Review of Philosophy and Psychology (2022) 13:47–63
https://doi.org/10.1007/s13164-021-00545-6

Is the biological adaptiveness of delusions doomed?

Eugenia Lancellotta[1]

Accepted: 22 March 2021 / Published online: 6 April 2021

Abstract

Delusions are usually considered as harmful and dysfunctional beliefs, one of the primary symptoms of a psychiatric illness and the mark of madness in popular culture. However, in recent times a much more positive role has been advocated for delusions. More specifically, it has been argued that delusions might be an (imperfect) answer to a problem rather than problems in themselves. By delivering psychological and epistemic benefits, delusions would allow people who face severe biological or psychological difficulties to survive in their environment - although this has obvious epistemic costs, as the delusion is fixed and irresponsive to compelling counterevidence. In other words, it has been argued that delusions are biologically adaptive. The adaptiveness of delusions has been compared by Ryan McKay and Daniel Dennett to a shear pin, a mechanism installed in the drive engine of some machines which is designed to shear whenever the machine is about to break down. By breaking, shear pins prevent the machine from collapsing and allow it to keep functioning, although in an impaired manner. Similarly, when delusions form, they would allow a cognitive or psychological system which is about to collapse to continue its functioning, although in an impaired manner. However, this optimistic picture of delusions risks being undermined by both theoretical and empirical considerations. Using Sarah Fineberg and Philip Corlett's recent predictive coding account as a paradigmatic model of the biological adaptiveness of delusions, I develop two objections to it: (1) principles of parsimony and simplicity suggest that maladaptive models of delusions have an upper hand over adaptive models; and (2) the available empirical evidence suggests that at least some delusions stand good chances of being psychologically adaptive, but it is unlikely that they also qualify as biologically adaptive.

✉ Eugenia Lancellotta
EXL737@student.bham.ac.uk

[1] University of Birmingham, 515 King Edwards Wharf, Sheepcote Street, Birmingham B16 8AB, UK

1 Introduction

Delusions are characterized as fixed and irrational beliefs which are not amenable to change in the light of evidence to the contrary and are held despite what everyone else believes (APA 2013, 87). They can be found in a number of psychiatric illnesses, ranging from schizophrenia to delusional disorder, depression, OCD, Body Dysmorphia, and Anorexia Nervosa. They are also generally held to be the outcome of a dysfunction of some sort (Davies et al. 2001; Fletcher and Frith 2008; Ellis and Young 1990) and to cause harm and psychological distress to the person who experiences them (McKay et al. 2005; Garety and Hemsley 1987).

However, already in the last century a more positive role was advocated for delusions. Drawing on Freud's approach to psychosis (1924/1986), so called psychodynamic or psychoanalytic accounts claimed that delusions might play a psychologically protective or defensive function, relieving painful and difficult emotions (Bell 2003). For example, Capgras and Carette (1924) saw the delusion of a young woman who believed that her father was an imposter as the result of a defence mechanism, aimed at hiding to the patient's consciousness her incestuous desires towards her father. Notwithstanding some exceptions, (e.g. Bentall and Kaney 1996; Enoch and Trethowan 1991), psychodynamic accounts of delusions and psychoanalytic theory more in general are usually dismissed by contemporary mainstream psychiatry as a relic of folk psychology with no scientific basis (Ellis 2003; Stone and Young 1997). This is mirrored by the DSM definition of delusions, which is based on one of the nowadays most influential paradigm of delusion formation and maintenance, the neuropsychological paradigm: this paradigm sees delusions as the by-product of neurological or physiological deficits or dysfunctions (e.g. Miyazono 2015; Davies et al. 2001; Langdon and Coltheart 2000).

The old psychodynamic idea that delusions might be an (imperfect) answer to a problem rather than problems in themselves has been more recently revisited by other theories of delusion formation and maintenance. Even more daringly than their psychodynamic ancestors, such theories have argued that the benefits of delusions can extend further than the psychological realm. By delivering both psychological and epistemic benefits, delusions would allow people who face severe biological or psychological difficulties to be better off and to survive in their environment - although with some epistemic costs, as the delusion is fixed and irresponsive to compelling counterevidence. In other words, such theories have argued that delusions are not only psychologically adaptive, in line with psychodynamic accounts, but also epistemically beneficial to some extent and biologically adaptive. In an influential paper, McKay and Dennett (2009) - who deny the biological adaptiveness of delusions but are open to their possible psychological adaptiveness - have compared the adaptiveness of delusions to a shear pin, a mechanism installed in the drive engine of some machines which is designed to shear whenever the machine is about to break down. By breaking, shear pins prevent the machine from collapsing and allow it to keep functioning, although in an impaired manner. Similarly, when delusions are formed and maintained, they would allow a cognitive or psychological system which is about to collapse to continue its functioning, although in an impaired manner. Despite McKay and Dennett ultimately arguing in their paper that delusions are not biologically adaptive, the shear pin metaphor has been particularly successful among supporters of the biological adaptiveness of delusions. Aaron Mishara and Philip Corlett et al. (2009) and, more recently, Sarah Fineberg and Corlett (2016, F&C from now on), have put forward the most developed model of the biological adaptiveness of delusions within a predictive coding framework, focusing on schizophrenic and neurological delusions such as Capgras.[1]

In this paper, I argue that claims about the biological adaptiveness of delusions are partly undermined both by theoretical and empirical considerations, which seem to suggest that, while some delusions might be psychologically adaptive, at least in the short term, it is unlikely that they are biologically adaptive. Taking F&C's model as my target, I develop two objections to the claim that delusions are biologically adaptive: (1) principles of simplicity and parsimony suggest that the maladaptive view of delusions should be preferred over the adaptive/shear pin model proposed by F&C; and (2) though still scant and difficult to find, the available empirical evidence speaks in favour of the psychological adaptiveness of some delusions but against their biological adaptiveness.

The paper is structured as follows. After briefly introducing the notion of adaptiveness (2), I will illustrate how such a notion, as well as the related notion of epistemic innocence, have been applied to delusions (2.1). Then I will examine F&C's as the most developed existing model of the biological adaptiveness of delusions. After introducing the basics of predictive coding (3) - the account of delusion formation which F&C's model belongs to – I illustrate F&C's model (4) in detail and raise two objections to it. The first objection (5.1) is more theoretical in nature, stating that between a maladaptive and an adaptive model of delusion formation, the former should be preferred, as it is simpler and makes fewer and less controversial assumptions than the latter. The second objection (5.2) is based on a review of some of the existing empirical evidence about the psychological and biological benefits of delusions. It maintains that, contra F&C's model, such evidence speaks against the biological adaptiveness of delusions. I also point to some additional empirical research that could be developed to test F&C's model. I thus conclude (6) that, before we accept the F&C's model, we need more empirical evidence in its support, even if obtaining it will be no easy task.

2 Adaptiveness

'Adaptiveness' is a key term in evolutionary biology and psychology. A trait or mechanism is considered to be biologically adaptive when it performs the function it was designed for by natural selection (Boorse 1975; Wakefield 1992). For example, a heart which pumps blood is an adaptive mechanism, as hearts were designed by natural selection to pump blood. The goal of adaptive traits and mechanisms is to support the reproductive success and survival of an organism in a given environment. An adaptive

[1] Although F&C's model has been developed with schizophrenic and Capgras delusions in mind, in principle there is no reason why the model should not be extended to delusions in other conditions, as predictive coding is an all-encompassing account of human cognition. In fact, beyond schizophrenia and Capgras, predictive coding models have also been developed for other disorders where delusions and delusion-like ideas are present, such as OCD (Levy 2018), depression (Badcock et al. 2017) and Anorexia Nervosa (Gadsby and Hohwy 2020).

trait is closely connected to the environment in which it developed. As a consequence, some traits can be adaptive in one environment but not in another or lose their adaptiveness as a consequence of an environmental change. By analogy with biological adaptiveness, philosophers and psychologists talking about beliefs and delusions often speak of psychological adaptiveness[2] when a trait, mechanism or mental state delivers psychological benefits, either by enhancing feelings of pleasure, or by conferring purpose and meaning to one’s life (Bortolotti 2015; McKay and Kinsbourne 2010; McKay et al. 2005). Although biological and psychological adaptiveness often go hand in hand, they can also come apart. For example, activities which are biologically adaptive (such as sexual intercourse) can also be psychologically pleasant. However, conditions such as extremely low levels of anxiety - as pleasant as they might be - pose serious threats to genetic fitness, as they make individuals less sensitive to danger and hence decrease their chances of survival in a given environment (Lee et al. 2006). In other words, they are psychologically but not biologically adaptive.

The notion of adaptiveness or maladaptiveness has been applied to behaviours that are associated with mental illness, in the following three ways (Murphy 2005): (a) as a failure of some component of the mind/brain to fulfil its evolutionary function; (b) as a result of the mismatch between the ancestral and the present environment; (c) as the unpleasant social consequences of atypical traits which were and are still adaptive in the present environment—it is debated, however, if the label of mental illness still applies to such cases. While (a) seems quite straightforward in accounting for the rise and development of mental illnesses, (b) and (c) are more controversial, and require complex causal histories of the phenomena they aim to account for. According to option (b), traits or mechanisms which were adaptive in a past environment are no longer adaptive in a new environment and they thus fail to deliver the benefits they were originally designed for. According to option (c), atypical traits like the lack of prosocial emotions which characterize antisocial personalities are usually labelled as mental illnesses. However, this would be a mistake from an evolutionary standpoint. These traits would in fact allow the person to survive and successfully reproduce – even if at the expenses of others - and hence they might in fact be biologically adaptive despite their being socially despicable.

Delusions are usually conceived as a maladaptive phenomenon, a breakdown of some component in the brain/mind machinery—in accordance with view (a). From a merely biological standpoint, delusions would be maladaptive: for example, people with schizophrenia, where delusions are largely present, tend to marry and reproduce less than controls (Nanko and Moridaira 1993). Also, paranoid ideation without psychosis in the general population is associated with depression, anxiety and suicide attempts, suggesting that delusional ideas have a negative impact on psychological wellbeing but also survival (Na et al. 2019). From a psychological standpoint, delusions are maladaptive as they hijack cognitive resources and, depending on the content, cause worry and unhappiness to the people experiencing them (e.g. Garety and Hemsley

[2] Here I use the notions of psychological adaptiveness and psychologically adaptive traits in the sense that the scholarly debate on delusions employs them, which should not be confused with the way in which evolutionary psychologists use the same terms. For evolutionary psychologists (see e.g. Schmitt and Pilcher 2004), psychologically adaptive traits (or psychological adaptations) are all those cognitive and behavioural traits that favour the fitness of an organism in a given environment. In this paper, psychologically adaptive traits are simply all those traits that provide some psychological benefit to the beholder.

1987) For example, people with persecutory delusions might be worried that people want to harm them and, as a consequence of that, live in fear and interrupt social contacts.

However, the proposal that delusions might be psychologically or even biologically adaptive has gained momentum. Expanding on the psychodynamic idea that delusions are the beginning of a solution rather than problems in themselves, delusions have been considered as part of ‘a shear pin mechanism’ which delivers biological or psychological benefits to organisms that find themselves in a situation of severe biological or psychological difficulty. In the next section, I am going to illustrate the metaphor and its application to delusions in more detail. I will also introduce the notion of epistemic innocence, that has been utilized by F&C’s model in support of the claim that delusions are biologically adaptive.

2.1 The Adaptiveness of Delusions: Psychological Adaptiveness, Epistemic Innocence and Biological Adaptiveness

According to the inventors of the metaphor, Ryan McKay and Daniel Dennett “*A shear pin is a metal pin installed in, say, the drive train of a marine engine. The shear pin locks the propeller to the propeller shaft and is intended to “shear” should the propeller hit a log or other hard object*” (McKay and Dennett 2009, 497). By breaking and disabling some of the parts of a system which is about to collapse, the function of the shear pin would be to prevent the system’s complete breakdown. In this way, the system keeps functioning in an impaired manner, but the breakdown is prevented.

McKay and Dennett consider that delusions might be akin to a shear pin mechanism, whose function would be to prevent the complete collapse of the cognitive system of an individual who finds himself in severe biological or psychological difficulties. However, they ultimately reject this idea, concluding that delusions are not the by-product of a shear pin mechanism. They acknowledge that some delusions could be psychologically adaptive, but also that this is not sufficient to make them biologically adaptive.

Drawing on McKay and Dennett, Lisa Bortolotti argues that in virtue of their psychological adaptiveness, some delusions are also epistemically innocent. A belief is epistemically innocent if 1. It delivers a significant epistemic benefit, such as the acquisition and retention of true beliefs; 2. This benefit cannot be otherwise achieved, i.e. at a minor epistemic cost (Bortolotti 2020; Bortolotti 2015). Motivated delusions would prevent severe depression by protecting from overwhelmingly negative emotions caused by adverse events (Bortolotti 2015); depressive delusions would restore a coherent sense of self by resolving the clash between pre-existing negative schemata about the self and contradictory evidence (Antrobus and Bortolotti 2016); elaborated and systematized delusions in schizophrenia would resolve the uncertainty and relieve the anxiety brought about by anomalous experiences and predictive errors (Bortolotti 2016). In all these cases, delusions would relieve uncertainty, low mood, and enhance one’s self-esteem: in other words, they would be psychologically adaptive. However, these psychological benefits would also translate into epistemic ones, as having anxiety relieved and mood enhanced would allow someone to be more engaged with the external environment, which in turn favours the acquisition and retention of true beliefs (condition 1 of epistemic innocence). Moreover, this epistemic benefit would not be otherwise achievable, i.e. by entertaining a non-delusional belief, because believers would not have access to alternative beliefs due to various dysfunctions, impairments or biases (condition 2 of epistemic innocence). It should be highlighted that the contact with the environment and the psychological relief that delusions provide is far from optimal, as delusions are irrational, irresponsive to rational arguments to the contrary, and often distressing depending on their content. However, this state of imperfect contact with reality and of precarious psychological equilibrium which delusions provide would still be better than one in which the difficulties that people undergo are not responded to by the delusions. For example, it is undeniable that schizophrenic delusions cause a great deal of psychological distress and, by limiting the social life of an individual, also negatively affect his contact with the environment. However, the epistemic and psychological consequences of not having the delusion for someone who is undergoing anomalous experiences and predictive errors would be even worse than entertaining the delusion that, for example, one is persecuted by the CIA.

Revisiting the notion of epistemic innocence as well as a previous proposal by Mishara and Corlett (2009) - who, in response to McKay and Dennett, argued that delusions could be the by-product of a shear pin mechanism - F&C argue that delusions are biologically adaptive: this is because being more in contact with the environment, as posited by epistemic innocence, is also key to survival and reproduction.

In what follows, I will explain how F&C conceive of the biological adaptiveness of delusions. Before doing that, however, it is necessary to illustrate the basics of predictive coding, as their model firmly relies on the principles of this popular theory of human cognition.

3 Predictive Coding

Fineberg and Corlett's model belongs to predictive coding theories of delusion formation and maintenance. Predictive coding is an influential model of human cognition. Its basic assumption is that the brain is a predictive machine ruled by a simple principle: to minimize uncertainty by building an internal model of the world where the gap between the expected and actual sensory input is reduced to minimum (e.g. Hohwy 2014; Fletcher and Frith 2008).

The mismatch between actual and expected sensory inputs is signalled by prediction errors (PEs). According to predictive coding, the final goal of the brain would be to get rid of PEs in the long run in order to achieve a consistent model of the world where expected sensory inputs reliably predict actual inputs (Williams 2018). PEs can be eliminated in various ways: actual inputs which mismatch expected inputs can be used to update the person's prior beliefs, engendering new beliefs and hence promoting learning; inputs can be discarded and prior beliefs retained; finally, PEs can also be eliminated or reduced by actions whose aim is to avoid the generation of PEs in the first place. Take Alice, who believes herself to be extremely overweight, but who in reality is very thin. Whenever she looks at her image in the mirror, Alice will get a disconfirmation of her belief that she is fat. PEs which signal the mismatch between her beliefs and her visual inputs are thus generated. To eliminate the PEs, Alice has three options: a. update her previous beliefs in light of her visual inputs and adopt the new belief that she is in fact thin; b. discard the visual inputs generated by the mirror image and continue to believe that she is fat; c. continue to believe that she is fat by stopping looking at the mirror altogether or by attending to specific body parts which she deems to be more fat or unattractive (as it happens in some cases of Anorexia Nervosa): in this way, as the source of the PEs is entirely avoided, PEs are also completely eliminated.

The criteria according to which Alice might choose one way or the other to get rid of PEs depend on the estimation of the precision of the PEs about body size. Estimation of precision indicates the degree of confidence that one has towards the reliability of PEs. If PEs are estimated to be highly precise (or reliable), that means that the set of beliefs or predictions that one holds are deemed not reliable enough to explain a given state of affairs. As a consequence, an update of beliefs will take place. On the contrary, if PEs are estimated to be highly imprecise, previous beliefs will be retained, as the generation of the PEs will be ascribed to some noise in the environment rather than to the unreliability of the set of beliefs. In general, precision is fundamental in the revision or maintenance of one's model of the world, to the point that, according to predictive coding, psychopathologies are entirely explicable as errors in precision weighing or estimation (Hohwy 2017). Here it is important to highlight the following point. The final goal of the predictive brain is to get rid of PEs in the long term, achieving consistency between one's set of beliefs and perceptual inputs. However, this does not necessarily entail that the resulting model of the world will mirror how the world really is, as there is not a preferential way to reduce PEs. The way one chooses to minimize PEs in the long run might or might not lead to the formation of veridical beliefs about the world, as this is the result of complex processes concerning precision estimation.

The predictive coding framework has also been characterized as a Bayesian inferential hierarchy.

Bayes' theorem stipulates what the best way of updating beliefs under conditions of uncertainty is:

$$\mathrm{p(h/e) = p(e/h)p(h)/p(e)}$$

More specifically, the probability of a hypothesis given the available evidence p(h/e) is proportional to its likelihood p(e/h) – how well the hypothesis predicts the sensory data - divided by its prior probability p(h) – the probability of the hypothesis prior to the sensory data. This would be the ideal way of updating beliefs under conditions of uncertainty. However, according to supporters of the Bayesian brain, the theorem would also bear a descriptive value, as many inferential processes in the brain would actually approximate Bayes' theorem (Knill and Pouget 2004).

Another important feature of predictive coding is that there is no clear-cut distinction between perceptions and beliefs (Corlett and Fletcher 2015; Fletcher and Frith 2008; Teufel and Fletcher 2016). Rather than being two discrete entities, perception and belief would represent two different levels of the same inferential hierarchy, organized according to spatiotemporal levels of increasing complexity. While perceptions would capture causal regularities at a more limited space and smaller duration of time, lower down in the hierarchy, beliefs would do it at the higher levels of the hierarchy, at a broader space and time length. However, although in everyday language we might choose to refer to the lower levels as perceptions and to the higher levels as beliefs, as Corlett and Fletcher put it, "*at another level of analysis – the one we think is more useful – no distinction is called for*" (Corlett and Fletcher 2015, 96–97). The disavowal of the distinction between perceptions and beliefs has a huge impact on the classification of psychotic symptoms: for example, hallucinations and delusions should not be considered as distinct entities but rather they can be explained "*in terms of a disturbed hierarchical Bayesian framework, without recourse to separate consideration of experience and belief*" (Fletcher and Frith 2008, 48). Finally, while in other theories of delusion formation perceptions can influence beliefs but not vice-versa, in predictive coding the influence between perceptions and beliefs is mutual and continuous, so that we do not only believe what we see but we also see what we believe (McKay 2012). This phenomenon is known as cognitive penetration.

It is interesting to see how predictive coding relates to the other most popular theories of delusion formation and maintenance, namely one- and two-factor accounts. One-factor accounts postulate that delusions are caused by a single neuropsychological impairment, which often takes the form of an anomalous perception: delusions would be a rational explanation of such perception (Maher 1974), as for one-factor theorists the cognitive capacities of people with delusions would be intact. On the other hand, two-factor accounts claim that people with delusions undergo a double dysfunction, the former being a neuropsychological impairment under the form of an aberrant perception and the latter being a reasoning bias or deficit (Coltheart et al. 2010; Davies et al. 2001; Stone and Young 1997). In this case, delusions would be the result of this double impairment. Some assimilate predictive coding to a one-factor account ("*We posit a single factor, prediction error dysfunction for delusion formation and maintenance*"; Corlett et al. 2010, 361); others, however, argue that predictive coding is not incompatible with a two-factor account (Miyazono and McKay 2019; McKay 2012). In any case, it is clear that predictive coding cannot be easily reduced either to a one- or to a two-factor account. Contrary to both one and two factor accounts, in predictive coding perceptions and beliefs are not two water-shed entities but two levels of the same inferential hierarchy which mutually influence each other. In line with one-factor accounts, predictive coding posits that delusions are reached via a single impairment. The impairment in question would be represented by aberrant PEs, which cause the person to depart from ideal Bayesian norms of rationality. In patients with delusions, PEs would erroneously be estimated to be more or less precise than they should, or elicited when they should not, by normal and unsurprising events. When PEs are erroneously estimated to be highly precise, "*possibly as a consequence of dopamine dysregulation, events that are insignificant and merely coincident seem to demand attention, feel important and relate to each other in meaningful ways. Delusions ultimately arise as a means of explaining these odd experiences*" (Corlett et al. 2009, 1). In the light of such PEs, a revision of beliefs take place, and a delusional belief is adopted as a new belief which explains the anomalous PEs. At a later stage, aberrant PEs are explained in the light of the delusional belief, giving way to a process of reinforcement which strengthens the delusion even more (Hohwy 2013). However, PEs can also be mistakenly estimated to be less precise than they should: in this case, prior beliefs are wrongfully overvalued and perceptual evidence is prematurely discarded.

In what follows, I will illustrate how the adaptive model of F&C marries predictive coding and what contribution this interaction brings to the issue of the biological adaptiveness of delusions.

4 Fineberg and Corlett's Model

F&C's model sees delusions both as biologically adaptive and entirely explicable by a predictive coding framework.

On the model, delusions are the by-product of the shear pin, which is designed to break in times of what can be called a 'doxastic emergency'. More precisely, such an emergency would be triggered by the person undergoing aberrant PEs (often under the form of anomalous experiences). The biological adaptiveness of delusions would consist in the fact that, when explaining those anomalous experiences, delusions contribute to the elimination of PEs and hence to a person's survival, by allowing the learning system to resume its functioning; as a result, the person continues to stay engaged with reality. For example, in the Capgras syndrome, aberrant PEs are generated as the result of expecting feelings of familiarity when perceiving a familiar face but instead not having those feelings. As long as the experience remains unexplained, a big part of the cognitive resources of the person is absorbed into the process of making sense of it, and cannot be employed to actively engage with other aspects of the outside world. In other words, the learning system (or a big part of it) is momentarily blocked by the presence of the unexplained experience. When the delusion emerges as an explanation of the experience - leading to the adoption of the belief that the dear one has been replaced by an imposter – those cognitive resources become available again and can be employed to interact with and exploit the external world, increasing the chances of survival and reproduction of the individual.

However, arriving at an explanation in this way also has some costs. Although delusions free cognitive resources, they remove the delusional belief from the control of the most flexible (but cognitively expensive) part of a person's learning system, the goal-directed learning system. In this way, the delusional belief falls under the control of the habitual learning system. While the former "*involves learning flexible relationships between actions and outcomes*" (Mishara and Corlett 2009, 530), the latter involves a more fixed relationship between stimuli and responses, such that to a specific stimulus always corresponds a specific response. For F&C's model, when delusions are adopted, the goal- directed part of the learning system is disabled in order to avoid the complete breakdown of the system, while the more habitual part monopolizes the control of the person's delusional beliefs. This shift in control between the different parts of the learning system explains the inflexibility which typically marks delusional states. This is the unavoidable cost of the shear pin break: preserving the overall functionality of the learning system in spite of aberrant perceptions or PEs has the side effect that the system does not function in an optimal manner. The delusion would restore the overall functionality of the learning system but at the cost of the delusion being irresponsive to evidence and to reasonable argument to the contrary.

It is important to notice that, according to F&C's model, the biological adaptiveness of delusions is mediated by their epistemic benefits. In other words, delusions increase the chances of survival and reproduction of an individual by restoring the functionality of the learning system, although this implies some epistemic costs. Restoring learning is good for the individual, because it allows him to stay engaged with the environment and exploit it for purposes of survival and reproduction.

> *Delusions [...] enable patients to stay engaged with the environment and to exploit its regularities, though the patient may be inflexible and unresponsive to corrective feedback* (Fineberg and Corlett 2016, 76)
> *Delusions form when the shear pin breaks, permitting continued engagement with an overwhelming world, and ongoing function in the face of paralysing difficulty* (Fineberg and Corlett 2016, 73)

The restored functionality of the learning system also translates into psychological benefits, easing the sense of confusion and anxiety which characterizes the presence of unexplained and anomalous experiences or PEs. In turn, at a physiological level, relief from anxiety manifests in a dramatic fall of the stress-related hormone cortisol.

5 Objections

Here I am going to consider two objections to the F&C's model, emerging from theoretical and empirical considerations.

5.1 The Maladaptive View Is Simpler

As I have pointed out in 3, according to predictive coding, delusions are generated by a single dysfunction, affecting both lower and upper levels of the Bayesian hierarchy (the equivalent of the folk-psychological notions of perceptions and beliefs); this dysfunction would cause the person to depart from Bayesian norms of rationality. How is it then possible for delusions to be dysfunctional and adaptive at the same time, as F&C argue? In other words, how can delusions be a response to a dysfunction if predictive coding generally holds that they are the result of a dysfunction?

In F&C's model, delusions are not dysfunctional: they are the by-product of a mechanism – the shear pin – which is functioning exactly as designed when it breaks and gives rise to delusions. Although the shear pin is activated by a dysfunction (under the form of anomalous PEs), its operation and the rise of delusions are not dysfunctional. This assumption seems to be somehow incompatible with the aetiology of delusions that predictive coding provides. In predictive coding, delusions are in fact dysfunctional, as they are the by-product of an abnormality – i.e. anomalous PEs - which disrupts inference at every level of the Bayesian hierarchy.

> *The unusual perceptual experiences and beliefs in psychosis can be explained by one core atypicality, namely a shift in the balance of Bayesian inference within a hierarchically-organised information processing system* (Teufel and Fletcher 2016, 5)

> *These theorists [of predictive coding] disavow any strict conceptual separation between experience and belief. There is just one basic abnormality—aberrant prediction error signalling—that disrupts inference across the board.* (McKay 2012, 349)

While in predictive coding PEs seem to directly generate delusions, by disrupting inference at every level of the Bayesian hierarchy, in F&C's model delusions are generated by the shear pin, which is triggered in response to the anomalous PEs. This seems to be an important point of conflict between predictive coding accounts of delusion formation and F&C's model. If, as predictive coding assumes, delusions are dysfunctional, this claim is incompatible with the claim that delusions are the by-product of an adaptive shear pin break. While in predictive coding delusions are something which goes against biological adaptiveness, in F&C's model delusions are part of the shear pin mechanism, which is designed by natural selection to support biological adaptiveness.

F&C's model clearly deviates from the standard/maladaptive version of predictive coding accounts of delusion formation insofar as delusions are not generated by anomalous PEs but by the shear pin, which stops the anomalous PEs from propagating to the upper levels of the Bayesian hierarchy.

> *Delusion formation. In response to prodromal* [i.e. when people present the anomalous PEs but not yet the delusion] *confusion and stress, the "doxastic shear-pin" breaks. Delusions appear in an aha- moment, when explanatory insight occurs and flexible processing is disabled.* (Fineberg and Corlett 2016, 76)

Nonetheless, there seems to be an important reason why standard/maladaptive predictive coding models of delusions should have the upper hand over their shear pin/ adaptive predictive coding counterparts in accounting for the rise and maintenance of delusions. Let's take Fred, who believes himself to be persecuted by the CIA.

While a maladaptive model of delusions holds that the delusion of persecution is adopted because when undergoing certain anomalous PEs, Fred cannot help but adopt the delusional belief, the shear pin view argues that the delusion of persecution is adopted in response to anomalous PEs which have the potential to seriously impair Fred's learning and cognitive system: it is when the shear pin kicks in to tackle the PEs that delusions ensue. The maladaptive view holds that 1. people with delusions undergo anomalous PEs; 2. that the PEs are sufficient to cause delusions to arise. The conclusion is that delusions are the maladaptive result of anomalous PEs.

Compared to the maladaptive view, the shear pin view holds two additional assumptions, i.e. 1. that, if not tackled, the PEs that people with delusions undergo have the potential to break down the overall functioning of their cognitive system and 2. that delusions are the by-product of a shear pin mechanism, which is supposed to tackle the anomalous PEs. It follows that, as having a functioning cognitive system is vital to surviving and reproducing, delusions are the adaptive by-product of a shear pin mechanism. In principle, between two or more models, the one that must be preferred for explaining a given phenomenon is the simplest, i.e. the one which makes the fewest assumptions. It is easy to see how, being the simplest model between the two, the maladaptive model of predictive coding should have the upper hand over the shear pin model endorsed by F&C, unless there are some aspects of delusions which the former cannot explain and hence motivate a recourse to the latter. It is thus fundamental for supporters of the shear pin version of predictive coding to specify those aspects of delusions that a maladaptive model would neglect and that a shear pin model would instead capture. F&C's model provides an answer to this issue, which is contained in the additional assumptions made by the model: the aspects that maladaptive/standard predictive coding models of delusions overlook relate to the fact that delusions are biologically adaptive, i.e. that they favour survival and reproductive success thanks to their epistemic benefits, by keeping a person who is undergoing anomalous PEs in some form of contact with his environment rather than in no contact at all.

However, such a claim requires strong empirical evidence in its support; otherwise, all things being equal, the maladaptive version of predictive coding accounts should be preferred over its adaptive/shear pin counterpart in virtue of its superior simplicity. In what follows, however, I will illustrate that the available empirical evidence, though still scant, does not raise good prospects for the biological adaptiveness of delusions.

5.2 The Maladaptive View Is more Compatible with Available Evidence

Having established that the F&C's shear pin model should pass the test of empirical evidence in order to be reasonably preferred over maladaptive models of delusions, what kind of studies should one conduct to verify whether the model is accurate?

I envisage that the issue can only be settled through studies which compare people who undergo anomalous PEs and have delusions with people who present the same PEs but no delusions. If delusions really are an adaptive response to anomalous PEs of some sort, and it is an assumption of F&C that, if not eliminated, these PEs seriously undermine the contact of a person with reality and thus his chances of survival and reproduction, it follows that people undergoing the same PEs as people with delusions but who present no delusions should be less in contact with reality (and thus stand less chances of survival) than people presenting both the PEs and the delusions.

Studies of this type can take up two forms. The first applies to psychosis. It would consist in comparing people who have already developed a delusion with people who are still in its prodromal stages to see which group is more in contact with reality and hence potentially stands more chances of survival and reproduction. Comparative studies of this kind are difficult to run, as in psychosis it is hard to observe delusions in isolation from other confounding psychotic symptoms. Nonetheless, there exist some studies which have compared people with first-episode psychosis with people who are at high risk for psychosis on measures of quality of life and cognitive functioning, finding that people at high risk for psychosis display a lower mood but also a higher degree of cognitive functioning than people in their first episode (Fusar-Poli et al. 2012; Broome et al. 2009; Bechdolf et al. 2005).

The findings of these studies show that, while psychosis ameliorates the psychological wellbeing of people still in its prodromal stage, undergoing aberrant PEs, it worsens their cognitive functioning. This would seem to speak against F&C's model: what emerges from these studies is that psychosis seems to be psychologically adaptive, as it boosts the mood of the people in the prodromal phase, but it is not biologically adaptive as intended by F&C, as it negatively affects those cognitive capacities which in the model would be so important for keeping in contact with the external environment. However, one should be careful to jump to easy conclusions by extending these results to delusions. As psychosis is a broader construct than delusions, it cannot be ruled out that the psychological relief and the cognitive deterioration brought about by psychosis are ascribable to other phenomena than delusions. In other words, it could yet be proved that it is due to delusions that people in the full-blown psychotic phase are more detached from their environment than people experiencing only anomalous PEs in the prodromal phase. Although there is still room to refute the thesis that delusions are maladaptive, these studies seem to point in a different direction from F&C's model.

The second way in which the shear pin model might be tested extends beyond psychosis. It would consist in comparing people with delusions with people who present the same psychological or neurological dysfunctions of the deluded group but no delusions. Translated into the language of predictive coding, this would equal to comparing people who present anomalous PEs and delusions with people who present only the former. If the shear pin model is correct, the latter group should be less in contact with reality and less fit for survival than the deluded group, as in the non-delusional group the shear pin would not be operative. At first sight, this looks like a more viable way to test the shear pin hypothesis than the previous one. After all, there are plenty of conditions – such as Anorexia Nervosa, depression, OCD, BDD – which present both a delusional and non-delusional variant; it seems sensible to think that, despite the presence of delusions, the delusional and non-delusional forms of each condition share the same kind of dysfunctions, which in turn give rise to the anomalous PEs. This would make the shear pin hypothesis look quite easily testable, as it would be sufficient to take, say, some individuals with depression but no delusions and compare them to their delusional counterparts to see which ones are in fact more in contact with their environment.

However, recent debates have shown how difficult can be to prove that two groups of people undergo exactly the same kind of dysfunctions. A recent discussion about Capgras and ventromedial frontal patients, involving Corlett and McKay, is a prime example in this sense (McKay 2019; Corlett 2019). Famously, Capgras and ventromedial frontal patients have been held to share the same kind of neurological impairment (i.e. a disruption in the autonomic response to familiar faces, which would cause the faces of beloved ones to be perceived as unfamiliar), with only the former presenting delusional ideation. However, Corlett has recently cast doubt on this assumption, highlighting that ventromedial frontal patients present vaster impairments of the ventromedial prefrontal cortex than Capgras patients. It would follow that ventromedial frontal patients and Capgras patients are not comparable groups, as their neurological impairments do not coincide. This debate shows how difficult it is in practice to establish if two individuals or groups of people suffer from a dysfunction which is exactly the same, and hence to prove that one group is biologically and epistemically better off than the other. Similarly, how is it possible to know if people with anorexia or depression suffer from the same kind of impairments of their delusional counterparts? Is a person with anorexia delusionally convinced of being extremely fat undergoing the same kind of neurological dysfunction or of anomalous PEs as a person with anorexia only thinking, feeling or fearing that she is fat, but in a less than delusional form?[3]

The F&C model could resist these objections by claiming that if people don't form the delusion, then that indicates that we are not talking about the same levels or kinds of PE. However, this reply is problematic. If the assumption is that people who do not form delusions are not undergoing the same kinds of PEs as people with delusions, this

[3] In a pilot study on the adaptiveness of delusions in OCD, I have tried to address this issue by comparing the epistemic functioning not of different people but of the same person before and during delusional ideation. If delusions are the by-product of an adaptive mechanism, a person that undergoes a dysfunction Y at time t1 (before delusions arise) should be epistemically worse off than at time t2, when delusions develop in response to Y.

must be shown to be true, otherwise the theory becomes unfalsifiable. Moreover, if it turned out to be very difficult to empirically prove that people with delusions undergo different kinds of PEs of people who are not deluded, then the simplest model should be preferred – and, as I have showed, this is not F&C's but its maladaptive counterpart.

Although the above considerations suggest that the shear pin hypothesis might be in practice hard to prove - either because delusions are hard to observe in isolation from other psychotic symptoms or because it is hard to tell if two people or groups suffer from exactly the same dysfunctions - it is already possible to point out that the biological adaptiveness of delusions seems to be put into question by the fact that the cognitive capacities of people with psychosis – who almost always present delusional ideation – seem to be deteriorated if compared to the cognitive capacities of people in the prodromal phase of psychosis. As good cognitive capacities are essential to keep in touch with the outside environment and exploit it for purposes of survival and reproduction, psychosis seems to be detrimental to biological adaptiveness. Although it is still premature to extend these results to delusions, these findings do not raise good prospects for the biological adaptiveness of delusions. The shear pin hypothesis needs to be built on stronger empirical evidence, and obtaining it will be no easy task.

6 Conclusion

Delusions are a complex and multifaceted phenomenon, which offers itself to various interpretations. Much of psychiatry and philosophy of psychiatry still sees delusions as a maladaptive phenomenon, something which impedes the psychological and biological flourishing of an individual. However, there is also another, though not much told, story about delusions. For some accounts, rather than a problem or a dysfunction in themselves, delusions would be an imperfect answer to a situation which is already compromised from a biological or psychological standpoint. Among these accounts, the most daring is that of Fineberg and Corlett, which argues in favour of the biological adaptiveness of delusions within a predictive coding framework. While in standard predictive coding accounts delusions are generated by anomalous PEs, in the F&C's model, they are generated by the shear pin, which is activated in response to the anomalous PEs. Hence while in standard predictive coding accounts delusions are maladaptive, the by-product of a dysfunction, in F&C's model delusions are the outcome of an adaptive process of belief formation.

However, the maladaptive view of delusions offered by standard predictive coding is simpler compared to the shear pin/adaptive model put forward by F&C: principles of parsimony and simplicity would thus suggest that the maladaptive view should be preferred to the adaptive one, unless there are some aspects of delusions which cannot be explained without resorting to the latter. In F&C's model the aspects that the maladaptive view would not capture is that delusions provide epistemic benefits to their beholders, being the only way that individuals have to be in contact with the outside environment despite the biological and psychological difficulties that they undergo. For this reason, delusions would be biologically adaptive. However, this issue cannot be settled by purely theoretical means: only empirical studies of a comparative kind can clarify whether delusions do present epistemic benefits of the kind envisaged by F&C's model and hence whether resorting to an adaptive rather than to a maladaptive view is justified. However, some studies run on people with psychosis and on people in the prodromal phase of psychosis seem to point towards a different direction: that psychosis ameliorates psychological wellbeing but has a negative impact on those cognitive capacities which are deemed to be so essential for the successful exploitation of the external environment by F&C's model. This would suggest that psychosis is psychologically but not biologically adaptive.

Although one should be careful to extend these results to delusions - as psychosis is a broader construct than delusions - it seems that these studies speak more in favour of a biologically maladaptive than of an adaptive view of delusions. Despite the fact that comparative studies of this kind are difficult to run, because it is difficult to separate delusions from other confounding factors and because it is hard to prove that two groups of people undergo the exact same kind of dysfunction, it is the only way for F&C's model to gain the upper hand over their maladaptive counterparts. The latter in fact rest on a theoretically more solid ground, as they resort to a simpler model to explain the rise and maintenance of delusions.

References

American Psychiatric Association. 2013. *DSM-V: Diagnostic and statistical manual of mental disorders (5th Rev ed.)*. Washington, DC: APA.

Antrobus, M., and L. Bortolotti. 2016. Depressive delusions. *Filosofia Unisinos* 17 (2): 192–201.

Badcock, P.B., C.G. Davey, S. Whittle, N.B. Allen, and K.J. Friston. 2017. The depressed brain: An evolutionary systems theory. *Trends in Cognitive Sciences* 21 (3): 182–194.

Bechdolf, A., R. Pukrop, D. Köhn, S. Tschinkel, V. Veith, F. Schultze-Lutter, S. Ruhrmann, C. Geyer, B. Pohlmann, and J. Klosterkötter. 2005. Subjective quality of life in subjects at risk for a first episode of psychosis: A comparison with first episode schizophrenia patients and healthy controls. *Schizophrenia Research* 79 (1): 137–143.

Bell, D. 2003. *Paranoia*. Cambridge: Icon.

Bentall, R., and S. Kaney. 1996. Abnormalities of self-representation and persecutory delusions: A test of a cognitive model of paranoia. *Psychological Medicine* 26 (6): 1231–1237.

Boorse, C. 1975. On the distinction between disease and illness. *Philosophy and Public Affairs* 5: 49–68.

Bortolotti, L. 2020. *The epistemic innocence of irrational beliefs*. Oxford: Oxford University Press.

Bortolotti, L. 2015. The epistemic innocence of motivated delusions. *Consciousness and Cognition* 33: 490–499.

Bortolotti, L. 2016. The epistemic benefits of elaborated and systematised delusions in schizophrenia. *British Journal for the Philosophy of Science* 67 (3): 879–900.

Broome, M.R., P. Matthiasson, P. Fusar-Poli, J.B. Woolley, L.C. Johns, P. Tabraham, E. Bramon, L. Valmaggia, S.C.R. Williams, M.J. Brammer, X. Chitnis, and P.K. McGuire. 2009. Neural correlates of executive function and working memory in the 'at- risk mental state'. *The British Journal of Psychiatry* 194: 25–33.

Capgras, J., and P. Carette. 1924. Illusion de sosie et complexe d'Oedipe. *Annales Medico-Psychologiques* 82: 48–68.

Coltheart, M., P. Menzies, and J. Sutton. 2010. Abductive inference and delusional belief. *Cognitive Neuropsychiatry* 15 (1): 261–287.

Corlett, P. 2019. Factor one, familiarity and frontal cortex: A challenge to the two-factor theory of delusions. *Cognitive Neuropsychiatry* 24 (3): 165–177.

Corlett, P. (2018). Delusions and prediction error. In Bortolotti, L. (Ed). Delusions in context. Palgrave Macmillan.

Corlett, P., and P. Fletcher. 2015. Delusions and prediction error: Clarifying the roles of behavioural and brain responses. *Cognitive Neuropsychiatry* 20 (2): 95–105.

Corlett, P., J. Taylor, X. Wang, P. Fletcher, and J. Krystal. 2010. Toward a neurobiology of delusions. *Progress in Neurobiology* 92 (3): 345–369.

Corlett, P., J. Krystal, J. Taylor, and P. Fletcher. 2009. Why do delusions persist? *Frontiers in Human Neuroscience* 3: 12.

Davies, M., M. Coltheart, R. Langdon, and N. Breen. 2001. Monothematic delusions: Towards a two- factor account. *Philosophy, Psychiatry, and Psychology* 8 (2/3): 133–158.

Ellis, H.D. 2003. Book review: Uncommon psychiatric syndromes. *Cognitive Neuropsychiatry* 8: 77–79.

Ellis, H.D., and A.W. Young. 1990. Accounting for delusional misidentifications. *The British Journal of Psychiatry* 157 (2): 239–248.

Enoch, M., and W. Trethowan. 1991. *Uncommon psychiatric syndromes*. 3rd ed. Oxford: Butterworth-Heinemann.

Fineberg, S., and P. Corlett. 2016. The doxastic shear pin: Delusions as errors of learning and memory. *Cognitive Neuropsychiatry* 21 (1): 73–89.

Fletcher, P., and C. Frith. 2008. Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience* 10 (1): 48–58.

Freud, S. 1986. Neurosis and psychosis. (J. Strachey, trans). In *The essentials of psychoanalysis: The definitive collection of Sigmund Freud's writing* , ed. A. Freud. London: Penguin.

Fusar-Poli, P., G. Deste, R. Smieskova, S. Barlati, A.R. Yung, O. Howes, R.D. Stieglitz, A. Vita, P. McGuire, and S. Borgwardt. 2012. Cognitive functioning in prodromal psychosis: A meta-analysis. *Archives of General Psychiatry* 69 (6): 562–571.

Gadsby, S. 2018. Self-deception and the second factor: How desire causes delusion in anorexia nervosa. *Erkenntnis* 85: 609–626.

Gadsby, S., & Hohwy, J. (2020). Why use predictive processing to explain psychopathology? The case of anorexia nervosa. In *The Philosophy and Science of Predictive Processing* (Eds.) S. Gouveia, R. Mendonça, & M. Curado. Bloomsbury.

Garety, P., and D. Hemsley. 1987. Characteristics of delusional experience. *European Archives of Psychiatry and Neurological Sciences* 236 (5): 294–298.

Hohwy, J. 2017. Priors in perception: Top-down modulation, Bayesian perceptual learning rate, and prediction error minimization. *Consciousness and Cognition* 47: 75–85.

Hohwy, J. 2014. *The predictive mind*. Oxford: Oxford University Press.

Knill, D., and A. Pouget. 2004. The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences* 27 (12): 712–719.

Langdon, R., and M. Coltheart. 2000. The cognitive neuropsychology of delusions. *Mind and Language* 15 (1): 184–218.

Lee, W., M. Wadsworth, and M. Hotopf. 2006. The protective role of trait anxiety: A longitudinal cohort study. *Psychological Medicine* 36 (3): 345–351.

Levy, N. 2018. Obsessive–compulsive disorder as a disorder of attention. *Mind & Language* 33: 3–16.

Maher, B.A. 1974. Delusional thinking and perceptual disorder. *Journal of Individual Psychology* 30: 98–113.

McKay, R. 2012. Delusional inference. *Mind & Language* 27 (3): 330–355.

McKay, R. 2019. Measles, magic and misidentifications: a defence of the two-factor theory of delusions. *Cognitive Neuropsychiatry* 24: 183–190.

McKay, R., and M. Kinsbourne. 2010. Confabulation, delusion, and anosognosia: Motivational factors and false claims. *Cognitive Neuropsychiatry* 15 (1–3): 288–318.

McKay, R., and D. Dennett. 2009. The evolution of misbelief. *Behavioral and Brain Sciences* 32 (06): 493–561.

McKay, R., R. Langdon, and M. Coltheart. 2005. "Sleights of mind": Delusions, defences, and self-deception. *Cognitive Neuropsychiatry* 10 (4): 305–326.

Mishara, A.L., and P. Corlett. 2009. Are delusions biologically adaptive? Salvaging the doxastic shear pin. *Behavioral and Brain Sciences* 32 (6): 530–531.

Miyazono, K. 2015. Delusions as harmful malfunctioning beliefs. *Consciousness and Cognition* 33: 561–573.

Miyazono, K., and R. McKay. 2019. Explaining delusional beliefs: A hybrid model. *Cognitive Neuropsychiatry* 24 (5): 335–346.

Murphy, D. 2005. Can evolution explain insanity. *Biology and Philosophy* 20 (4): 745–766.

Na, E.J., K.W. Choi, and J.P. Hong. 2019. Paranoid ideation without psychosis is associated with depression, anxiety, and suicide attempts in general population. *The Journal of Nervous and Mental Disease* 207 (10): 826–831.

Nanko, S., and J. Moridaira. 1993. Reproductive rates in schizophrenic outpatients. *Acta Psychiatrica Scandinavica* 87 (6): 400–404.

Schmitt, D., and J. Pilcher. 2004. Evaluating evidence of psychological adaptation: How do we know one when we see one? *Psychological Science* 15 (10): 643–649.

Stone, T., and A.W. Young. 1997. Delusions and brain injury: The philosophy and psychology of belief. *Mind & Language* 12: 327–364.

Teufel, C., and P. Fletcher. 2016. The promises and pitfalls of applying computational models to neurological and psychiatric disorders. *Brain* 139 (10): 2600–2608.

Wakefield, J. 1992. The concept of mental disorder: On the boundary between biological facts and social values. *American Psychologist* 47 (3): 373–388.

Williams, D. 2018. Hierarchical Bayesian models of delusion. *Consciousness and Cognition* 61: 129–147.