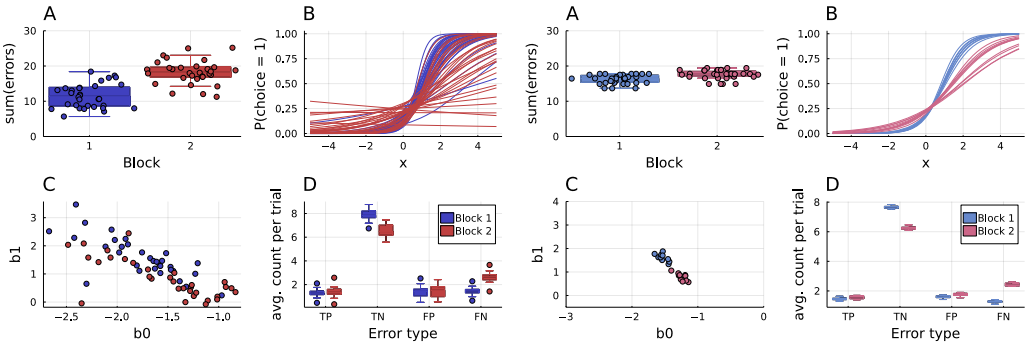

Figure 5.7: **Observed and simulated behavior:** The left side shows the observed and the right side the simulated responses. Panel A shows the distribution of errors for blocks one and two and each subject. Panel B shows logistic curves fit to each participants choices (see text) and panel C shows the corresponding coefficients. Panel D shows the average counts for the occurance of different errors: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

there can only be a maximum of four directions. Still, one can derive a measure of sensitivity from this analysis. This allows us to see differences between the two blocks. Panels B and C in figure 5.7 shows the fitted curves and corresponding estimates for base-rates (b_0) and sensitivities (b_1). We can see a negative relation (panel C), and a lower sensitivity in the second block (panel B). Furthermore, panel D shows a significantly higher false negative rate in the second block. Note that the second block did differ in multiple regards. Firstly, the true number of underlying drifts was unknown to the subjects. Secondly, each trial had four drifts with unequal weights. We compare evidence for both alternative by simulating our model for both blocks, either with or without knowledge of the number. We encoded this knowledge via a prior belief, that pulled the Markov chain towards the correct number of drifts. We imagine a similar proces within our participants as they remind themselves of their knowledge shortly before they were asked to enter their responses.

To further investigate the drop in performance using our model. For our mod-

elling approach, we simulated the model with the same parameters for both blocks, but with the difference that we encoded knowledge about the true number of latent drifts. The right half of 5.7 shows simulated data from parameters chosen to match the observations. This simulation assumes the same processing capacity (search length or number of iterations of the Markov chain model), perceptual accuracy (level of noise assumptions) and temperature (cooling) schedule. Figure 5.8 shows, for a couple example trials the proportion of responses of subjects against the proportions calculated from responses that were simulated from the model. As can be seen, there is at least a qualitative correspondence. With regard to the PDI

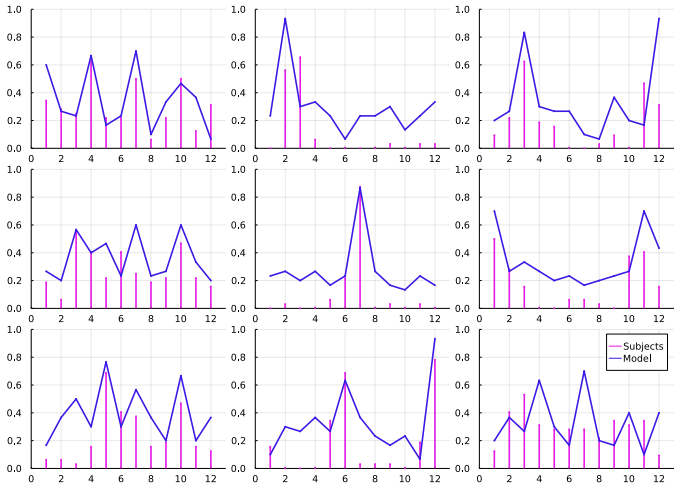

Figure 5.8: **Observed choices and model:** Shown are, for a number of example trials the proportion of subjects that chose each of the 12 possible response options and the posterior distribution over these of our model.

scores of the participants, I found that PDI did not relate to the differences in performance between blocks or to the types of errors. The variation of PDI scores in our sample was rather low, with a mean of 4.4 and a standard deviation of 3.6.

5.2.3 Experiment 2

While the first experiment has given us some idea of how people perform the task, its static structure with a single response did not allow us to identify biases in the inference process that might relate to early commitments towards certain decisions. In the second experiment, I allowed subjects to control the presentation of the stimulus. That is, the stimulus animation advanced only while participants were holding down the left mouse key. In between these "observation phases", the participants were instructed to adjust the response interface (identical to the one in experiment 1) as soon as they detected a discrepancy between the currently set response and the correct response as implied (noisily) by stimulus. The aim was, to observe how, firstly the belief develops over time, and secondly, how having made certain responses influences subsequent evidence accumulation. One hypothesis, inspired from previous accounts would be that people show a bias towards sticking with their own choices.

Participants

We recruited 30 participants online via Amazon's MTurk marketplace. Of these, 13 participants were rejected due to failure of attention checks or performances indices indicating behavior not distinguishable from random choice. Participants completed the task, several demographic questions and the Peters et al. Delusions Inventory (PDI).

Task

A trial consists of the specification of an initial guess by the subject and then interleaved phases of evidence accumulation and response adjustment. The subjects could start and stop the playing the animation and thus control the amount of

evidence receive before adjusting their response. The subjects were instructed to play the animation til the end once they were sure of their response.

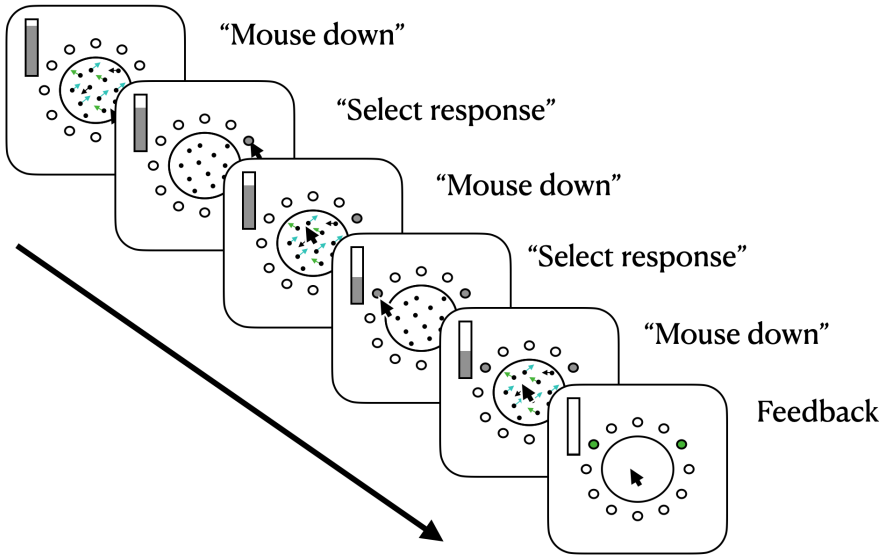

Figure 5.9: **Trial structure of dynamic task:** A trials consisted of a dynamic interface, where the subject could alternate between collecting data from the stimulus (by clicking and holding down the left mouse button on the stimulus aperture) and adjusting their response (by clicking the empty circles around the stimulus display representing the twelve possible directions) to match the structure underlying the stimulus. Once, the time budget, shown as the grey rectangle was exhausted, the trial ended. The budget was only decreasing while the subject were collecting data from the stimulus.

The structure of the task was similar to that of Experiment 1 above. Subjects underwent a number of training trials, where the true response was given to them and they simply had to reproduce it. After the training, there were two blocks of 20 and 30 trials, respectively. As in experiment 1, in the first blocks, the correct number of drifts was displayed and the correct response was shown at the end of the trial. In the second block, the number was unknown to the participants and they were not given feedback, but instead were asked for confidence ratings regarding their response: "how close do you feel you were to the correct response".

Results

Firstly, I replicated the results from experiment 1. For this, I analyzed the responses given by the participants at the end of the trials. Figure 5.10 shows the same statistics as computed in experiment 1. As can be seen, the performance showed a similar pattern, with worse performance in the second block.

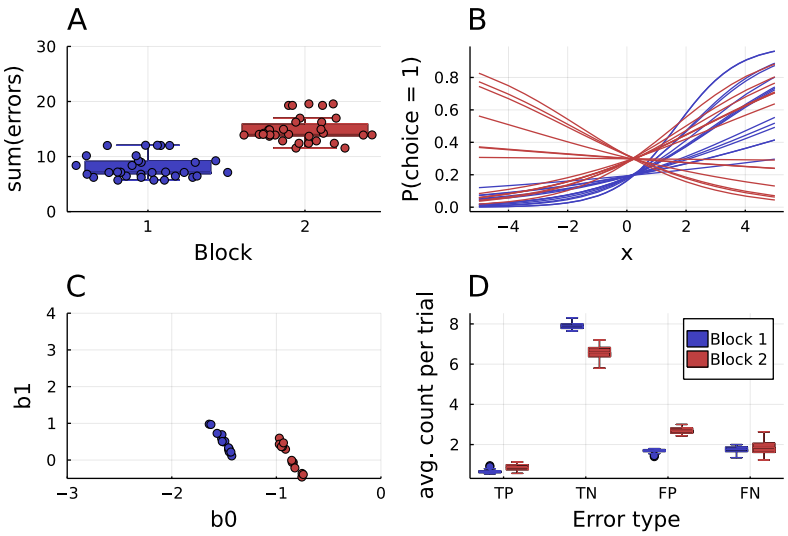

Figure 5.10: **Observed behavior:** Panel A shows the distribution of errors for blocks one and two and each subject. Panel B shows logistic curves fit to each participants choices (see text) and panel C shows the corresponding coefficients. Panel D shows the average counts for the occurance of different errors: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

Secondly, I looked at differences in behavior associated with the participants PDI scores. One hypothesis was that we might find choice-consistent evidence reweighting. This would show in a suboptimal results of inference. Figure 5.11 shows the evolution of the error of each participant's responses over time for all trials of the second block. Which each adjustment of their response, the error (distance to true stimulus structure) changes. One can see the different speeds at which people adjust their response, and how optimal they are, with some participants showing

trajectories that do not decrease in error towards the end.

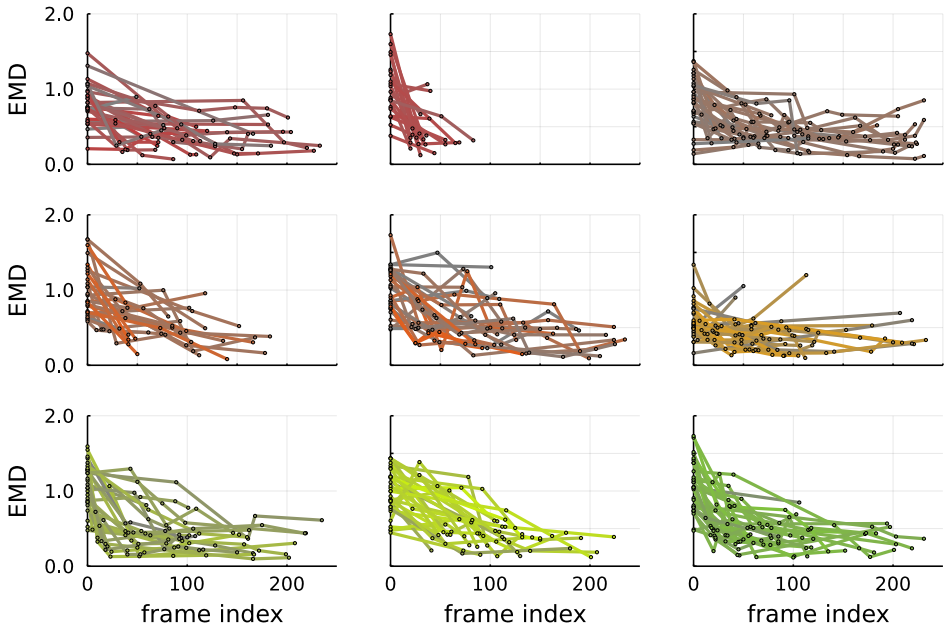

Figure 5.11: **Temporal evolution of choices:** Shown are the distances between choices and stimuli (measured by the EMD) over the course of a trial for the 12 subjects, ordered increasingly according to their scores on the PDI questionnaire (values given in text). Every circle indicates an adjustment of the response by a subject and each line connects the adjustments in a trial. On the x-axis is the index of the frames when adjustments where made (up to the maximum of 240). The saturation of the color of any single chain was mapped to the confidence judgement of the corresponding trial.

The panels of the 5.11 are sorted by PDI score (also indicated by color with red indicating low score (0) and green high). After exclusion, either due to poor performance or failed attention checks, only 9 subjects remained, and of these most scores were low (the collected scores were $(0, 0, 2, 2, 2, 4, 7, 7, 9)$; and they correspond to the panels of the plot from top to bottom and left to right). The color saturation of the lines corresponds to the confidence judgements of that trial, and what is evident from the figure is that the subjets with higher scores tend to have higher confidence. This is supported also by a positive correlation of PDI score with

average confidence ($\rho = 0.51$ boostrap C.I.: $(0.15, 1.0)$). However, given our small sample I could not address our hypotheses regarding the relation of PDI scores and behavioral differences in our task.

5.2.4 Discussion

Over two experiments, I have investigated evidence accumulation under structural uncertainty. I have developed a model of incremental inference, that was able to qualitatively account for observed behavior. While I have not yet been able to find and characterize specific differences in evidence-accumulation between groups of people with low and high PDI scores, I have taken the first steps in developing and validating novel tasks and analyses and found prelimiary evidence for a relation of PDI and over-confidence in experiment 2. Advantages of this tasks are that the incoporation of multiple directions keeps the stimulus interesting and allows for the plausible inference of false positives. A disadvantage is the difficulty of inference, due to a lack of a closed-form likelihood and the flexibility of the choice interface which leads to greater complexity, which can only be tamed by careful modelling. One remaining problem is that our theory does not lead to strong assumptions about the subjects inferential computations, which could lead to predictions about the timings of the choice adjustment in the dynamic task of experiment 2.

One goal of designing the task presented here was to test our model from chapter 2. However, we were not able to test this, for two reasons. Firstly, by restricting the space of possible directions to the 12 pre-determined response options, it was implausible that participants would generate highly precise and spurious causes and thereby maintain an initially formed belief. Secondly, due to the sampling-based inference process, and weak determination throught the task, we were not able to recover the expected precision hyper parameter. In future iterations, the design should be based on a-priori simulations in order to ensure that differences

in the prior beliefs about the precision of causes matters and is recoverable from task behavior.

Another goal was to explore potential differences in inferences about structure during evidence accumulation. Given our more involved task, it was easier to detect inattentive or other forms of responding that indicate that participants completed the task in a way that was not intended. This should be considered an advantage of the task. However, the testing of our hypotheses regarding delusional ideation have been stymied by the exclusion of many of the subjects initially recruited and illustrates the the difficulty of obtaining online data. Our problem is made worse not only by the want to sample a specific population. In future work, we aim to improve the modelling of behavior in the dynamic task of experiment 2 by casting the problem as a POMDP. Further, an interesting avenue would we to combine these tasks with functional imaging such as M/EEG techniques and decoding techniques.

Chapter 6

General Discussion

In this thesis, I have aimed at developing a formal grounding for the study of the processes supporting human (mis-)belief, and, in particular, how they might be altered in delusional ideation. What has been achieved is only the very first step. Further work will need to more fully develop the theoretical framework, and improve upon the early versions of experimental paradigms that may ultimately allow to empirically test it. Compared with previous conceptions of delusional ideation, we have provided a quantitative account. It seems right, at least intuitively, to frame this in terms of the learning of a world model.

In *chapter 2*, we have laid out a framework that allows to describe delusional inference within the context of an open-ended process of learning the latent structure of the world, based on the Dirichlet process. This is a Bayesian solution to the structure learning problem, has been used as a prior in various approaches to structure learning, and has been considered as a model for prefrontal cortex function [52]. Given the context, which allows for an infinite number of latent causes acting in the world, the agent may not only form a delusional belief in response to a suspicious coincidence, but may also abuse the ability to "think up" new explanations in order to explain-away events that may otherwise lead to a weakening

or disconfirmation of some existing belief.

While this framework was described and simulated under simple Gaussian assumptions, *chapter 3* dealt with the problem of learning within highly structured hypothesis spaces. Central to this was the assumption of a higher-order generative model, a probabilistic context-free grammar, that spans the space of hypotheses. Learning can then be achieved by searching this space through a process of making local adjustments (based on Metropolis-Hastings Markov Chain Monte Carlo methods). This grammar-based approach can allow for modelling hypothesis generation, and potentially explanatory biases.

In *chapters 4-5*, we report several empirical project dealing with inferential processes in general population samples with varying levels of delusional ideation. Each of these deals with inference under ambiguous structure, and requires inferences in large hypothesis spaces. In *chapter 4*, we probed an explicit searching task, with choice options being cells in a rectangular grid. We found that the behavior of people with higher scores on a delusional ideation questionniare were described by a model with less exploration, possibly due to greater inertia, indicating a tendency for taking very small steps in the searching process. Further, they showed differences in metacognition, with a tendency for over-confidence and weakened sensitivity for the amount of information they had gathered during their search.

In *chapter 5*, we explored two novel tasks that required subjects to learn latent structures. The first section dealt with a rule learning task, that was developed on the basis of the account presented in chapter 3. There, an explicit rule expressed in a formal language (modified first-order logic) that described the relation between a set of context stimuli and the correct response was to be learned. Our model was based on the account of chapter 3, though adjusted for binary context variables and the block structure of the task. On the whole, in chapter 5 we were not able

make strong inferences. This was both due to difficulties in collecting a sample that showed good understand of the task while at the same time scoring in the middle and higher ranges of the "Peters delusions inventory" questionnaire.

The current work has only scratched the surface of the phenomenon of delusional belief. What are the next steps for the future of a computational characterization of the phenomenology of a delusional world model? Future work will not only have to empirically test the accounts laid out in this thesis. An obvious remaining project is to combine the concepts of chapters 2 and 3. The empirical validation will require further specification and adaptation of the framework to specific contexts or tasks, and their combination with imaging and pharmacology. Combining the concepts of chapters 2 and 3, we can envision an agent learning explanations for the events in its environment that are represented in a grammar-based language of explanation.

This ties in with the rationalist account of Campbell [121], that regards delusions as having the epistemic status of beliefs expressed by Wittgensteinian "framework propositions". Such framework beliefs are not ordinary beliefs about facts, but instead form the background for any inquiry into truth or falsity. That is, what confirmation vs. disconfirmation of a hypothesis is defined within the space spanned by the framework beliefs. In inferential terms, such beliefs are hyper-priors in a hierarchical model that all updates of lower-level beliefs are conditioned on. They specify the model space, which is searched in order to explain any observed event. In the Bayesian framework, hyper-priors can be updated, but this needs vastly more data than updates of lower-level beliefs that are more closely tied to observations. In this view delusion formation may be seen as Kuhnian paradigm shifts, that is, they represent departures from the previous framework of knowledge, with terms changing their meaning and new terms and rules becoming available for use in the composition of new hypotheses and explanations. This view also

helps to understand how social isolation can result from a breaking of communication: when sense-making by different people is based on different frameworks, the same terms come to mean different things and it can become impossible to find a common ground.

With a model available of how systems of framework beliefs form and evolve, one could test hypotheses about the conditions for the learner to develop a maladaptive explanatory framework that, while fundamentally wrong, somehow does account for the agent's experiences, or might be maintained for other reasons (that will have to be more precisely specified then). For example, one may compare a motivated reasoning account vs. a purely inferential one, or one-factor against two-factor accounts. Simply the attempt to formalize concepts can help to structure future investigations.

Although it may be difficulty to estimate an individual's grammar purely from behavioral data, this could be possible through the combination of multiple tasks. By collecting a wide range of information through multiple tasks about single subjects, we may come to understand how explanations are constructed, much like expressions in a language. Further, in combination with imaging and pharamcological interventions, we have the everything we need to understand the phenomenon. For example, the authors in [122] studied changes in the semantic associations in patients with panic disorder. They were able to measure the effects of CBT on the associations between panic-triggers and symptoms words through a semantic priming paradigm in the scanner. Similarly, one might be able to investigate and modify the semantic networks that patients with delusions search through during inference of delusional explanations. Here, the crucial difference from non-patients may not only the associations of certain triggers with symptoms, but rather the way (the algorithms) by which explanations are crafted. These may be found to often follow certain schemas, which can be conceived of as the priors of sturctured hypotheses,

or grammars-of-explanation. Based on this, new meta-cognitive treatments may be developed that target specific biases and allow to update the grammar, lowering the availability of delusional interpretations.

Delusions likely defy simplistic etiological or mechanistic descriptions [123], and are due to multiple causes and involving multiple levels of description [124]. Therefore, it must be kept in mind that our considerations of the cognitive processes still need to be made to fit with the whole phenotype of psychotic illness. Cognitive decline, working memory deficits and thought disorder ought to play a role in any account of delusion in the context of schizophrenia. Other aspects, such as cultural transmission and communication in networks and the view of beliefs as social signalling instruments should be considered in the context of conspiracy beliefs. The complex interplay of many these factors might be what characterizes delusional beliefs. Here, we have only considered a single node in the network of factors, which still contains a lot of complexity of its own.

Conclusion

Understanding delusions, and mis-belief more generally, is an important scientific goal. Clearly, this task is not easy, but progress is possible. Treating delusions as beliefs and studying them using formal, computational approaches while building on recent findings from neuroscience is the way forward. This view is on the rise, replacing previous views that held that delusions did not emerge from the same psychological processes involved in normal beliefs and were thus un-understandable and that hindered progress. Proponents of the older view included Jaspers [2], one of the fathers of modern psychiatry, who argued that primary delusions were essentially inexplicable and psychologically irreducible. Similarly, Fodor [125] suggested that beliefs were not suitable to scientific study in his so-called: “First Law of the Nonexistence of Cognitive Science”. His argument was based on the distinc-

tion of “input systems”, that is, modules that perform specific function without interference, and and “central processes”, that operate on the input modules, such as reasoning and belief. Fodor held that scientific investigation was only possible for cognitive processes that are like “input systems”, that central processes would resist scientific study and analysis, as they might be not be decomposable into simpler parts. Today, these views are clearly falsified [126], with much progress in decision-making and higher-order cognition both in cognitive science and neuroscience, although much work remains to be done. While delusions do indeed involve many factors and are clearly messier that visual perception, this does not mean that they cannot be understood. There is no reason why we will not understand them eventually, as our conceptual tools and theories evolve. Formalization and theory building will be central and this thesis has aimed at taking a step in this direction.

The way humans construct their realities is better understood as first-person story-telling than as objective logical derivation [127]. In future work, by combining models of world learning with grammar-based priors, we may be able to study the belief-based mechanisms of reality construction. An important role, that has heretofore been neglected, is played by the process of hypothesis generation. Its basis are generative models and priors that can represent structure, and thus serve as models for the idiosyncratic structures of the world view of an individual.

Appendix A

Details of model and inference algorithm for chapter 2

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

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

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

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

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

$$\mu_k | \mu_\mu, \tau_\mu \sim \mathcal{N}(\mu_\mu, \tau_\mu) \tag{A.4}$$

$$\tau_k | \mu_\tau, \tau_\tau \sim \mathcal{HN}(\mu_\tau, \tau_\tau) \tag{A.5}$$

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

$$y_i \sim \mathcal{N}(\mu_k, \tau_k),\ i = 1, \ldots, n. \tag{A.7}$$

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

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

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:

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

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

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

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

and temporary candidate parameters for the m new components drawn their respective priors $\mu_k \sim \mathcal{N}(\mu_\mu, \tau_\mu)$ and $\tau_k \sim \mathcal{HN}(\mu_\tau, \tau_\tau)$, $k = K < k < K + m$.

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 $\mathbf{z}$ and the cluster parameters ϕ according to Algorithm 8 in [53].

Simulation details

For the simulations for Figure 2.3, 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 $\phi_j^* \sim \pi(\mu_k, \tau_k | \mu_\mu, \tau_\mu, \mu_\tau, \tau_\tau)$ 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 [53]. 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^*, $i = 1, \ldots, 50$ 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 re-sample the cluster labels according to the algorithm detailed above. The simulation was performed with the following hyperparameter settings: $\mu_\mu = 0.0$, $\tau_\mu = 1/10$, $\tau_\tau = 10$ and with the prior only differing for $\mathcal{HN}((\mu_\tau^{(j)}, \tau_\tau))$, 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.

Appendix B

Definition of Context-free grammars for chapter 3

A context-free grammar is defined by a 4-tuple (V, Σ, R, S), with V a finite set of *variables*, Σ a finite set of *terminals*, R a set of *rules*, each of which consist of a variable and a string of variables and terminals, and $S \in V$ is the *start variable*. One can use a grammar to describe a language by generating strings of that language in the following manner.

1. Write down the start symbol.

2. Find a variable that is written down and a rule that starts with that variable. Replace the variable with the right-hand side of the rule.

3. Repeat step 2 until no variables remain.

$$V := \{S, C, D, P, Loc, Col, \wedge, \vee\},$$

Σ is $\{c_1, c_2, c_3, red, green, blue\}$ and rules are as given in section 3.3. This grammar describes a basic programming language for expressions containing logical connectives and predicates. For example, the string $color(c_1) = red \wedge color(c_2) = green$ can

be generated from the grammar. The sequences of substitutions to obtain a string is called a *derivation*. The derivation of the above example is shown in B.1.

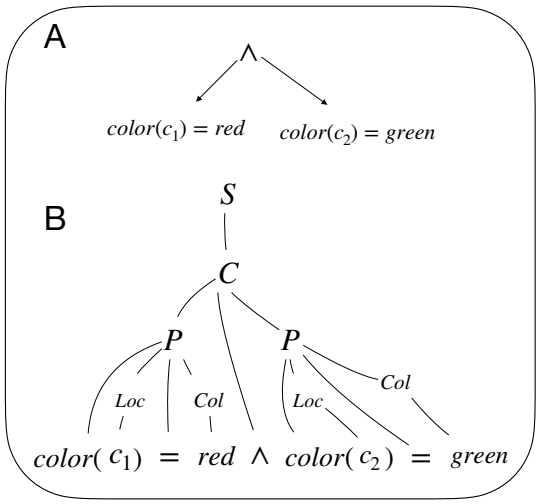

Figure B.1: A: Example of an expression that evaluates to true in the example from 3.1. B: Corresponding derivation from the language defined by the grammar in section 3.3.

Bibliography

[1] Thomas Gilovich. *How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life*. Free Press, 1991.

[2] Karl Jaspers. *General Psychopathology*. JHU Press, November 1997.

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

[4] Jerome C Wakefield. The concept of mental disorder: Diagnostic implications of the harmful dysfunction analysis. *World Psychiatry*, 6(3):149–156, October 2007.

[5] Kengo Miyazono. Delusions as Harmful Malfunctioning Beliefs. *Consciousness and Cognition*, 33:561–573, May 2015.

[6] Michael H. Connors and Peter W. Halligan. Delusions and theories of belief. *Consciousness and Cognition*, 81:102935, May 2020.

[7] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman. How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, 331(6022):1279–1285, March 2011.

[8] Charles Kemp, Joshua B. Tenenbaum, Sourabh Niyogi, and Thomas L. Griffiths. A probabilistic model of theory formation. *Cognition*, 114(2):165–196, February 2010.

[9] Rick A. Adams, Klaas Enno Stephan, Harriet R. Brown, Christopher D. Frith, and Karl J. Friston. The Computational Anatomy of Psychosis. *Frontiers in Psychiatry*, 4, May 2013.

[10] G. E. Berrios. Delusions as "Wrong Beliefs": A Conceptual History. *British Journal of Psychiatry*, 159(S14):6–13, November 1991.

[11] William James. *The Principles of Psychology.* New York : Holt, 1890.

[12] Brendan A. Maher. Delusional thinking and perceptual disorder. *Journal of Individual Psychology*, 30(1):98–113, 1974.

[13] Tony Stone and Andrew W. Young. Delusions and Brain Injury: The Philosophy and Psychology of Belief. *Mind & Language*, 12(3-4):327–364, sep 1997.

[14] W Hirstein and V S Ramachandran. Capgras syndrome: A novel probe for understanding the neural representation of the identity and familiarity of persons. *Proceedings of the Royal Society B: Biological Sciences*, 264(1380):437–444, March 1997.

[15] Robyn Langdon and Max Coltheart. The Cognitive Neuropsychology of Delusions. *Mind & Language*, 15(1):184–218, 2000.

[16] Brandon K. Ashinoff, Nicholas M. Singletary, Seth C. Baker, and Guillermo Horga. Rethinking delusions: A selective review of delusion research through a computational lens. *Schizophrenia Research*, page S0920996421000657, March 2021.

[17] P. Read Montague, Raymond J. Dolan, Karl J. Friston, and Peter Dayan. Computational psychiatry. *Trends in Cognitive Sciences*, 16(1):72–80, January 2012.

[18] Klaas Enno Stephan and Christoph Mathys. Computational approaches to psychiatry. *Current Opinion in Neurobiology*, 25:85–92, April 2014.

[19] Xiao-Jing Wang and John H. Krystal. Computational Psychiatry. *Neuron*, 84(3):638–654, November 2014.

[20] Christoph Mathys. How Could We Get Nosology from Computation? In A. David Redish, Joshua A. Gordon, and J. Lupp, editors, *Computational Psychiatry: New Perspectives on Mental Illness*, volume 20 of *Strüngmann Forum Reports*, pages 121–135. MIT Press, Cambridge, MA, 2016.

[21] Rick A. Adams, Quentin J. M. Huys, and Jonathan P. Roiser. Computational Psychiatry: Towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery & Psychiatry*, 87(1):53–63, January 2016.

[22] Quentin J. M. Huys, Tiago V. Maia, and Michael J. Frank. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, 19(3):404–413, March 2016.

[23] Joshua B. Tenenbaum and Thomas L. Griffiths. Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24(4):629–640, August 2001.

[24] Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79, jan 1999.

[25] Karl Friston. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):815–836, April 2005.

[26] Philipp Sterzer, Rick A. Adams, Paul Fletcher, Chris Frith, Stephen M. Lawrie, Lars Muckli, Predrag Petrovic, Peter Uhlhaas, Martin Voss, and Philip R. Corlett. The Predictive Coding Account of Psychosis. *Biological Psychiatry*, 84(9):634–643, November 2018.

[27] Pierre Maurice Marie Duhem. *La théorie physique: son objet, et sa structure*. Chevalier & Rivière, 1906.

[28] W. V. Quine. Two Dogmas of Empiricism. *The Philosophical Review*, 60(1):20–43, 1951.

[29] M. Strevens. The Bayesian Treatment of Auxiliary Hypotheses. *The British Journal for the Philosophy of Science*, 52(3):515–537, sep 2001.

[30] Edwin T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[31] Samuel J. Gershman. How to never be wrong. *Psychonomic Bulletin & Review*, 26(1):13–28, February 2019.

[32] D. R. Hemsley and P. A. Garety. The formation of maintenance of delusions: A Bayesian analysis. *The British Journal of Psychiatry*, 149(1):51–56, July 1986.

[33] Max Coltheart, Peter Menzies, and John Sutton. Abductive Inference and Delusional Belief. *Cognitive Neuropsychiatry*, 15(1-3):261–287, January 2010.

[34] William J. Speechley, Jennifer C. Whitman, and Todd S. Woodward. The Contribution of Hypersalience to the "Jumping to Conclusions" Bias Associated with Delusions in Schizophrenia. *Journal of Psychiatry & Neuroscience : JPN*, 35(1):7–17, January 2010.

[35] Robert Dudley, Peter Taylor, Sophie Wickham, and Paul Hutton. Psychosis, Delusions and the "Jumping to Conclusions" Reasoning Bias: A Systematic Review and Meta-Analysis. *Schizophrenia Bulletin*, 42(3):652–665, May 2016.

[36] Todd S. Woodward, Steffen Moritz, Carrie Cuttler, and Jennifer C. Whitman. The Contribution of a Cognitive Bias Against Disconfirmatory Evidence (BADE) to Delusions in Schizophrenia. *Journal of Clinical and Experimental Neuropsychology*, 28(4):605–617, May 2006.

[37] Daniel Freeman, Philippa A. Garety, David Fowler, Elizabeth Kuipers, Paul E. Bebbington, and Graham Dunn. Why Do People With Delusions Fail to Choose More Realistic Explanations for Their Experiences? An Empirical Investigation. *Journal of Consulting and Clinical Psychology*, 72(4):671–680, 2004.

[38] Philippa A. Garety, Daniel Freeman, Suzanne Jolley, Graham Dunn, Paul E. Bebbington, David G. Fowler, Elizabeth Kuipers, and Robert Dudley. Reasoning, Emotions, and Delusional Conviction in Psychosis. *Journal of Abnormal Psychology*, 114(3):373–384, aug 2005.

[39] Benjamin F. McLean, Julie K. Mattiske, and Ryan P. Balzan. Association of the Jumping to Conclusions and Evidence Integration Biases With Delusions in Psychosis: A Detailed Meta-Analysis. *Schizophrenia Bulletin*, 43(2):344–354, mar 2017.

[40] Annabel Broyd, Ryan P. Balzan, Todd S. Woodward, and Paul Allen. Dopamine, Cognitive Biases and Assessment of Certainty: A Neurocognitive Model of Delusions. *Clinical Psychology Review*, 54:96–106, jun 2017.

[41] Michael V. Bronstein, Gordon Pennycook, Jutta Joormann, Philip R. Corlett, and Tyrone D. Cannon. Dual-process theory, conflict processing, and delusional belief. *Clinical Psychology Review*, 72:101748, aug 2019.

[42] K. Friston. A Theory of Cortical Responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):815–836, apr 2005.

[43] Paul C. Fletcher and Chris D. Frith. Perceiving Is Believing: A Bayesian Approach to Explaining the Positive Symptoms of Schizophrenia. *Nature Reviews Neuroscience*, 10(1):48, January 2009.

[44] P.R. Corlett, J.R. Taylor, X.-J. Wang, P.C. Fletcher, and J.H. Krystal. Toward a Neurobiology of Delusions. *Progress in Neurobiology*, 92(3):345–369, nov 2010.

[45] Philip R Corlett, Garry D Honey, and Paul C Fletcher. Prediction Error, Ketamine and Psychosis: An Updated Model. *Journal of Psychopharmacology (Oxford, England)*, 30(11):1145–1155, November 2016.

[46] Katharina Schmack, Ana Gòmez-Carrillo de Castro, Marcus Rothkirch, Maria Sekutowicz, Hannes Rössler, John-Dylan Haynes, Andreas Heinz, Predrag Petrovic, and Philipp Sterzer. Delusions and the Role of Beliefs in Perceptual Inference. *Journal of Neuroscience*, 33(34):13701–13712, aug 2013.

[47] Alan Jern, Kai-min K. Chang, and Charles Kemp. Belief polarization is not always irrational. *Psychological Review*, 121(2):206–224, 2014.

[48] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, dec 2006.

[49] Finale Doshi-velez. The Infinite Partially Observable Markov Decision Process. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 477–485. Curran Associates, Inc., 2009.

[50] John R Anderson. The adaptive nature of human categorization. *Psychological review*, 98(3):409, 1991.

[51] Samuel J. Gershman and David M. Blei. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12, feb 2012.

[52] Anne Collins and Etienne Koechlin. Reasoning, Learning, and Creativity: Frontal Lobe Function and Human Decision-Making. *PLoS Biology*, 10(3):e1001293, mar 2012.

[53] Radford M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 2000.

[54] Ryan Mckay. Delusional Inference. *Mind & Language*, 27(3):330–355, June 2012.

[55] Shitij Kapur. Psychosis as a State of Aberrant Salience: A Framework Linking Biology, Phenomenology, and Pharmacology in Schizophrenia. *American Journal of Psychiatry*, 160(1):13–23, January 2003.

[56] P.R. Corlett, G.D. Honey, and P.C. Fletcher. From Prediction Error to Psychosis: Ketamine as a Pharmacological Model of Delusions. *Journal of Psychopharmacology*, 21(3):238–252, May 2007.

[57] Philip R. Corlett, John H. Krystal, Jane R. Taylor, and Paul C. Fletcher. Why Do Delusions Persist? *Frontiers in Human Neuroscience*, 3, 2009.

[58] Aaron C. Courville, Nathaniel D. Daw, and David S. Touretzky. Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, 10(7):294–300, July 2006.

[59] A. D. Redish and A. Johnson. A Computational Model of Craving and Obsession. *Annals of the New York Academy of Sciences*, 1104(1):324–339, apr 2007.

[60] Samuel J. Gershman, David M. Blei, and Yael Niv. Context, Learning, and Extinction. *Psychological Review*, 117(1):197–209, 2010.

[61] Peter Fonagy and Elizabeth Allison. The role of mentalizing and epistemic trust in the therapeutic relationship. *Psychotherapy*, 51(3):372–380, 2014.

[62] Florian Schlagenhauf, Quentin J. M. Huys, Lorenz Deserno, Michael A. Rapp, Anne Beck, Hans-Joachim Heinze, Ray Dolan, and Andreas Heinz. Striatal dysfunction during reversal learning in unmedicated schizophrenia patients. *NeuroImage*, 89:171–180, April 2014.

[63] James A. Waltz. The Neural Underpinnings of Cognitive Flexibility and Their Disruption in Psychotic Illness. *Neuroscience*, 345:203–217, mar 2017.

[64] Rick A. Adams, Gary Napier, Jonathan P. Roiser, Christoph Mathys, and James Gilleen. Attractor-like Dynamics in Belief Updating in Schizophrenia. *Journal of Neuroscience*, pages 3163–17, September 2018.

[65] Seth C. Baker, Anna B. Konova, Nathaniel D. Daw, and Guillermo Horga. A distinct inferential mechanism for delusions in schizophrenia. *Brain*, 142(6):1797–1812, June 2019.

[66] Maël Donoso, Anne G. E. Collins, and Etienne Koechlin. Foundations of Human Reasoning in the Prefrontal Cortex. *Science*, 344(6191):1481–1486, June 2014.

[67] Karl J. Friston, Marco Lin, Christopher D. Frith, Giovanni Pezzulo, J. Allan Hobson, and Sasha Ondobaka. Active Inference, Curiosity and Insight. *Neural Computation*, 29(10):2633–2683, October 2017.

[68] Ryan Smith, Philipp Schwartenbeck, Thomas Parr, and Karl J. Friston. An Active Inference Approach to Modeling Structure Learning: Concept Learning as an Example Case. *Frontiers in Computational Neuroscience*, 14, 2020.

[69] N. Chomsky. Three models for the description of language. *IRE Transactions on Information Theory*, 2(3):113–124, September 1956.

[70] Tomer D Ullman, Noah D Goodman, and Joshua B Tenenbaum. Theory Acquisition as Stochastic Search. page 6.

[71] Steven T. Piantadosi, Joshua B. Tenenbaum, and Noah D. Goodman. The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, 123(4):392–424, July 2016.

[72] Noah Goodman, Joshua Tenenbaum, Jacob Feldman, and Thomas Griffiths. A Rational Analysis of Rule-Based Concept Learning. *Cognitive Science: A Multidisciplinary Journal*, 32(1):108–154, January 2008.

[73] Ian Ballard, Eric M Miller, Steven T Piantadosi, Noah D Goodman, and Samuel M McClure. Beyond Reward Prediction Errors: Human Striatum Updates Rule Values During Learning. *Cerebral Cortex*, 28(11):3965–3975, November 2018.

[74] Sergey Levine. Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review. *arXiv:1805.00909 [cs, stat]*, May 2018.

[75] Beren Millidge, Alexander Tschantz, and Christopher L. Buckley. Whence the Expected Free Energy? *Neural Computation*, 33(2):447–482, February 2021.

[76] Michael Sipser. *Introduction to the Theory of Computation*. Boston : PWS Pub. Co., 1997.

[77] Herbert B. Enderton. *A Mathematical Introduction to Logic*. Harcourt/Academic Press, San Diego, 2nd ed edition, 2001.

[78] Steven T. Piantadosi, Joshua B. Tenenbaum, and Noah D. Goodman. Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, 123(2):199–217, May 2012.

[79] Harvey A. Whiteford, Louisa Degenhardt, Jürgen Rehm, Amanda J. Baxter, Alize J. Ferrari, Holly E. Erskine, Fiona J. Charlson, Rosana E. Norman, Abraham D. Flaxman, Nicole Johns, Roy Burstein, Christopher JL Murray, and Theo Vos. Global burden of disease attributable to mental and substance use disorders: Findings from the Global Burden of Disease Study 2010. *The Lancet*, 382(9904):1575–1586, November 2013.

[80] S. F. Huq, P. A. Garety, and D. R. Hemsley. Probabilistic Judgements in Deluded and Non-Deluded Subjects. *The Quarterly Journal of Experimental Psychology Section A*, 40(4):801–812, November 1988.

[81] P. A. Garety, D. R. Hemsley, and S. Wessely. Reasoning in Deluded Schizophrenic and Paranoid Patients: Biases in Performance on a Probabilistic Inference Task. *The Journal of Nervous and Mental Disease*, 179(4):194, April 1991.

[82] Jonathan D Cohen, Samuel M McClure, and Angela J Yu. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):933–942, May 2007.

[83] Daniella Laureiro-Martínez, Stefano Brusoni, and Maurizio Zollo. The neuroscientific foundations of the exploration-exploitation dilemma. *Journal of Neuroscience, Psychology, and Economics*, 3(2):95–115, November 2010.

[84] Eric Schulz and Samuel J. Gershman. The Algorithmic Architecture of Exploration in the Human Brain. *Current Opinion in Neurobiology*, 55:7–14, April 2019.

[85] Robert C. Wilson, Andra Geana, John M. White, Elliot A. Ludvig, and Jonathan D. Cohen. Humans Use Directed and Random Exploration to Solve the Explore–Exploit Dilemma. *Journal of experimental psychology. General*, 143(6):2074–2081, December 2014.

[86] Daniel J. Navarro, Ben R. Newell, and Christin Schulze. Learning and Choosing in an Uncertain World: An Investigation of the Explore–Exploit Dilemma in Static and Dynamic Environments. *Cognitive Psychology*, 85:43–77, March 2016.

[87] Charley M. Wu, Eric Schulz, Maarten Speekenbrink, Jonathan D. Nelson, and Björn Meder. Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, 2(12):915–924, December 2018.

[88] Maarten Speekenbrink and Emmanouil Konstantinidis. Uncertainty and Exploration in a Restless Bandit Problem. *Topics in Cognitive Science*, 7(2):351–367, 2015.

[89] Christopher G. Lucas, Thomas L. Griffiths, Joseph J. Williams, and Michael L. Kalish. A rational model of function learning. *Psychonomic Bulletin & Review*, 22(5):1193–1215, October 2015.

[90] Eric Schulz, Charley M. Wu, Quentin J. M. Huys, Andreas Krause, and Maarten Speekenbrink. Generalization and Search in Risky Environments. *Cognitive Science*, 0(0).

[91] Eric Schulz, Charley M. Wu, Azzurra Ruggeri, and Bjoern Meder. Searching for Rewards like a Child Means Less Generalization and More Directed Exploration. *bioRxiv*, page 327593, May 2018.

[92] James A. Waltz, Robert C. Wilson, Matthew A. Albrecht, Michael J. Frank, and James M. Gold. Differential Effects of Psychotic Illness on Directed and Random Exploration. *Computational Psychiatry*, 4:18–39, August 2020.

[93] Emmanuelle Peters, Stephen Joseph, Samantha Day, and Philippa Garety. Measuring Delusional Ideation: The 21-Item Peters et Al. Delusions Inventory (PDI). *Schizophrenia Bulletin*, 30(4):1005–1022, January 2004.

[94] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. John Wiley & Sons, September 2009.

[95] Steffen Moritz, Todd S. Woodward, Jennifer C. Whitman, and Carrie Cuttler. Confidence in Errors as a Possible Basis for Delusions in Schizophrenia:. *The Journal of Nervous and Mental Disease*, 193(1):9–16, January 2005.

[96] Ryan P. Balzan. Overconfidence in psychosis: The foundation of delusional conviction? *Cogent Psychology*, 3(1), January 2016.

[97] Christina Andreou, Steffen Moritz, Kristina Veith, Ruth Veckenstedt, and Dieter Naber. Dopaminergic Modulation of Probabilistic Reasoning and

Overconfidence in Errors: A Double-Blind Study. *Schizophrenia Bulletin*, 40(3):558–565, May 2014.

[98] Steffen Moritz, Nora Ramdani, Helena Klass, Christina Andreou, David Jungclaussen, Sarah Eifler, Susanne Englisch, Frederike Schirmbeck, and Mathias Zink. Overconfidence in incorrect perceptual judgments in patients with schizophrenia. *Schizophrenia Research: Cognition*, 1(4):165–170, December 2014.

[99] D. W. Joyce, B. B. Averbeck, C. D. Frith, and S. S. Shergill. Examining belief and confidence in schizophrenia. *Psychological Medicine*, 43(11):2327–2338, November 2013.

[100] Quentin J. M. Huys, Níall Lally, Paul Faulkner, Neir Eshel, Erich Seifritz, Samuel J. Gershman, Peter Dayan, and Jonathan P. Roiser. Interplay of Approximate Planning Strategies. *Proceedings of the National Academy of Sciences*, 112(10):3098–3103, March 2015.

[101] Neil R. Bramley, Peter Dayan, Thomas L. Griffiths, and David A. Lagnado. Formalizing Neurath's Ship: Approximate Algorithms for Online Causal Learning. *Psychological Review*, 124(3):301–338, 2017.

[102] Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38, 2013.

[103] Zachary Wojtowicz and Simon DeDeo. From Probability to Consilience: How Explanatory Values Implement Bayesian Reasoning. *arXiv:2006.02359 [cs, q-bio, stat]*, June 2020.

[104] T Lombrozo. Simplicity and probability in causal explanation. *Cognitive Psychology*, 55(3):232–257, November 2007.

[105] Tania Lombrozo. Explanatory Preferences Shape Learning and Inference. *Trends in Cognitive Sciences*, 20(10):748–759, October 2016.

[106] Joseph Jay Williams, Tania Lombrozo, and Bob Rehder. The hazards of explanation: Overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General*, 142(4):1006–1014, 2013.

[107] Alexandre Salvador, Luc H. Arnal, Fabien Vinckier, Philippe Domenech, Raphaël Gaillard, and Valentin Wyart. Premature commitment to uncertain beliefs during human NMDA receptor hypofunction. *bioRxiv*, page 2020.06.17.156539, June 2020.

[108] Clifford M. Cassidy, Peter D. Balsam, Jodi J. Weinstein, Rachel J. Rosengard, Mark Slifstein, Nathaniel D. Daw, Anissa Abi-Dargham, and Guillermo Horga. A Perceptual Inference Mechanism for Hallucinations Linked to Striatal Dopamine. *Current Biology*, 0(0), February 2018.

[109] Philip R. Corlett, Guillermo Horga, Paul C. Fletcher, Ben Alderson-Day, Katharina Schmack, and Albert R. Powers. Hallucinations and Strong Priors. *Trends in Cognitive Sciences*, 0(0), December 2018.

[110] Guillermo Horga and Anissa Abi-Dargham. An integrative framework for perceptual disturbances in psychosis. *Nature Reviews Neuroscience*, pages 1–16, November 2019.

[111] Tarryn Balsdon, Valentin Wyart, and Pascal Mamassian. Confidence controls perceptual evidence accumulation. *Nature Communications*, 11(1):1753, 04 2020.

[112] T. Balsdon, P. Mamassian, and V. Wyart. Separable neural signatures of confidence during perceptual decisions. *bioRxiv*, page 2021.04.08.439033, April 2021.

[113] Max Rollwage, Alisa Loosen, Tobias U. Hauser, Rani Moran, Raymond J. Dolan, and Stephen M. Fleming. Confidence drives a neural confirmation bias. *Nature Communications*, 11(1):2634, May 2020.

[114] Alan A Stocker and Eero P Simoncelli. A Bayesian Model of Conditioned Perception. page 8.

[115] Long Luu and Alan A Stocker. Post-decision biases reveal a self-consistency principle in perceptual inference. *eLife*, 7:e33334, May 2018.

[116] Bharath Chandra Talluri, Anne E. Urai, Konstantinos Tsetsos, Marius Usher, and Tobias H. Donner. Confirmation Bias through Selective Overweighting of Choice-Consistent Evidence. *Current Biology*, 28(19):3128–3135.e8, October 2018.

[117] Bharath Chandra Talluri, Anne E. Urai, Zohar Z. Bronfman, Noam Brezis, Konstantinos Tsetsos, Marius Usher, and Tobias H. Donner. Choices change the temporal weighting of decision evidence. *Journal of Neurophysiology*, 125(4):1468–1481, April 2021.

[118] Sonia Bansal, Gi-Yeul Bae, Benjamin M. Robinson, Britta Hahn, James Waltz, Molly Erickson, Pantelis Leptourgos, Phillip Corlett, Steven J. Luck, and James M. Gold. Association Between Failures in Perceptual Updating and the Severity of Psychosis in Schizophrenia. *JAMA Psychiatry*, December 2021.

[119] Samuel Joseph Gershman and Yael Niv. Perceptual Estimation Obeys Occam's Razor. *Frontiers in Psychology*, 4, 2013.

[120] Johannes Bill, Hrag Pailian, Samuel J. Gershman, and Jan Drugowitsch. Hierarchical structure is employed by humans during visual motion percep-

tion. *Proceedings of the National Academy of Sciences*, 117(39):24581–24589, September 2020.

[121] John Campbell. Rationality, Meaning, and the Analysis of Delusion. *Philosophy, Psychiatry, & Psychology*, 8(2):89–100, 2001.

[122] Yunbo Yang, Ulrike Lueken, Jan Richter, Alfons Hamm, André Wittmann, Carsten Konrad, Andreas Ströhle, Bettina Pfleiderer, Martin J. Herrmann, Thomas Lang, Martin Lotze, Jürgen Deckert, Volker Arolt, Hans-Ulrich Wittchen, Benjamin Straube, and Tilo Kircher. Effect of CBT on Biased Semantic Network in Panic Disorder: A Multicenter fMRI Study Using Semantic Priming. *American Journal of Psychiatry*, 177(3):254–264, March 2020.

[123] Kenneth S. Kendler. Toward a Philosophical Structure for Psychiatry. *American Journal of Psychiatry*, 162(3):433–440, March 2005.

[124] Kenneth S. Kendler and James Woodward. Top-down causation in psychiatric disorders: A clinical-philosophical inquiry. *Psychological Medicine*, 51(11):1783–1788, August 2021.

[125] Jerry A. Fodor. *The Modularity of Mind*. MIT Press, April 1983.

[126] Gregory L. Murphy. On Fodor's First Law of the Nonexistence of Cognitive Science. *Cognitive Science*, 43(5):e12735, 2019.

[127] Jerome Bruner. The narrative construction of reality. *Critical inquiry*, 18(1):1–21, 1991.