When people hold weird beliefs and can't give them up: Predictive processing and the case of strange, rigid beliefs

Author:

Alexander Kaltenboeck

Word Count:

7999 words

Name of Degree:

MSc. Mind, Language & Embodied Cognition

The University of Edinburgh

Year of Presentation:

2016

Acknowledgements

I would like to thank my supervisor, Andy Clark, for his constant support during the completion of this paper, for his valuable input, and for many interesting and intellectually stimulating discussions.

I would also like to thank Veronika Breunhoelder, for helpful suggestions when discussing my ideas, and for her support whenever I needed it.

Finally, I would like to thank my parents who have always supported me in pursuing my intellectual interests.

Abstract

This paper analyses the phenomenon of strange, rigid beliefs through the lens of predictive processing (PP). By "strange, rigid beliefs" I refer to abstract beliefs about the world for which, according to a rational and scientific worldview, there is no evidence available, yet which people struggle to abandon even when challenged with strong counterarguments or counterevidence.

Following recent PP accounts of delusion formation, I show that one explanation for such strangely persistent beliefs can be a breakdown of the predictive machinery itself. However, given how common strange, rigid beliefs are, I argue that there must be another kind of explanation too – one that does not presuppose a malfunction of the prediction engine.

This will lead me to develop an alternative account that I will call "hijacking beliefs". Using the example of supernatural beliefs, I will argue that certain abstract beliefs, when adopted under the right circumstances, are especially hard to dislodge for a predictive mind, as they are evidentially self-protective. Such beliefs may be consistent with a wide range of experiences and therefore hard to falsify, and might also bias future perception, action, and model-updating in ways that make them immune to rational revision.

Keywords: Predictive Processing; Delusion; Hijacking belief

Table of contents

Introduction

A prevailing default assumption in philosophy of mind (leaving non-representational positions such as "radical enactivism" [1] aside) is that, in order to successfully navigate through the world, an agent needs to maintain a proper of model of its external reality. Such a model can be conceived of as a set of beliefs that represent states of affairs in the world, and in order to keep track of changes in its environment, an agent has to update those beliefs as its experiences unfold.

What appears puzzling then are cases of what I will call "strange, rigid beliefs". With this I mean beliefs about the world which, from a naturalistic, rational, or scientific point of view, are clearly wrong or unlikely to be true (hence they are "strange"), yet which people struggle to abandon even when they face evidence against their correctness (hence they are "rigid"). Two paradigmatic cases of strange, rigid beliefs are supernatural or religious beliefs, and clinical delusions. In both cases, people apparently adopt beliefs about the world for which there seems to be no evidence available, and are reluctant to give up these beliefs when challenged with counterevidence or counterargument. How can we explain this phenomenon?

The aim of the following work is to analyse the phenomenon of strange, rigid beliefs through the lens of predictive processing (PP), a promising new theory of human mental functioning. I will discuss two different accounts of how a predictive mind might come to hold a strange, rigid belief. The first account will be based on a recent theory of delusion formation and presumes a breakdown of the prediction engine itself. The second account will assume normal mental functioning and hypothesizes that certain beliefs might have characteristics that allow them to "hijack" a healthy predictive mind such that they are unlikely to be abandoned once they have become adopted.

This work will consist of four chapters:

Chapter one will introduce the reader to the basic concepts of PP. It will describe how the human mind is thought to work according to this theory, how the various functional building blocks of the theory are thought to be instantiated by the brain, and what overall philosophical conception of the mind arises from such a framework.

Chapter two will discuss a recent theory of how clinical delusions can be explained within the PP framework. It will become clear that certain cases of strange, rigid beliefs might be the result of a glitch in the prediction engine.

Chapter three will put forward an alternative, speculative account of how a normally functioning predictive mind might come to hold strange, rigid beliefs. The key idea

will be that certain sets of beliefs, once they have become adopted, can influence future perception, action, and updating of an agent's model of the world such that it becomes unlikely that they are given up again.

The last chapter will address two further critical questions regarding hijacking beliefs: First, is holding a scientifically informed model of the world just another case of being hijacked? And second, how can we potentially protect our predictive minds against hijacking beliefs?

Chapter 1: How predictive minds work

"Predictive Processing" (PP) (also referred to as "hierarchical predictive coding") is a new theory of human mental functioning [2-7].[1] This theory depicts the brain essentially as a sophisticated prediction machine that constantly attempts to predict its own incoming sensory data based on a complex model of the world that it maintains [5-8]. If predictions and sensory input match, then the brain has fulfilled its task. However, if there is a significant mismatch, then a "prediction error" ensues, which signals to the brain that the current predictions are not able to fully account for the sensory input. The brain then is forced to resolve this prediction error either by changing its predictions (i.e. putting forward a different guess, or slowly changing its model of the world), or by changing its sensory input (i.e. actively seeking out the sensory input that is currently predicted). Predictions and sensory input are then compared again, and the whole process is repeated until they match and prediction error is minimized. Prediction error minimization, the PP story suggests, is the fundamental operating strategy of the human brain [5, 7].

In the long run, the brain can best minimize prediction error if it maintains an accurate model of how (exteroceptive, interoceptive, and proprioceptive) sensory input is caused [5, 6, 8, 9]. Within the PP framework, the brain's model of the world is construed to be "generative" because it allows the system to generate by itself the sensory input (i.e. the pattern of neuronal activation) that would result from specific distal worldly causes [5, 8]. The generative model is thought to consist of multiple hierarchical levels, where the different levels hold beliefs about the world at different spatiotemporal scales: Lower levels of the hierarchy are thought to represent more "immediate" (rapidly changing) features of the world with a relatively small spatiotemporal resolution (and with representations of the "actual" sensory input at the lowest level of the hierarchy), whereas higher levels are thought to represent more abstract (invariant) features of the world with a relatively large spatiotemporal resolution [5, 6]. At the highest levels, beliefs about very abstract and invariant regularities of the world ("hyperpriors") are thought to exist [5]. The beliefs held within the generative model are thought to be probabilistic representations that are not necessarily consciously accessible [8]. This is important to highlight, since within

[1] All descriptions of the basic principles of PP put forward in this chapter are based on (and can be found in greater detail in) the introductory book on the topic by Andy Clark [6].

this work, I will be mostly concerned with beliefs that can be consciously endorsed and verbally expressed. The details of where such beliefs reside within the generative model have not been fully fleshed out within the PP theory yet (Andy Clark, personal communication). However, a reasonable assumption seems to be that they reside at a higher level of the generative model.

The generative model is thought to be instantiated by a hierarchical neural network with two distinct types of units at each level: "Prediction units", which have been linked to deep pyramidal cells in the cortex, and "prediction error units", which have been linked to superficial pyramidal cells [8]. Prediction units are construed to represent beliefs about the world, while prediction error units are hypothesized to represent prediction error [8]. The prediction units at any level of the generative model attempt to predict the beliefs represented by the prediction units at the level below [8]. The prediction error units at the level below receive these predictions and compare them to the beliefs represented by the prediction units at their own level [8]. If the two don't match, then the prediction error unit passes forward a prediction error signal to the prediction units at the level above, causing them to refine their predictions such that they can better anticipate the beliefs at the level below [8]. This whole process takes place at all levels of the hierarchy and continues until the overall prediction error is minimized [8]. When this has happened, then the system is thought to have found the most probable hypothesis about the distal worldly cause(s) of the current sensory input in a Bayesian way [8]. This means that the final hypothesis of what the system believes to be out there in the world (posterior belief) is derived taking into account both, the sensory signal (sensory evidence) as well as what the system already knows about the world (prior beliefs) [8] (for a brief introduction to Bayesian inference refer to [6, 7]). Notice that what travels forward (or bottom-up) in the system is only the prediction error signal, i.e. just the parts of the sensory signal that are not predicted yet [5]. This constitutes a characteristic feature of the PP framework.

It is important to highlight here that within the PP framework "believing" and "perceiving" are crucially intertwined: What one perceives is constituted by what one currently believes about the world (at various spatiotemporal scales), and beliefs about the world are constantly refined depending on how well they can predict the current sensory input [5, 6, 10]. Based on its beliefs, the mind, according to PP, constantly creates a sort of "fantasy" [7] about the world (and internal states of the body [9]), and this fantasy is what we call perception (or emotional experience in the case of internal bodily signals [6, 9]). This "fantasizing", however, is by no means

arbitrary, as it is constantly put to test in how well it accommodates incoming sensory information.

An important problem for a predictive system then arises from the fact that sensory signals contain not only genuine information about the world, but depending on the specific circumstances, also a varying degree of noise [5-7]. In order to prohibit this noise from having undue influence on its beliefs (and its "fantasy") about the world, a predictive system must be able to distinguish it from genuinely informative signals [5-7]. The PP framework assumes that the brain solves this problem by maintaining not only beliefs about worldly states of affairs, but also about the reliability of incoming sensory information [5-7]. These reliability estimations are used to weight prediction error signals accordingly: The more reliable sensory input is deemed to be, the more seriously a resulting prediction error is taken when it comes to updating one's beliefs about the world [5-7].

To illustrate, consider the following example: Assume you have the hypothesis, that there's a spider in your bathroom, but for some reason, you can't check the correctness of this belief yourself. The only information you have access to is the verbal report of another person who has checked your bathroom. In one case, this person is a trusted friend who has never given you any reason to doubt his or her reports. In the second case, the other person is your evil neighbour who you know is an occasional liar and who loves giving you wrong information here and there. Now, imagine that in both cases you receive the identical information that there's no spider in the bathroom. Given your prior knowledge of the world you will certainly estimate your friend's report as much more reliable than that of your neighbour. Therefore, you might be willing to give up the belief that there's a spider in the bathroom based on your friend's report, but you might be reluctant to abandon this belief solely on the basis of your neighbour's report. The brain, PP suggests, essentially does the same thing: It estimates how reliable it deems a certain sensory input to be, and factors in this reliability when it forms and updates its beliefs about the world.

From a statistical perspective, what gets estimated at each level of the generative model is the precision of the prediction error signal (the inverse variance of the probability distribution that represents the prediction error) [5-7]. On the neurobiological level, these precision-estimations are thought to be conveyed by dopaminergic top-down projections [8] that regulate the synaptic gain (the "volume" [5]) of prediction error units. Top-down projections therefore are construed to not only transmit prior beliefs about states of affairs in the world, but also beliefs about the

what kind of information and which sources of input are considered reliable under specific circumstances.

Precision-estimation based weighting of prediction error is an important means for the brain to flexibly and context-dependently adjust the relative impact that top-down knowledge and bottom-up sensory information have during perceptual inference and formation of beliefs about the world [5-7]. If the current sensory input is estimated to be highly reliable, then prediction error resulting from this input is deemed to be of importance, and hence will have strong impact on perceptual inference and the updating of beliefs [5-7]. However, if incoming sensory signals are estimated to be unreliable, then they will not have any significant impact [5-7]. In such cases, the brain will rely more on its prior knowledge to reach a conclusion what it believes to be out there in its environment [5-7]. Notice that significant failures of the prediction engine can arise when precision-estimations get things wrong. When genuinely informative signal becomes incorrectly treated as noise, or when noise becomes treated as genuine signal, then this can lead the brain to perceive and believe things to be in the world that actually aren't there. We will encounter important examples of this in the chapters that follow.

Chapter 2: Strange, rigid beliefs as the result of a breakdown of the prediction engine

"*I had to make sense, any sense, out of all these uncanny coincidences. I did it by radically changing my conception of reality.*" Peter Chadwick [11] about his personal experience of a psychotic crisis (as quoted in [12])

Delusions are a common feature of psychiatric disorders, such as schizophrenia or dementia [13], and represent a paradigm case of strange, rigid beliefs. Although no universally accepted definition of delusions exists, most authors would probably agree that delusions are the result of a pathological process that leads an individual to adopt certain beliefs and hold them with undue conviction such that they become unsusceptible to counterevidence or counterargument. Often their content is clearly wrong, unlikely to be true, or bizarre and not understandable for people with the same sociocultural background [13]. To illustrate, two examples of delusions are the belief that one is persecuted and harmed by aliens (persecutory delusion), or that one is actually already deceased (Cotard delusion) [13]. Researchers have long tried to come up with a theoretical explanation of why people develop such beliefs and popular accounts depict delusions as resulting from abnormal perceptual experiences, reasoning biases, impairments in hypothesis evaluation, certain motivations, or a combination of these factors [13-16].

Recently, researchers have attempted to come up with accounts of delusion formation within the explanatory framework of PP [5, 6, 10, 12, 17]. The basic hypothesis of these accounts is that delusions represent abstract beliefs (presumably residing at a high level of the brain's generative model) that result from false generation and/or undue weighting of prediction error signals at lower levels of the hierarchical generative model [5, 6, 10, 12, 17]. The pathophysiological mechanism that is thought to cause this faulty error signalling is a malfunction of the dopaminergic system [10, 12]. In order to account for the false and/or improperly precise prediction errors, higher levels of the generative model will then be forced to adjust their beliefs accordingly, and these beliefs, when induced by what seems to be highly reliable prediction error, will require strong counterevidence to be given up again [5, 6, 10, 12].

Notice that in most cases the false and/or unduly weighted prediction error signals will probably arise at lower, sub-personal levels of the generative model and hence

their effects at first will not be consciously accessible [12]. However, when deemed precisely enough, they will work their way up through the hierarchy of the generative model, and ultimately also induce adjustments of beliefs that can be consciously endorsed and verbally expressed [6, 12].

This ongoing pathological generation and/or undue weighting of prediction error signals constitutes steady feedback to the brain that the current model of the world is still wrong and requires further adjustments [6, 12]. Ultimately, this will lead the brain to come up with ever more complex explanations for the error signal and adopt ever more exotic beliefs about the world [6, 12]. Frith and Friston [12] (p. 11–12) capture this dynamic by comparing it to a faulty dashboard warning light of a car:

"... assume that an error warning light is unduly sensitive to fluctuations in the engine's performance from normal levels. This would correspond to a pathologically highly [sic] *precision at the sensory level, leading to a dashboard warning light that is almost continuously illuminated. I am led to falsely believe that there is indeed something wrong with the engine. I take my car to the garage and they report that nothing is wrong. However, the light is still on and keeps on signalling an error. So, this leads me to falsely believe that the garage is incompetent. I report them to the "good garage guide" who investigate and conclude that the garage is not incompetent. Now I believe that the "good garage guide" is corrupt."*

Delusions, these accounts suggest, still represent the brain's best guess about states of affairs in the world, which, however, is now based on faulty input. The brain's beliefs about the world are no longer constrained by error signals that reliably indicate whether they can properly account for sensory input. Rather, the error signals now have become biased and distorted, leading the brain's model of the world to veer away from reality.

This whole process might also be captured by reports of first-person experience of delusion development (see for example Chadwick [11]). Affected people commonly report that, to them, arbitrary features of the environment, which one would normally tend to ignore (for example, the specific arrangement of the furniture in a room), appeared strikingly salient and in demand for explanation (for example, that someone is sending hidden messages through the arrangement of the furniture). These phenomenological reports, if the PP story of delusions is on track, might represent the subjective experience of inappropriate assignment of precision to some (arbitrary)

prediction errors, and the subsequent attempt of the system to account for these by changing its beliefs about the world [6, 12].

The rigid and fixed character of delusional ideas might then be brought about by a number of reasons. First, based on prior knowledge, the delusional belief might represent the best way to account for the ceaseless flood of wrong prediction errors, and alternative explanations might simply not be able to withstand the barrage of pathological error signals [5, 6]. What appears an unlikely (or impossible) hypothesis to the healthy mind (e.g. secret conspiracies, contact with aliens etc.), might become the most probable (or even only) explanation for a mind that faces ongoing distorted prediction error signals [5, 6]. Second, the faulty error signals that induce delusional beliefs are thought to be assigned unduly high precision, hence, the beliefs they induce will also be held with strong conviction [10, 12]. In turn, this means that (perhaps impossibly) strong counterevidence would be required for the delusional belief to be given up again. Third, once adopted, delusional beliefs can influence future perception, such that subjective experience aligns with them and lends further support to them [6, 10, 17]. Fourth, due to the ongoing faulty signalling of prediction error, as well as supporting perceptual experience, the delusional belief might get frequently reactivated and reconsolidated, which might further decrease the chances that it is given up again [17, 18].

If these PP accounts of delusion formation are indeed on track, then one way how a predictive mind can come to hold a strange, rigid belief seems to be a glitch in the predictive engine. However, strange, rigid beliefs are far too common to assume that neurobiological malfunction can explain all cases. Psychiatrists know a similar problem, as delusion-like beliefs and hallucination-like experiences (especially in a supernatural context) are much more common in the general population than the prevalence of psychotic disorders would suggest [19, 20]. Therefore, if the PP story is indeed on track, there must be another way how strange, rigid beliefs can be formed, and this explanation should presuppose a normally functioning predictive mind.

Chapter 3: Can a predictive mind get hijacked by its own beliefs?

This chapter will put forward a speculative account of how a healthy, normally functioning predictive mind can come to hold beliefs about the world for which (according to a scientific worldview) there is no evidence available, and which the agent struggles to abandon even when exposed to information that contradicts these beliefs. The key idea, that I will argue for, is that certain beliefs, once adopted by a predictive mind, might influence future perception, action, and updating of the generative model such that it becomes unlikely for an agent to get rid of these beliefs. I will refer to beliefs with such properties as "hijacking beliefs".

For the purpose of illustration, I will use the example of supernatural beliefs, which I think represent a paradigm case of hijacking beliefs. Beliefs in supernatural entities (i.e. entities that aren't acknowledged by a naturalistic or scientific worldview) are widespread amongst the general population and can be observed in every human culture [21]. For the naturalistic philosopher of mind these beliefs represent a puzzling phenomenon for at least three different reasons: First, how can a supernatural entity become part of an agent's model of the world, when apparently no such entity exists? Second, how can it furthermore be that seemingly healthy, neurotypical individuals holding such beliefs frequently and convincingly report perceptual encounters[2] with these entities (e.g. feeling the presence of a god, or hearing spirits whispering in the wind)? Third, why do some individuals show such a strong persistence in sticking with their supernatural beliefs, even when challenged with good counterargument or counterevidence?

Scholars have long argued about possible answers to these questions and a variety of different explanations have been put forward (for some historically important approaches see for example Freud [23], James [24], Feuerbach [25]; for more recent treatments, see for example Dawkins [26], Dennett [27]). Recently, cognitive scientists have joined this endeavour and tried to explain features of supernatural beliefs by appealing to various concepts from their field [21, 28-34]. To my knowledge, the PP framework has not been invoked in this discussion yet. It is to this task that I shall turn now.

---

[2] Here I intend to focus on the experiences of ordinary believers, not the kinds of rare and sensational cases of allegedly mystical revelations that can be explained relatively easily as resulting from exceptional neurological circumstances (e.g. temporal lobe seizures) [22].

How hijacking beliefs might enter the generative model

In the standard story of PP, the acquisition of a model of the world is firmly connected to sensory input. Beliefs (at various levels of abstraction) are constantly formed and evaluated in how well they can account for the incoming sensory data. This is a hierarchical process: When (less abstract) beliefs at lower levels can't sufficiently explain a sensory signal, then the resulting prediction error travels upwards through the hierarchy and leads to changes of (more abstract) beliefs at higher levels until the overall prediction error is minimized. This account of how a model of the world is formed and maintained is probably true for all mammals [6].

Humans, however, seem to have privileged access to another information channel that allows them to directly change high level beliefs in their generative model without having to make the corresponding sensory experiences first. This unique ability seems to be grounded in our linguistic skills. I can come to hold the high level belief that my new neighbour owns a black dog by seeing her going for a walk with the animal. I can, however, also learn about my new neighbour's black dog simply by being told about it. In the latter case, I directly adopt an abstract, high level belief without having had any perceptual experience of the dog myself. A similar phenomenon can be observed in the emotional domain: Human beings can learn that a certain abstract stimulus is associated with pain either by direct experience, or simply by being told about it. In both cases, when exposed to the stimulus in question, brain activity patterns as well as elicited emotional responses seem to be quite similar [35, 36].

Arguably, a significant part of our abstract, high level beliefs about the world is not derived from individual perceptual experience, but from what others tell us. How this ability can be explained within the PP framework is yet to be explored. My own guess would be that language processing works the same way as perceptual inference, in that the unfolding of a linguistic expression is predicted based on the current model of the world (including beliefs about the rules of language), and the model is then refined based on the prediction error that arises from a mismatch between the predictions and what is actually said. My generative model, for example, might only be able to accurately account for my friend saying "Your neighbour has a black dog" by adopting the high level belief that this is indeed the case.

Notice that the idea that humans can influence each other's high level beliefs about the world through verbal communication has recently also been extended by Roepstorff and Frith [37] to top-down motor control. They argue that in many cases our seemingly freely willed behaviour might in fact be influenced significantly by

abstract "scripts" that are shared between humans through language in a "top-top" fashion.

However, it is certainly not the case that we believe anything that we are told. As with sensory input, where the reliability (precision) of a signal is constantly estimated and factored in when it comes to perceptual inference and updating of the generative model, a predictive mind arguably also has to keep track on how trustworthy it estimates the linguistic expressions from other individuals to be. These estimations will likely be informed by our prior knowledge of the world (e.g. what we know about the person who is talking to us, what kind of interaction we are currently engaged in etc.). Only linguistic input deemed reliable enough will have significant influence on an agent's model of the world.

Linguistic transmission seems to play a particularly important role when it comes to the acquisition of supernatural beliefs. Special cases of individual, seemingly mystical experiences aside, most people acquire their supernatural beliefs through communication with other humans [38]. In fact, it is highly questionable whether a child without any contact to other humans would ever develop core religious ideas by itself [39], and one of the most important influences when it comes to religious beliefs are certainly one's own parents [26, 27]. A child's caregivers arguably are amongst its primary sources of knowledge about the world, and precision-estimations likely deem information coming from them as highly reliable (children might have a "programmed-in gullibility" [40] towards such information). Therefore, a child, when told by its caregivers about the existence of some supernatural entity, will likely adopt a corresponding high level belief into its model of the world.

However, a high level belief, even when acquired from what is estimated to be a reliable source, might have a short lifespan in the generative model if the predictions it makes can't properly account for the sensory input. If I tell you that there's a coffee stain on your, in fact, perfectly clean T-shirt, then you might trust me and adopt this high level belief into your model of the world. However, such a belief provides a precise way of how to falsify it: It predicts that if you look down on your shirt, you should see a stain. If you do that, and can't spot any signs of coffee on your shirt, the resulting prediction error will lead you to update your model and give up the now falsified belief.

However, the coffee-stain-on-your-T-shirt-belief makes very different predictions than those implied by supernatural beliefs. Usually, supernatural beliefs feature an agent-like entity equipped with some sort of super-human power. Such beliefs,

however, are consistent with a wide range of perceptual experiences. "Agency" is an abstract category that we assign an entity to when we can't easily anticipate its behaviour (we say that it has "a will of its own") [37]. Therefore, when something is construed to be an agent, we will predict that it is not easy to forecast how exactly it is going to behave under given circumstances. When the agent is furthermore thought to be equipped with super-human powers, then it becomes even less clear how to exactly predict encounters with it, as it could always use its abilities to conceal its existence. Put in a nutshell, this means that supernatural beliefs do not make predictions that can be falsified easily (if at all) by everyday experience. In a way, such beliefs resemble the kinds of claims that pseudosciences commonly make [41, 42]. To illustrate, take the example of an omnipotent, invisible god. The belief in the existence of such an entity is virtually consistent with any kind of sensory experience an agent makes. Just as with pseudoscientific claims (e.g. horoscopes) there's no specific perceptual situation that you could find yourself in, and no part of the world that you could actively sample, such that you could definitely make an experience that contradicts the predictions implied by this belief.[3]

Once adopted through linguistic transmission, these super-consistent supernatural beliefs therefore have a good chance to stay in a predictive mind's generative model, as they will not give rise to sampling-based prediction errors. However, once they have become part of an agent's model of the world, they can then induce a crucial dynamic.

As outlined previously, in order to make use of the noisy and context-variable sensory input that the predictive mind receives from the world, it has to maintain beliefs about what kind of information it considers reliable under given circumstances. These precision-estimations determine when a prediction error is thought to carry genuine information and hence should be used to refine the agent's beliefs about the world, and when the error is deemed to be the result of noise in the sensory data and hence can be safely ignored. What if a set of hijacking beliefs also contained expectations about the reliability of future input? Or, put differently, what if such a set of beliefs also changed a predictive mind's precision-estimations?

[3] The idea that supernatural beliefs are often irrefutable is of course not new. However, to my knowledge, the significance that this has for a predictive mind that is essentially in the business of testing its own hypotheses about the world has not been fully appreciated yet.

Seeing ghosts: When believing becomes perceiving

In order to figure out what's out there in the world during perceptual inference, the mind, as PP suggests, relies crucially on both, prior knowledge as well as actual sensory input. The balance between these two influences is adjusted by means of precision-estimation based weighting of prediction error, and depends on the particular context and the expected noise in the incoming signals [5-7].

Prior knowledge about the world can be crucial for successful perception. Thus, take the example of sine-wave speech discussed by Clark [6]. This skeletal outline of recorded speech, stripped of its normal features and acoustics, represents dynamical changes as pure tone whistles only (for a demonstration, see: http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/SWS/). Without prior knowledge, most people can't make out any intelligible verbal statement in the sine-wave utterance. However, when they are also shown the original record, such that they know what the sine-wave utterance is supposed to mean, then, when shown the sine wave statement again, their experience has changed. What first sounded like nothing more than a peculiar sequence of tones, now becomes an intelligible verbal utterance. This is a perfect demonstration of how, under specific circumstances, predictive minds need to rely on their prior knowledge of the world in order to make the best out of a sensory signal.

However, although prior knowledge might be necessary to extract meaningful patterns out of sensory data under some circumstances, it can also lead one's experience of the world astray when it gets assigned undue weight during perceptual inference. A nice example of such overshooting is demonstrated in a study by Merckelbach and van de Ven [43] (reported by Clark [6]): In their experiment, the authors played audiofiles of white noise to study participants. Beforehand, however, they had manipulated the participants' expectations by telling them that the file would contain a faint version of Bing Crosby's famous song "White Christmas" and that this could be hearable in the first, second, or last third of the recording. Although there was in fact no hidden song, almost one third of the participants indicated that they had heard it. This shows how prior beliefs, when assigned unduly high weight, can cause us to perceive things that aren't really there [6].

It now seems a reasonable hypothesis that a similar dynamic could ensue from the previously discussed irrefutable supernatural beliefs. When these abstract beliefs are assigned high weight (e.g. because they are acquired through an information stream that is deemed highly reliable, or because the hijacking belief set itself contains undue

precision-estimations), an agent's experience of the world might be biased such that it aligns with these beliefs.

Thus, consider the fact that many believers in supernatural entities report having some sort of perceptual encounters with the entities they believe in. As with the case of hearing a song in white noise, these experiences might result from high level beliefs that are assigned undue precision. During interoceptive inference [9], for example, naturally ambiguous internal body signals might then become felt as the presence of a god, some kind of sacred awe, or a state of being possessed. Similarly, in the exteroceptive domain believers might hear their ancestors' whispers in the wind, see demons rushing through a water stream, see auras surrounding other people, or have encounters with the Virgin Mary in all kinds of occasions (including seeing her face on a toast [44]).

An unduly weighted supernatural belief might therefore give rise to apparent perceptual encounters with the very entities it represents to exist in the world (see Pezzulo [45] for a similar argument that aims to explain why some predictive minds have alleged encounters with the "bogeyman" at night). The consequence of this dynamic, however, is that a set of beliefs that is already hard to falsify, now even gets (seemingly) supported by an agent's perceptual experiences.

When contradictory information becomes noise

However, the dynamic of hijacking beliefs might have yet another aspect worth considering. Precision-estimations are not only important during perceptual inference, but they might also more generally describe which sources of information are deemed trustworthy, and might also prescribe ways of how to actively sample the environment in order to harvest the most reliable information that can support (or falsify) a given belief [5, 6, 46]. Here too, hijacking beliefs might intervene such that they reduce the likelihood that they are given up again.

First, precision-estimations adopted with a set of hijacking beliefs might assign low reliability to sources that can potentially provide information that contradicts the hijackers. Consider for example the way many religious systems adjudge sources of information that potentially conflict with their beliefs (e.g. critical thinkers, non-believers, scientific education etc.) as sources of error, or evil attempts of misinformation. If a predictive mind's model of the world deems such sources as unreliable, then information coming from them will not have an effect on its beliefs. As a consequence supernatural believers might stick with their beliefs, even in the face

of what might seem to be clear counter-arguments or counter-evidence. What appears to be a conclusive argument from a critic's point of view, becomes treated as nothing else than negligible noise by the hijacked mind. This fits nicely with the experience of various scholars, who sometimes seem to despair of how stubbornly people adhere to their supernatural beliefs even in the face of seemingly good counter-arguments [26]. Furthermore, hijacking beliefs (or rather the precision-estimations that come with them) might also bias an agent's active sampling of the environment in their favour. Active agents, PP suggests, are driven to sample their environment such that they encounter the most reliable information that can support (or potentially falsify) a given belief about the world [6]. Visual saccades, for example, have been discussed to function as such a kind of "perceptual experiment", where prior beliefs and precision-estimations prescribe where to saccade next in order to harvest high quality information regarding the correctness of the current perceptual hypothesis [6, 46]. If high precision is assigned only to sources that support the hijacking belief sets, then a confirmation bias [47] can ensue, leading an agent to actively harvest only information that does not threaten its hijackers. This phenomenon can be observed in the case of supernatural beliefs as well. For instance, many religious systems feature behavioural norms that lead to exactly such biased sampling of the environment. Examples of this include staying away from non-believers or "heretics", interacting exclusively with fellow believers, or reading only specific kinds of literature. In a similar vein, some norms might even encourage believers to actively restructure their environment, as by erecting religious architecture, or imposing ritualised religious practices on whole communities. This biased sampling, of course, decreases the chances that a hijacked agent encounters information that conflicts with its beliefs even further.

Hijacking beliefs: A summary

Using the example of supernatural beliefs, I have argued that, under the right circumstances, a predictive mind can fall victim to its own beliefs. These beliefs, once acquired, can bias future perception, action, and updating of the generative model such that the agent has little chance of getting rid of them again. The exact dynamic can be summarised as follows:

1. A set of (abstract, high level beliefs) is incorporated under conditions estimated to provide highly reliable information (e.g. through verbal communication with caregivers). Hence, the beliefs are assigned high precision and strong

counterevidence is needed for an agent to give them up again (they are held with strong conviction).

2. These beliefs, however, make predictions that are consistent with a wide range of sensory input. Hence, it is unlikely that an agent comes across situations that cause enough prediction error to lead it to abandon the beliefs.
3. A set of hijacking beliefs might furthermore also encompass beliefs about the reliability of future sensory input (i.e. precision-estimations) that assign high reliability only to information that accords with the content of the hijacking beliefs.
4. During perceptual inference, certain beliefs might also be assigned high weight such that subjective experience accords with them and lends further support to them.
5. Finally, hijacking beliefs might influence an agent's active sampling of the environment such that it only harvests information that is in line with them. In a similar vein, hijacking beliefs might also lead an agent to actively restructure its environment, such that it becomes even less likely for the agent to encounter potentially contradicting information.

Notice, that none of these processes must happen on a conscious level. In the case of supernatural entities, the corresponding beliefs might reside on a high, consciously accessible and verbally expressible level of the generative model. However, supernatural beliefs are certainly not the only instances of hijacking beliefs and other cases might also befall mainly lower levels. One such example might be depicted by a recent PP account of "functional" (i.e. without a classical pathophysiological substrate) motor and sensory symptoms [8]. According to this theory – similar to the idea of hijacking beliefs – certain neurological impairments can result from the adoption of wrong sub-personal beliefs that are assigned undue precision, and that subsequently bias perception or action in a way such that the affected agent makes experiences that support these beliefs.

To conclude, we can highlight that the account put forward in this chapter provides a first sketch of how a healthy predictive mind might come to hold strange, rigid beliefs. Though in these cases the prediction engine functions perfectly normal, by coming across the wrong kind of hijacking information a predictive mind can get caught in an epistemic trap, where chances for escape become extremely low. If these ideas are

indeed on track, then the possibility to become hijacked by its own beliefs might represent just another instance of the "dark sides" of the predictive mind [49].[4]

[4] Andy Clark has pointed out to me that formally Brown's [48] "complete class theorem" shows that, given the right priors, any behaviour can become Bayes optimal. The fine-grained dynamics of hijacking beliefs, might therefore also be seen as an illustration of this fundamental, yet somehow depressing, logical fact.

Chapter 4: Two further worries regarding hijacking beliefs

In the previous chapters I have discussed two different theoretical accounts of how a predictive mind can come to hold strange, rigid beliefs. I have argued that though in some cases a breakdown of the prediction engine can underlie their formation, this might not be a necessary requirement, as a predictive mind can also adopt and maintain strange, rigid beliefs just by coming across the wrong kind of (Bayesian) hijacking information – as may be the case for widely held supernatural beliefs. In the following paragraphs I will briefly discuss two further worries that might be raised with regards to this idea.

The case of individual scientific beliefs

The key characteristic of hijacking beliefs is that, once acquired, they are hard to get rid of again, since they are consistent with a wide range of sensory input, and influence an agent's future experience such that information that contradicts them becomes unlikely to be encountered, or to have an effect on the agent's model of the world. For the purpose of illustration, I have used the example of supernatural belief systems in the previous chapter. However, these are certainly not the only instances of beliefs hijacking a predictive mind. Other potential examples might be conspiracy theories, political ideologies, stereotypes, or folie à deux [50, 51], where similar mechanisms seem to be at work.

A sceptic might now point out that, in the average agent, scientific education also installs a kind of hijacking belief system. She might say that, as in the case of supernatural beliefs, scientific education also transmits beliefs in the existence of entities that are not falsifiable by everyday sensory experience (such as cells, molecules, or subatomic particles), and that scientific education also changes precision-estimations such that high reliability is only assigned to sources of information that are likely to support these beliefs (e.g. articles in scientific journals, academic lectures etc.).

However, such a criticism misconstrues matters. It does seem right that during (verbally communicated) scientific education abstract beliefs about the world as well as about the reliability of certain sources of information become installed in a predictive mind. However, contrary to the (Bayesian) hijacking picture, the model of the world that is set up during scientific education remains plastic and is strongly open for change and revision in the light of new evidence. Unlike the hijacking case, a scientific world view does not contain beliefs that are quasi set in stone and that are

protected against falsification by the way they are construed, and the way they influence the reliability assigned to incoming information. Rather, beliefs featured in a scientific worldview aim to offer exact prescriptions of how they can be perceptually falsified. For example, I might not be able to see a neuron during my everyday experience, however, my belief in neurons makes a clear prediction under which circumstances I should encounter one: All I need to do is to look at some brain tissue through the right microscope. This distinguishes it clearly from the belief in, for example, an invisible fairy (where would I look for that?).

Furthermore, the reliability estimations that go along with scientific worldviews do not exclude sources of information that could potentially provide evidence that falsifies scientific beliefs. Quite the contrary, a scientific worldview assigns high precision to (and even encourages to actively harvest) information from sources that can potentially change very quickly large parts of what one believes about the world, such as for example the monthly publications in distinguished scientific journals. Information, in order to be spread by this sources, must fulfil a number of criteria (testability, evidential support, (blind) peer review etc.) that makes sure that it does not induce any hijacking dynamic. The sources of information that are indeed deemed unreliable within a scientific worldview (e.g. holy books, prophecies, oracles etc.) are exactly those that do not subject themselves to these protective criteria.

Contrary to the hijacking case, one can therefore say, what gets installed in a predictive mind through scientific education is a model of the world that remains deliberately susceptible to revision and that encourages an agent to actively seek out high quality information that could potentially contradict (and thereby refine) its beliefs about the world.

Can we protect predictive minds against hijacking beliefs?

Before I go on to conclude my discussion of strange, rigid beliefs through the lens of PP, a final – more practical – question deserves consideration here: Assuming my theory of hijacking beliefs is indeed on track, can we potentially protect our minds from falling victim to such a vicious dynamic?

It seems to me that the crucial aim of any potentially protective measures must be to prevent a predictive mind from adopting a set of hijacking beliefs in the first place. This opens up two (mutually not exclusive) possible approaches.

First, we could aim to reduce our exposure to hijacking beliefs as far as possible. Practically this would mean to avoid contact to potential sources of hijacking

information and – on a larger scale – to prohibit others from spreading it (similar to certain legislations that prevent advertisements from spreading bluntly wrong or misleading information). However, as one can easily see, such an approach is unlikely to be successful alone, not only because new beliefs with hijacking potential are probably created every day, but also because it touches sensitive ethical issues such as the right to free speech.

Another approach, that seems more promising, could aim for installing a certain mindset (a kind of "mental inoculation") in our predictive brains that makes it hard for hijacking beliefs to gain hold in the generative model. Such a mindset would first have to acknowledge the existence of hijacking beliefs and their effects, such that an agent can identify them when it comes across them, and deliberately not adopt them. More generally, the protective mindset could also assign low reliability to all those sources of information that can potentially induce (or have been identified to spread) beliefs with hijacking potential. And finally, the mental inoculation should also encourage an agent to expose itself to as as much different high quality abstract information and sensory experiences as possible, such that wrong beliefs about the world have a good chance of being falsified and abandoned.

From a practical perspective, philosophical and/or scientific education, together with frequent social and intellectual exchange, and an open-minded attitude towards worldly experiences seem well suited to realise exactly such a protective mindset.

Conclusion

In this work I have analysed the phenomenon of strange, rigid beliefs through the lens of predictive processing. Using the example of clinical delusions, I have shown that one way how predictive minds might come to hold strange, rigid beliefs is a breakdown of the prediction engine itself. However, given how common strange, rigid beliefs are, I have hypothesized that there must be another explanation which does not assume a malfunction of the prediction machine. This has led me to put forward the alternative account of hijacking beliefs. Hijacking beliefs, I have argued, are abstract beliefs about the world which are transmitted through verbal communication and which, due to their self-protective character, are hard to get rid of once they have become part of the generative model. They might be consistent with a wide range of experiences and hence are unlikely to be falsified, and they could also change perception and action such that contradictory information is unlikely to be encountered, or, if encountered, is not taken seriously. This picture of (Bayesian) hijacking thus shows how a healthy, neurotypical mind could develop a behaviour resembling that of clinical delusions simply by being exposed to the wrong kind of information. If these ideas are on track, then at least two questions open up for future research: First, how do distinct kinds of hijacking beliefs differ from each other? And second, how we can potentially intervene in order break the spell once a predictive mind has become hijacked?

References

1. Hutto, D.D., *Knowing What? Radical Versus Conservative Enactivism.* Phenomenology and the Cognitive Sciences, 2005. **4**(4): p. 389-405.
2. Friston, K., *The free-energy principle: a unified brain theory?* Nature Reviews Neuroscience, 2010. **11**(2): p. 127-138.
3. Friston, K., *The free-energy principle: a rough guide to the brain?* Trends in Cognitive Sciences, 2009. **13**(7): p. 293-301.
4. Friston, K., J. Kilner, and L. Harrison, *A free energy principle for the brain.* Journal of Physiology-Paris, 2006. **100**(1–3): p. 70-87.
5. Clark, A., *Whatever next? Predictive brains, situated agents, and the future of cognitive science.* Behavioral and Brain Sciences, 2013. **36**(03): p. 181-204.
6. Clark, A., *Surfing uncertainty: Prediction, action, and the embodied mind*. 2015: Oxford University Press.
7. Hohwy, J., *The predictive mind*. 2013: Oxford University Press.
8. Edwards, M.J., et al., *A Bayesian account of 'hysteria'*. Brain, 2012. **135**(11): p. 3495-3512.
9. Seth, A.K., *Interoceptive inference, emotion, and the embodied self.* Trends in Cognitive Sciences, 2013. **17**(11): p. 565-573.
10. Fletcher, P.C. and C.D. Frith, *Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia.* Nature Reviews Neuroscience, 2009. **10**(1): p. 48-58.
11. Chadwick, P.K., *The stepladder to the impossible: a first hand phenomenological account of a schizoaffective psychotic crisis.* Journal of Mental Health, 1993. **2**(3): p. 239-250.
12. Frith, C. and K.J. Friston, *False perceptions and false beliefs: understanding schizophrenia.* Neurosciences and the Human Person: New Perspectives on Human Activities, 2013: p. 1-15.
13. Bortolotti, L., *Delusion*, in *The Stanford Encyclopedia of Philosophy*, E.N. Zalta, Editor. 2016.
14. Maher, B.A., *Delusional thinking and perceptual disorder.* Journal of individual psychology, 1974. **30**(1): p. 98-113.
15. Langdon, R. and M. Coltheart, *The cognitive neuropsychology of delusions.* Mind & Language, 2000. **15**(1): p. 184-218.
16. McKay, R., R. Langdon, and M. Coltheart, *Models of misbelief: Integrating motivational and deficit theories of delusions.* Consciousness and Cognition, 2007. **16**(4): p. 932-941.
17. Corlett, P.R., et al., *Toward a neurobiology of delusions.* Progress in Neurobiology, 2010. **92**(3): p. 345-369.
18. Corlett, P.R., et al., *Why do delusions persist?* Frontiers in human neuroscience, 2009. **3**(12): p. 1-12.
19. Pechey, R. and P. Halligan, *The prevalence of delusion-like beliefs relative to sociocultural beliefs in the general population.* Psychopathology, 2011. **44**(2): p. 106-115.
20. Johns, L.C. and J. Van Os, *The continuity of psychotic experiences in the general population.* Clinical Psychology Review, 2001. **21**(8): p. 1125-1141.
21. Boyer, P., *Religion Explained: The Human Instincts That Fashion Gods, Spirits and Ancestors*. 2002: Vintage.
22. Shermer, M., *The Believing Brain: From Spiritual Faiths to Political Convictions – How We Construct Beliefs and Reinforce Them as Truths*. 2012: Robinson.
23. Freud, S., *Totem und Tabu*. 2014: Nikol.
24. James, W., *The varieties of religious experience*. 2008: Routledge.

25. Feuerbach, L., *Das Wesen des Christentums*. 2013: Reclam.
26. Dawkins, R., *The god delusion*. 2016: Black Swan.
27. Dennett, D.C., *Breaking the spell: Religion as a natural phenomenon*. 2006: Penguin Books.
28. Guthrie, S., *Faces in the clouds: A new theory of religion*. 1993: Oxford University Press.
29. Guthrie, S., *Why gods? A cognitive theory*, in *Religion in mind: Cognitive perspectives on religious belief, ritual, and experience*, J. Andresen, Editor. 2001, Cambridge University Press.
30. Barrett, J.L., *Cognitive science, religion, and theology: From human minds to divine minds*. 2011: Templeton Press.
31. Barrett, J.L., *Exploring the natural foundations of religion.* Trends in Cognitive Sciences, 2000. **4**(1): p. 29-34.
32. Barrett, J.L., *Cognitive science of religion: What is it and why is it?* Religion Compass, 2007. **1**(6): p. 768-786.
33. Barrett, J.L., *Cognitive Science of Religion*, in *Encyclopedia of Sciences and Religions*. 2013, Springer. p. 409-412.
34. Blackmore, S., *The Meme Machine*. 2000: Oxford University Press.
35. Phelps, E.A., et al., *Activation of the left amygdala to a cognitive representation of fear*. Nature Neuroscience, 2001. **4**(4): p. 437-441.
36. Olsson, A. and E.A. Phelps, *Learned Fear of "Unseen" Faces after Pavlovian, Observational, and Instructed Fear.* Psychological Science, 2004. **15**(12): p. 822-828.
37. Roepstorff, A. and C. Frith, *What's at the top in the top-down control of action? Script-sharing and 'top-top'control of action in cognitive experiments.* Psychological Research, 2004. **68**(2): p. 189-198.
38. Gervais, W.M., et al., *The cultural transmission of faith. Why innate intuitions are necessary, but insufficient, to explain religious belief.* Religion, 2011. **41**(3): p. 389-410.
39. Banerjee, K. and P. Bloom, *Would Tarzan believe in God? Conditions for the emergence of religious belief.* Trends in Cognitive Sciences, 2013. **17**(1): p. 7-8.
40. Dawkins, R. *Viruses of the Mind*. 1991 [cited 2016 16/08]; Available from: http://www.inf.fu-berlin.de/lehre/pmo/eng/Dawkins-MindViruses.pdf.
41. Hansson, S.O., *Science and Pseudo-Science*, in *The Stanford Encyclopedia of Philosophy*, E.N. Zalta, Editor. 2015.
42. Popper, K., *The logic of scientific discovery*. 2002: Routledge.
43. Merckelbach, H. and V. van de Ven, *Another White Christmas: fantasy proneness and reports of 'hallucinatory experiences' in undergraduate students.* Journal of Behavior Therapy and Experimental Psychiatry, 2001. **32**(3): p. 137-144.
44. BBC. *'Virgin Mary' toast fetches $28,000*. 2004 [cited 2016 14/08]; Available from: http://news.bbc.co.uk/1/hi/4034787.stm.
45. Pezzulo, G., *Why do you fear the bogeyman? An embodied predictive coding model of perceptual inference.* Cognitive, Affective, & Behavioral Neuroscience, 2014. **14**(3): p. 902-911.
46. Friston, K., et al., *Perceptions as hypotheses: saccades as experiments.* Frontiers in Psychology, 2012. **3**: p. 1-20.
47. Nickerson, R.S., *Confirmation bias: A ubiquitous phenomenon in many guises.* Review of General Psychology, 1998. **2**(2): p. 175-220.
48. Brown, L.D., *A Complete Class Theorem for Statistical Problems with Finite Sample Spaces.* The Annals of Statistics, 1981. **9**(6): p. 1289-1300.
49. Clark, A. *The dark side of the predictive mind*. The Brains Blog 2015 [cited 2016 13/08]; Available from: http://philosophyofbrains.com/2015/12/17/the-dark-side-of-the-predictive-mind.aspx.

50. Freeman, L.P., R.E. Cox, and A.J. Barnier, *Transmitting delusional beliefs in a hypnotic model of folie à deux.* Consciousness and Cognition, 2013. **22**(4): p. 1285-1297.
51. Wehmeier, P.M., N. Barth, and H. Remschmidt, *Induced Delusional Disorder. A Review of the Concept and an Unusual Case of folie à famille.* Psychopathology, 2003. **36**(1): p. 37-45.