EuJAP | Vol. 16 | No. 2 | 2020
UDC: 165:159.972
https://doi.org/10.31820/ejap.16.2.2

# DELUSIONS IN THE TWO-FACTOR THEORY: PATHOLOGICAL OR ADAPTIVE?

Eugenia Lancellotta
University of Birmingham

Lisa Bortolotti
University of Birmingham

Original scientific article – Received: 15/6/2020 Accepted: 13/7/2020

## *ABSTRACT*

*In this paper we ask whether the two-factor theory of delusions is compatible with two claims, that delusions are pathological and that delusions are adaptive. We concentrate on two recent and influential models of the two-factor theory: the one proposed by Max Coltheart, Peter Menzies and John Sutton (2010) and the one developed by Ryan McKay (2012). The models converge on the nature of Factor 1 but diverge about the nature of Factor 2. The differences between the two models are reflected in different accounts of the pathological and adaptive nature of delusions. We will explore such differences, considering naturalist and normativist accounts of the pathological and focusing on judgements of adaptiveness that are informed by the shear-pin hypothesis (McKay and Dennett 2009). After reaching our conclusions about the two models, we draw more general implications for the status of delusions within two-factor theories. Are there good grounds to claim that delusions are pathological? Are delusions ever adaptive? Can delusions be at the same time pathological and adaptive?*

***Keywords***: *Delusions; adaptiveness; pathology, two-factor theories; delusion formation*

Correspondence: L.Bortolotti@bham.ac.uk

## 1. Introduction

Delusions are symptoms of mental disorders. Does that mean that they inherit from disorders their *pathological* status? Or should they be seen instead as emergency responses to a critical situation and thus described as *adaptive*? Could they be simultaneously pathological *and* adaptive? In this paper we are interested in the answers that the two-factor theory of delusions provides to such questions.

We are aware that delusions come in different forms and contents and that the two-factor theory has interesting things to say about all types of delusions—and other kinds of beliefs too. However, in this paper we shall refer to monothematic delusions and in particular the Capgras delusion as our standard example. This is for two reasons: (1) the two-factor theory was initially put forward to account for monothematic delusions,[1] even though its scope has been gradually extended to account for a wider range of phenomena;[2] (2) the Capgras delusion is the standard example in the papers proposing the two models of the two-factor theory we have chosen to focus on.

### 1.1. Delusions: The Pathological and the Adaptive

Delusions are unusual beliefs that are considered as symptomatic of a number of mental disorders, such as schizophrenia and delusional disorder. Monothematic delusions revolve around one theme and their content is often wildly implausible: someone with Capgras delusion believes that their spouse has been replaced by an impostor who looks just like the spouse; someone with Cotard delusion believes that they are disembodied or dead; someone with mirrored-self misidentification believes that they can see a stranger—and not their own image—in the mirror. The two-factor theory of delusion formation is a very influential theory proposing that monothematic delusions are caused by at least two factors. Factor 1 is a neuropsychological deficit responsible for anomalous data that may also result in an anomalous experience. Factor 2 is a cognitive process (described as either dysfunctional or biased) explaining either the initial endorsement of the delusional belief or the prolonged maintenance of the delusional belief in the face of mounting counterevidence. Multiple versions of the two-factor theory have been put forward, where the main difference between them lies in the description of Factor 2 and its role in the process of delusion formation.

[1] Some authors suggest that the two-factor theory is best suited to account for monothematic delusions, and that has been built around the Capgras delusions (e.g., Corlett 2019).

[2] See for instance the discussion of self-deception in McKay et al. (2005).

According to the two-factor theory, are delusions pathological? Are they adaptive? Following the most popular ways to characterise what counts as a disorder in the philosophy of medicine in general and in psychiatry in particular, a belief counts as ‘pathological’ when it is either (1) the output of a dysfunctional process (*naturalism*); (2) harmful (*normativism*); or (3) the output of a dysfunctional process and harmful (*harmful-dysfunction account*) (Bortolotti 2020). Beliefs are sometimes regarded as pathological when they deviate from some norm to which they are expected to conform—but that use of the term ‘pathological’ is an extension and we shall not consider it here.

Beliefs are usually called ‘adaptive’ if they enhance a person’s wellbeing, purpose in life, or good functioning (*psychological adaptiveness*); or if they enhance an individual’s chances of survival and reproduction (*biological adaptiveness*). It has been shown that arguments for the biological adaptiveness of delusions are less common and overall less persuasive than claims about their psychological adaptiveness (McKay and Dennett 2009; Lancellotta and Bortolotti 2019) and when some delusions are presented as psychologically adaptive, their contribution to wellbeing or good functioning is often regarded as partial or temporary. We will spend more time on the psychological adaptiveness claim simply because the biological adaptiveness thesis has been defended (to our knowledge) only within the predictive-processing account of delusion formation (Fineberg and Corlett 2016) and not within the two-factor theory. To make our task more manageable, we shall confine our attention to forms of psychological adaptiveness that are explained by a shear-pin mechanism (McKay and Dennett 2009).

## 1.2. The Shear-pin Hypothesis

According to the “shear-pin” hypothesis (McKay and Dennett 2009), some false beliefs that prevent a cognitive system from being overwhelmed can count as adaptive (*adaptive misbeliefs*). This might happen for instance when people experience such a traumatic event that they would succumb to suicidal thoughts if their negative emotions were not managed. One example is anosognosia (“denial of illness”), where a person, who has lost the use of a limb as a result of physical trauma, denies paralysis or does not acknowledge the full extent of the ensuing impairment (Ramachandran and Blakeslee 1998; McKay et al. 2005). Someone’s delusion that they can clap their hands when their right arm is paralysed would act as a motivated belief which serves to reduce the harmful impact of their new disability on their wellbeing and sense of self. McKay and Dennett (2009) suggest in their paper that, in situations of extreme stress, motivational influences are allowed to intervene in the process of belief evaluation. As a result, people come to believe what they desire to be true (“My arm is not paralysed”; “I can clap!”) and not what they have evidence for (“My arm is not moving because it is paralysed”). This is designed to permit the cognitive system to continue operating.

According to the shear-pin hypothesis, the situation in which adaptive misbeliefs emerge is already seriously compromised.

> What might count as a doxastic analogue of shear pin breakage? We envision doxastic shear pins as components of belief evaluation machinery that are “designed” to break in situations of extreme psychological stress (analogous to the mechanical overload that breaks a shear pin or the power surge that blows a fuse). Perhaps the normal function (both normatively and statistically construed) of such components would be to constrain the influence of motivational processes on belief formation. Breakage of such components, therefore, might permit the formation and maintenance of comforting misbeliefs – beliefs that would ordinarily be rejected as ungrounded, but that would facilitate the negotiation of overwhelming circumstances (perhaps by enabling the management of powerful negative emotions) and that would thus be adaptive in such extraordinary circumstances. (McKay and Dennett 2009, 501)

The person is already experiencing high levels of stress and can come to more serious harm unless their negative emotions are managed. Thus, adaptive misbeliefs prevent the situation from worsening. McKay and Dennett talk about the “extraordinary circumstances” in which motivational influences on belief are not just tolerated but desirable. Such influences intervene not by accident but by design, and this is what makes the resulting beliefs adaptive despite their falsehood.

McKay and Dennett consider the possibility that some delusions count as biologically adaptive misbeliefs but argue that in the case of delusions the extent to which desires are allowed to influence belief formation is excessive. They leave it open whether some delusions can count as psychologically adaptive.

### 1.3. The Two-factor Theory

According to Max Coltheart (2007), who is the founder of the two-factor theory, a satisfactory theory of delusions should be able to answer two questions about the genesis and maintenance of delusional beliefs:

1. Where does the delusion come from?
2. Why is the delusion adopted and then maintained in the face of disconfirming evidence?

Two-factor models of delusions provide an answer to these questions by advocating two factors in the generation and maintenance of a delusional belief (Coltheart 2007).

Factor 1 answers the first question and results in anomalous data/experience. Consider for example the Capgras delusion where the person comes to believe that a loved one has been replaced by an identical impostor. Factor 1 is an autonomic failure in the face recognition system, so when the person sees their spouse, the well-known face does not trigger the usual feelings of familiarity.[3] This generates an anomalous experience of a face which is recognised but does not *feel* familiar. On the model, Factor 1 explains the content of the delusion. Factor 1 varies from delusion to delusion and may even vary across individual cases of the same delusion. Two-factor theories hold that Factor 1 is necessary but not sufficient to explain the phenomenon of delusions. This is mainly due to the fact that there seem to be people who have the deficit playing the Factor 1-role but do not report delusional beliefs. To differentiate these cases from delusional ones, another factor (Factor 2) is required to explain the transition from the data resulting in an anomalous experience to the delusional belief. The move from not feeling that a well-known face is familiar to believing something like: “The person I see in front of me is not my spouse but an impostor” is due to a process of either endorsement or explanation of the content of the anomalous experience.

Whilst Factor 1 differs from one delusion (or person) to the next, Factor 2, broadly described as a problem in belief evaluation, is supposed to be constant across all delusions. However, two-factor theorists disagree on the precise nature of Factor 2. Some proposals identify Factor 2 with a lesion to the right dorsolateral prefrontal cortex (Coltheart et al. 2018) but there is disagreement about whether this locus is specific to delusions or shared with other neuropsychological conditions (see Tranel and Damasio 1994; Corlett 2019). Another open question about two-factor theories is whether Factor 2 contributes to the *adoption* or to the *maintenance* of the delusional belief.

[3] We are aware that the way of describing the conscious experience of people with Capgras when they look at their loved one is controversial, but we will not engage in questions about the nature of their experience as it is not relevant to our discussion. In this paper, we shall talk about their failing to experience a “feeling of familiarity”. Also, there is a debate about how to accurately characterise the content of the Capgras delusion. In this paper, we shall talk about people believing something like the following: “The person I see in front of me is not my beloved one but an impostor”.

Let us describe two competing models of the two-factor theory—the most influential and detailed—and map their differences.

### 1.4. The Coltheart Model

On what we shall refer to as *the Coltheart model* (Coltheart et al. 2010), Factor 1 is a neuropsychological deficit which results in anomalous data and can manifest at conscious level as an anomalous experience.

Factor 1 operates at the belief adoption stage. What happens at the belief adoption stage? The anomalous data are accounted for by a process of inference *to the best explanation* (abductive inference): given the very unusual nature of the data, the delusional explanation is the best possible explanation among a range of candidate hypotheses. Abductive inference is understood in Bayesian terms. Bayes' theorem stipulates the best way of choosing among candidate hypotheses to explain a given piece of evidence (O). A hypothesis (H) is more apt than another hypothesis (H') to explain O if its posterior probability is higher than the posterior probability of H'. The posterior probability of a hypothesis is the product of the hypothesis' prior probability (the probability of the hypothesis before O) and its likelihood (how likely it is to observe O if the hypothesis was true). On this account, given O, it is possible for H to be a better explanation than H' even if H has a low prior probability providing that the likelihood of H given O offsets its low prior probability.

Consider the Capgras delusion. In the Coltheart model, the impostor hypothesis ("That woman is not my wife but an impostor") can be a better explanation than the spouse hypothesis ("That woman is my wife") with regard to evidence O. Even if the impostor hypothesis has a lower prior probability than the spouse hypothesis, as impostors are not a frequent occurrence, its likelihood can be much greater than that of the spouse hypothesis, to the point of making its posterior probability higher than that of the spouse hypothesis. In this scenario, the impostor hypothesis is the most rational explanation for the absence of a feeling of familiarity: people have intact reasoning capacities when adopting the delusional hypothesis. Their reasoning is compromised when evidence against the delusional belief start accumulating.

Factor 2 is a cognitive deficit inhibiting the rejection of an endorsed belief even in the presence of strong counterevidence—Factor 2 makes the belief virtually impossible to revise. On this model, Factor 2 operates at the belief maintenance stage. What happens then, at the belief maintenance stage? On the Coltheart model, there is a second dysfunction responsible for the delusion (Factor 2) which amounts to a *deficit* in belief evaluation. This allows the delusional belief to be preserved in the face of evidence to the contrary.

In the case of Capgras delusion, the person faces overwhelming evidence against the impostor belief but that is not sufficient reason for the person to abandon or revise that belief. Evidence may include the testimony from relatives and friends confirming that the person accused to be an impostor is in fact the spouse. The person who adopted the delusional belief is unable to step back from it and to consider alternative explanations even when the belief receives serious challenges.

### 1.5. The McKay Model

Ryan McKay puts forward several objections to the Coltheart model which are important to understand his own proposal (McKay 2012), what we shall call *the McKay model*. As the objections are also relevant to our assessment of the status of delusions, we shall consider some of them here, albeit briefly.

First, the novel contribution in the Coltheart model (Coltheart et al. 2010) is that adopting the delusional hypothesis (e.g., the impostor hypothesis in the Capgras delusion) is Bayesian-rational because the hypothesis is the best explanation for the anomalous data. But for McKay the rationality of the endorsement of the delusional hypothesis is overestimated in the Coltheart model, because the model does not take into account how incredibly unlikely the state of affairs which makes up the content of the delusion is. As McKay says, it would be akin to a miracle if an impostor were to take the place of one’s spouse and be also perfectly identical to the spouse. Thus, it is not plausible to suppose that there is nothing problematic in the reasoning step that leads from the anomalous data and the resulting experience to the delusional belief.

Second, how do we account for the experiences of ventromedial frontal patients who, similarly to Capgras patients, experience an autonomic failure to familiar faces but who, differently from Capgras patients, do not adopt the impostor belief? In the Coltheart model, the assumption is that ventromedial frontal patients initially adopt the impostor belief—as the best possible explanation of the anomalous data which sometimes results in an anomalous experience—but do not maintain it. When faced with disconfirming evidence, differently from Capgras patients, they abandon the impostor belief. This can be accounted for if ventromedial frontal patients share Factor 1 with Capgras patients but not Factor 2.

McKay's objection to this proposal is that it is implausible that ventromedial frontal patients first adopt the impostor belief and then reject it. It is implausible that the spouse hypothesis is dismissed at the stage of belief adoption but then embraced once the person receives evidence against the impostor belief. The conjunction of new evidence (i.e. testimony from relatives and friends which contradicts the impostor belief) and old evidence (i.e. the absence of a feeling of familiarity which confirms the imposter belief and protestations for the alleged impostors that they are not impostors) does not favour the spouse hypothesis over the impostor belief in the circumstances. Why would the spouse hypothesis explain the total evidence any better than the impostor belief? More precisely, it is not clear why the testimony of others should radically change the distribution of likelihoods between the impostor belief and the spouse hypothesis, considering that, according to McKay, the spouse's testimony was presumably already dismissed at the stage of the adoption of the impostor belief.

A possible response in defence of the Coltheart model is that the testimony of the spouse does not count as evidence in favour of the spouse hypothesis: it is easy to see that a good impostor would still convincingly pretend to be someone's spouse even when explicitly confronted about it. The testimony of friends and family seems a more reliable source of evidence in favour of the spouse hypothesis. Hence, it might be the case that ventromedial frontal patients initially adopt the impostor belief because it is the one which best explains the evidence at hand—the absence of feelings of familiarity and the testimony of the spouse—but then correctly dismiss it in the face of the testimony of friends and family.

The third criticism of the Coltheart model is probably the most compelling. It concerns the *chronology* of Factor 1 and Factor 2. If people with Capgras delusion are unable to revise their impostor belief in the light of contradicting evidence because of Factor 2, this means that they cannot acquire Factor 2 prior or at the same time of Factor 1, otherwise they would be unlikely to abandon the spouse hypothesis and would dismiss the evidence for the impostor hypothesis (i.e., the absence of a feeling of familiarity). In other words, if people who develop the Capgras delusion are *conservative* with their existing beliefs at the maintenance stage, why should they be *revisionist* with their existing beliefs at the adoption stage? The Coltheart model seems to require that people with Capgras acquire Factor 2 *after* Factor 1, that is, after endorsing the impostor belief and before facing the testimony of family and friends which counts against it.

McKay overcomes this objection by putting forward his own model, according to which Factor 2 operates at the adoption stage, just like Factor

1: the impostor hypothesis is adopted because people suffer from a neuropsychological impairment responsible for the anomalous data and resulting in the anomalous experience (Factor 1), and because they have a bias towards *explanatory adequacy* (Factor 2) which leads them to accept hypotheses that seem to explain their experiences even when such hypotheses have low prior probability and conflict with their existing beliefs.

> An individual with a bias towards explanatory adequacy will update beliefs as if ignoring the relevant prior probabilities of the candidate hypotheses. (McKay 2012, 345)

The McKay model builds on previous work by Stone and Young (1997), Aimola Davies and Davies (2009), and McKay himself. It largely agrees with the Coltheart model about the nature of Factor 1. Factor 1 is a neuropsychological deficit and in the case of Capgras delusion it causes the absence of a feeling of familiarity towards well-known faces.

However, the model offers a different account of Factor 2. In the McKay model, Factor 2 is activated in the transition from the anomalous experience to the belief. Due to the explanatory adequacy bias, salient perceptual experience is taken at face value, causing the person to adopt a hypothesis which explains the experience in question but does not fit with the person’s previous beliefs (e.g., the impostor hypothesis in Capgras). Ventromedial frontal patients who may also fail to experience feelings of familiarity towards well-known faces (Factor 1) but who do not come up with the impostor belief may just lack the explanatory adequacy bias (Factor 2). In the model, Factor 2 is thus already present when the delusional belief is adopted whereas the Coltheart model is supposed to locate Factor 2 at the belief maintenance stage.

For McKay, given the extreme low prior probability of the impostor hypothesis, it is not rational to adopt it as an explanation of the anomalous experience, so some bias needs to be involved in the acceptance of the delusional belief. The delusion is adopted due to the fact that people discount the prior probabilities of the delusional hypothesis in favour of how well the hypothesis explains (‘fits’) the data. So, people who develop Capgras adopt the impostor belief despite its low prior probability because it matches the absence of a feeling of familiarity towards well-known faces better than the spouse hypothesis.

Here is a way of describing the difference between the McKay model and the Coltheart model: for McKay the delusion emerges when the impostor belief is adopted, as Factor 1 and Factor 2 have contributed by then to the person endorsing an unusual explanation for an unusual experience. For Coltheart and colleagues, the impostor belief is adopted as a result of Factor 1, but it becomes a delusion only when it grows resistant to counterevidence at the maintenance stage as a result of Factor 2.

### 1.6. Interim Summary and Plan

We have introduced two models of the two-factor theory, explaining how they differ (see table 1 for a summary). In section 2 we shall ask whether the models are compatible with delusions being pathological. In section 3 we shall ask whether they are compatible with delusions being adaptive.

| | *Factor 1* | *Factor 2* |
|---|---|---|
| **The Coltheart Model (Coltheart et al. 2010)** | A *neuropsychological deficit* manifesting in an unusual experience leads the person to adopt an unusual belief. | A *cognitive deficit in belief evaluation* leads the person to preserve the unusual belief in the face of counterevidence. |
| | *Factor 1 explains belief adoption and Factor 2 the belief maintenance.* | |
| **The McKay Model (McKay 2012)** | A *neuropsychological deficit* manifesting in an unusual experience contributes to the person adopting an unusual belief. | An *explanatory adequacy bias* contributes to the person adopting a belief with low prior probability. |
| | *Factor 1 and Factor 2 together explain the adoption of the delusional belief.* | |

Table 1: Differences in two influential versions of the two-factor theory of delusion formation

## 2. Are Delusions Pathological?

In this section we ask whether the claim that delusions are pathological beliefs is compatible with the two-factor models of delusions described in section 1, the Coltheart model and the McKay model. We structure the discussion around three ways in which we can understand what it means for delusions to be pathological, which map the notions of disorder defended in the philosophy of medicine: *naturalism* (the system is disordered if it is dysfunctional); *normativism* (the system is disordered if it causes harm); the *harmful-dysfunction* view (the system is disordered if it is dysfunctional and it causes harm).

### 2.1. The Naturalist View

For naturalists, the pathological nature of a delusional belief depends on whether the belief's aetiology involves a dysfunction. More precisely, the claim is that for a belief to be pathological, there must be a dysfunction in the mechanisms responsible for how the belief is adopted or maintained.

In statements about the two-factor theory of delusion formation, the words 'deficit' and 'dysfunction' are indeed used and delusions are recognised as pathological: "[W]e advocate a deficit model of delusion formation, that is, delusions arise when the normal cognitive system which people use to generate, evaluate, and then adopt beliefs is damaged" (Langdon and Coltheart 2000, 184). And again: "Essentially, we view delusion as a dysfunctional belief, a doxastic state of a particular pathological severity" (McKay et al. 2005, 315). We know by now that in the two-factor theory, the two factors are a neuropsychological deficit resulting in anomalous data/experience and, more relevant to assessing the pathology of a belief, a problem with reasoning. Factor 2 is described as a *cognitive bias* (e.g., Fine et al. 2007; Langdon et al. 2010; McKay 2012) or as a *cognitive deficit* (e.g., Coltheart 2007; Coltheart et al. 2010).[4]

In two-factor theories advocating cognitive biases, people reporting delusional beliefs are found to reason differently from people who do not, but the difference is not a disadvantage independent of the context in which the bias operates. This suggests that there is no deficit or dysfunction involved in forming the delusion given the anomalous nature of the experience. The presence of biases in the belief fixation process is not sufficient for the resulting belief to qualify as pathological, and indeed many non-pathological beliefs are the output of biased reasoning. The same bias can be beneficial in some contexts and detrimental in other contexts, and biased reasoning does not imply the presence of an underlying deficit. The McKay model is a good example of the bias approach: the problem identified in the inference from the experience to the belief (Factor 2) is an *explanatory adequacy* bias. People who have it tend to disregard a hypothesis's low prior probability if the hypothesis seems to explain well the data salient to them. The opposite tendency, often called *doxastic conservatism,* consists in resisting a hypothesis that does not fit with previous beliefs even if the hypothesis seems to explain well the data. It is a form of inertia where the person's existing model of the world is protected from change. Whether one bias or the other leads to

[4] If the only problem with the delusion was the anomalous data it explains, then one might come to the conclusion that the delusional belief itself is not pathological as there is nothing dysfunctional in the way in which belief fixation mechanisms operate.

better outcomes (the adoption and maintenance of true and rational beliefs) depends on the context. Thus, on naturalist grounds alone, delusions are not pathological in the McKay model.

In two-factor theories explicitly advocating a cognitive deficit or a doxastic dysfunction, Factor 2 is to be identified with such a deficit or dysfunction: examples would be the failure for the belief fixation system to inhibit implausible hypotheses or the failure for the belief maintenance system to abandon or revise a belief that has received disconfirmation by further evidence after its adoption. This suggests that the role of Factor 2 in the formation of delusions is sufficient for the delusion to count as pathological on naturalist grounds. The Coltheart model fits such a description: impostor beliefs may not be pathological when they are adopted, as the impostor hypothesis is the best explanation for the person's anomalous data/experience. However, the belief becomes pathological at the stage in which it is maintained in the face of powerful counterevidence, because its maintenance is due to a dysfunction affecting belief evaluation.

### 2.2. The Normativist and the Harmful-dysfunction View

Normativists agree that the pathological nature of a belief depends on whether the belief causes harm or otherwise leads to undesirable consequences for the agent—as judged by the agent or by society, depending on the preferred version of the view. Harms and disadvantages may include impaired functioning, loss of agency, negative emotions, failure to fulfil one's goals, and so on. It is plausible to claim that delusions (differently from many non-delusional irrational beliefs) are generally disruptive and can negatively affect a person's wellbeing causing impaired functioning, social isolation and withdrawal.

However, for a belief to be pathological, we would expect *the belief itself* to be the cause of harms or other disadvantages. It is not clear in the case of delusion whether the belief is the cause of the harm or disadvantage or is instead a response to a situation that is already critical for the person. The difficulty for normativism here is that what we know about so-called pathological beliefs does not usually enable us to determine whether the harm or disadvantage is caused by the beliefs themselves. Indeed, it may be caused by something else but ultimately explain why the beliefs are adopted or maintained; or it may just happen alongside the adoption and the maintenance of the belief.

For instance, on some accounts of delusions in schizophrenia, the delusion is seen as a response to the uncertainty in the prodromal phase of psychosis (e.g., Jaspers 1963; Mishara 2010). More relevant to monothematic delusions, in anosognosia the adoption of the belief that one's arm is not paralysed (say) can be seen as a reaction to the physical and psychological trauma the person experienced (e.g., Turnbull et al. 2014). In such a case, the delusion seems to be a response to a critical situation as opposed to the source of the harm or disadvantage (although the maintenance of the delusion may become a source of further harm or disadvantage). In the case of monothematic delusions like Capgras, it is not clear whether the delusion causes or is a response to harm or disadvantage: psychodynamic accounts of Capgras tended to see it as a motivated delusion, but more recent cognitive-deficit accounts do not make room for the delusion to be part of a defence mechanism (McKay et al. 2005).

There are cases in which unquestionable harm or disadvantage is associated with believing the delusional content (e.g. when the content is distressing, causing guilt, fear, or anxiety). There are also cases in which the harm or disadvantage is caused by the reaction of the surrounding social environment to the person reporting the belief: individuals whose beliefs have similar surface features may experience drastically different responses, ranging from being supported by their social circle to being vulnerable to exclusion and isolation. In sum, there is a significant link between delusions and harm or disadvantage even when a person's overall functioning is not impaired by the delusion (e.g., Jackson and Fulford 1997).

Where does this leave our two models? Are delusions pathological on normativist grounds for the two-factor theory? The most plausible answer is yes. McKay is explicit about delusions causing harm—functioning is disrupted by the extent of the mismatch between the content of the delusion and the reality as experienced by those who are non-delusional (McKay et al. 2005; McKay and Dennett 2009). Factor 1 and Factor 2 are both responsible for this mismatch, the data being anomalous and the delusional hypothesis being so implausible that it would be 'miraculous' for its content to turn out true. The Coltheart model does not explicitly discuss negative psychological consequences of the delusion but that delusions cause harm or disadvantage is often implied.

On views of the pathological nature of delusions according to which both a harmfulness condition and a dysfunction condition are combined (the so-called 'harmful-dysfunction' views inspired by the work of Jerome Wakefield), delusions still result as pathological on the Coltheart model but not on the McKay model unless Factor 2 is described as a cognitive dysfunction as opposed to a cognitive bias.

## 2.3. Summary of Section 2

The two-factor theory aims at providing an account of the pathological nature of delusions, so it is not surprising that the claim that delusions are pathological is compatible with both the Coltheart model and the McKay model (see table 2 for a summary).

| | *Naturalism* | *Normativism* | *Harmful Dysfunction* |
|---|---|---|---|
| **The Coltheart Model (Coltheart et al. 2010)** | The delusion is pathological because its maintenance is due to a cognitive dysfunction. | The delusion is pathological because its maintenance disrupts psychological functioning. | The delusion is pathological because its maintenance is due to a cognitive dysfunction and disrupts psychological functioning. |
| **The McKay Model (McKay 2012)** | The delusion is not pathological because it is due to a cognitive *bias,* not a cognitive dysfunction. | The delusion is pathological because it disrupts psychological functioning. | The delusion is not pathological because it disrupts psychological functioning but is not due to a cognitive dysfunction. |

*Table 2: Are delusions pathological?*

## 3. Are Delusions Adaptive?

In this section, we ask whether the claim that delusions are adaptive is compatible with the Coltheart model and the McKay model. In the philosophical, psychological, and psychiatric literature there have been recent explorations of the idea that some delusions may be adaptive *in some sense* (Lancellotta and Bortolotti 2019), psychologically (McKay and Dennett 2009), biologically (Fineberg and Corlett 2016), even epistemically (Bortolotti 2015; 2016).

As anticipated, we shall focus on the shear-pin hypothesis as the best (most detailed) conceptualisation of adaptiveness as applied to delusional beliefs. The shear-pin metaphor illustrates one of the ways in which delusions could be considered as adaptive. By disabling some of its parts, shear pins allow a system which is about to collapse to continue operating, albeit in an imperfect manner. In shear-pin accounts, an adaptive misbelief is the outcome of a process that is designed to prevent the collapse of the cognitive system. The misbelief is biologically adaptive if it enhances genetic fitness and psychologically adaptive if it contributes to wellbeing or good functioning. As we saw, after careful consideration, McKay and Dennett (2009) conclude that delusions are *not* biologically adaptive misbeliefs.[5] However, they do not rule out that some delusions can be psychologically adaptive.

Based on our analysis in section 2, both the Coltheart and the McKay models identify a factor responsible for anomalous data. In the Coltheart model the adoption of the belief is Bayesian-rational but its maintenance is due to a cognitive deficit; in the McKay model, the adoption of the delusion is due to a cognitive bias. Do such accounts leave room for delusions to be described as an adaptive emergency response?

### 3.1. The Coltheart Model and the Shear-pin Hypothesis

In the Coltheart model as applied to monothematic delusions such as Capgras, does the *adoption* of the unusual belief (1) emerge in the context of a crisis and (2) rescue the cognitive system from collapsing? As we saw, the unusual belief is an explanation—the best possible one—of the anomalous data brought about by Factor 1. When people lack feelings of familiarity towards a familiar face, the cognitive system produces a belief ("The woman in front of me is not my wife but is an impostor") which is false, but Bayesian-rational. The adoption of the unusual belief can hardly be interpreted as the response to a critical situation, and there seem to be no reason to believe that it would be rescuing the cognitive system from collapsing. This strongly suggests that the adoption of the unusual belief is not the outcome of a shear-pin mechanism.

Let's move now to the Coltheart model of belief *maintenance*. Does preserving the unusual belief in the face of counterevidence (1) emerge in the context of a crisis and (2) rescue the cognitive system from collapsing? In a delusion like Capgras and in the context of a deep tension between what one believes and what other people believe, remaining convinced that one's spouse has been replaced by an impostor could have some psychological benefits over believing that one has serious mental health

[5] Revisiting McKay and Dennett's shear-pin hypothesis in the light of their predictive-coding approach, Sarah Fineberg and Phil Corlett (2016) argue that the breakage of the shear pin and the consequent formation of the delusion allow an individual's cognitive system to keep functioning in the face of anomalous data. Such data, if left unexplained, would lead to the paralysis of the processes by which an individual engages in automated learning, significantly damaging the cognitive system. By explaining the anomalous data, the delusion allows automated learning to resume and the cognitive system to keep functioning. However, the cost is that all anomalous data are likely to be interpreted through the lens of the delusional belief which become more entrenched as the default explanation.

issues. Continuing to believe that one has veridical experiences and is the victim of a malicious third party (i.e., the impostor) would help preserve one's positive self-image, whereas acknowledging that one's experience is unreliable and gave rise to an implausible belief would not. In the light of this, Factor 2 could be interpreted as the sign that the shear pin has broken. If the goal is to salvage the cognitive system at the cost of disabling some of its parts, Factor 2 could be understood as the cost—the disabling of the capacity for belief evaluation.

However, the compatibility of the Coltheart model with the shear-pin hypothesis is compromised by the model branding Factor 2 as a cognitive dysfunction. Factor 2 emerges as a deficit in belief evaluation—an inability to revise one's existing beliefs in the face of disconfirming evidence. Due to such a deficit, the belief becomes resistant to counterevidence and is preserved. Factor 2 cannot be a shear-pin mechanism because it is characterised not as a design feature, but as a dysfunction, and thus the delusional belief cannot be regarded as adaptive.

What we can say, then, is that the shear-pin hypothesis is incompatible with belief adoption in the Coltheart model, because belief adoption does not respond to a crisis, and could be compatible with belief maintenance in the Coltheart model if the delusion were not branded as the outcome of a dysfunction. The delusion would be a design feature which prevents the system from collapsing.

### 3.2. The McKay Model and the Shear-pin Hypothesis

We saw that McKay sees the delusion as irrationally formed, that is, as a non-optimal explanation of the anomalous data caused by Factor 1. The main difference with the Coltheart model is that Factor 2 gets activated at the belief adoption stage rather than at the maintenance stage. Thus, we need not distinguish between belief adoption stage and belief maintenance stage in the McKay model because both Factor 1 and Factor 2 operate at the belief adoption stage and the unusual belief qualifies as a delusion then.

In the McKay model, then, do delusions (1) emerge in the context of a crisis and (2) rescue the cognitive system from collapsing? As with the Coltheart model, in the Capgras case the adoption of the delusion can hardly be interpreted as the response to a critical situation, and there seem to be no reason to believe that it would be rescuing the cognitive system from collapsing. Rather, the adoption of delusions is the outcome of a cognitive bias operating on anomalous data. When people with Capgras lack feelings of familiarity towards a familiar face, the cognitive system produces a belief (“The woman in front of me is not my wife but is an impostor”) which is false, but “fits” those feelings.

Can delusions more generally be seen as the output of a shear-pin mechanism in the McKay model? For the shear-pin hypothesis to apply, there needs to be a crisis the delusion is a response to (e.g., overwhelming negative emotions to manage) and this response prevents the cognitive system from collapsing. It is well known that unexplained anomalous experiences may generate uncertainty (Fineberg and Corlett 2016) and by providing an explanation of those experiences, delusions would contribute to relieve the ensuing anxiety. An example of a delusion that could be explained by the shear-pin hypothesis is the Reverse Othello syndrome (McKay et al. 2015). After recently becoming disabled, a man comes to believe that his previous partner is still in love with him and that they married, whereas his partner has moved on and is in another relationship. The realisation that his partner had left him on top of the many other changes caused by his new disability might have led the man to depression and even suicide, threatening the continued functioning of his cognitive system. In this case, it is easy to see how the shear-pin could intervene to avoid the collapse of the person’s cognitive system. The adoption of the delusion (e.g., “My partner and I still are in a happy relationship”) could be interpreted as a sign that the shear pin has broken: the man’s desires have been permitted to exercise a powerful influence on his beliefs (see also Mele 2006). In the instance of Reverse Othello syndrome examined by McKay (Butler 2000), the man then gradually abandoned the conviction in the delusional belief that his former partner still loved him and had become his wife which suggests that the delusion did not have long-term negative consequences for the man’s functioning. However, in an alternative hypothetical case in which the delusion persisted after the initial crisis had been managed, the delusion might have lost its adaptive role and become a serious hindrance.

Our conclusion is that the shear-pin hypothesis is compatible with the McKay model, because the adoption of the delusion is not due to a cognitive dysfunction, and the delusion can in some contexts be formed as a response to a crisis that prevents the cognitive system from collapsing. That said, the Capgras would not a be a good example of a delusion that is the outcome of a shear-pin mechanism and even for other types of delusions for which the shear-pin hypothesis is more plausible, it is not clear that the psychological benefits of adopting the delusion outweigh the potential long-term costs of maintaining the delusion.

### 3.3. Summary of Section 3

The two models of the two-factor theory we are discussing do not explicitly address the question whether delusions are adaptive, although Ryan McKay has considered the question elsewhere (McKay and Dennett 2009). It is an interesting issue, though, whether the two-factor theory is compatible with the claim that delusions are adaptive at least in the short-term, a claim that is not implausible for at least some delusions in some contexts.

We argued that the McKay model can make room for a shear-pin explanation of the adaptive nature of some delusions, whereas for the Coltheart model things get trickier (see table 3). We also observed that the overall plausibility of claims about delusional beliefs being adaptive cannot be generalised and depends on the content of the delusional belief and the context in which it emerges.

| **Delusions as adaptive outputs of a shear-pin breakage** | |
|---|---|
| **The Coltheart Model (Coltheart et al. 2010)** | The maintenance of the delusion in the face of counterevidence could be a response to a crisis that prevents the cognitive system from collapsing so it could be due to a shear-pin breakage. However, this is not compatible with the belief being the outcome of a cognitive dysfunction. |
| **The McKay Model (McKay 2012)** | The adoption of some delusions is a response to a crisis that prevents the cognitive system from collapsing so it could be due to a shear-pin breakage. This is compatible with those delusions being the outcome of a cognitive bias. |

*Table 3: Are delusions adaptive?*

## 4. Conclusions and Implications

We asked what two influential models of the two-factor theory of delusion formation have to say about the potential pathological nature and adaptiveness of delusions, with a special focus on monothematic delusions such as Capgras. Throughout, we made some observations which have implications for further investigations into the nature of delusions.

First, delusions can be pathological on a normativist reading of disorder, where delusions simply need to be harmful to count as pathological, although it is not clear that delusions are always the source of harm as opposed to a response to an existing crisis that causes harm (Bortolotti 2015). Some delusions may enable the person to cope with adversities and preserve their self-esteem (Gunn and Bortolotti 2018). In one case, Barbara started believing that God was communicating with her by telepathic messages because she was his child and she was good: "as God was talking to me he was making sure that I knew there was nothing wrong with me. And he's always there, whether I'm right, whether I'm wr… well, he, he says I'm never wrong, God says I'm never wrong". Barbara developed the delusion after hearing voices for some time and her delusional belief may be considered as an explanation for her unusual experiences. Furthermore, Barbara's belief that she was special and that God was supporting her followed a very difficult time in her life, when her unfaithful husband had left her permanently and she was feeling both vulnerable and guilty about earlier decisions she made in her life. In the short term, the delusion might have protected Barbara from negative feelings about herself and prevented a suicidal attempt which was on her mind.

It is even more dubious that we can base the pathological nature of delusions on a naturalist or harmful-dysfunction reading of disorder, where delusions need to be the outcome of a dysfunctional process to count as pathological. That is because we cannot easily show that the cognitive process responsible for delusion formation is a dysfunctional process in itself as opposed to a cognitive process that operates in non-ideal conditions (such as a process whose input is the outcome of a dysfunction, a process affected by biases or performance errors, etc.).

Second, whether delusions are the outcome of a shear-pin breakage is also very difficult to ascertain in general terms. It is possible that a shear-pin mechanism works to protect a person's cognitive functioning by relieving that person from the anxiety which comes with anomalous experiences, helping the person manage negative emotions, or salvaging the person's positive self-image. However, whether the alleged benefits ever outweigh, even temporarily, the costs of having the delusion is by no means obvious and needs further examination. Some progress could be made with the issue whether delusions are psychologically adaptive if it were possible to compare the psychological profile of people with delusions with the psychological profile of people who have the same experiences as people with delusions but develop no delusions. If delusions are an emergency response which is devised to help in the face of a crisis, then people facing the same crisis as people with delusions but with no delusions should be psychologically worse off. This would help clarify if delusions are the problem or the imperfect solution to a problem (Lancellotta forthcoming).

Finally, one interesting upshot of our investigation is that in a version of the two-factor theory of delusions the same belief can be adaptive and pathological (though not at the same time). This marks an important difference between the Coltheart model and the McKay model. In the McKay model, some delusions can prevent the person's cognitive system from breaking down at the time of their adoption (and thus be adaptive as the outcome of a shear-pin breakage) *and* disrupt the person's psychological functioning in the long-term (and thus count as pathological on a normativist account). However, in the Coltheart model, delusions cannot be adaptive *and* pathological, because by being the outcome of a dysfunctional process and counting as pathological in a naturalist and harmful-dysfunction sense, the possibility that they are also the outcome of a shear-pin mechanism which breaks by design is ruled out.

## REFERENCES

Davies, A. M., and M. Davies. 2009. Explaining pathologies of belief. In *Psychiatry as Cognitive Neuroscience*, eds. M. Broome and L. Bortolotti, ch. 15. Oxford: Oxford University Press.

Bortolotti, L. 2015. The epistemic innocence of motivated delusions. *Consciousness & 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.

Bortolotti, L. 2020. Doctors without disorders. *Aristotelian Society Supplementary* 94(1): 163–184.

Butler, P. V. 2000. Reverse Othello syndrome subsequent to traumatic brain injury. *Psychiatry* 63(1): 85-92.

Clutton, P., and S. Gadsby. 2018. Delusions, harmful dysfunctions, and treatable conditions. *Neuroethics* 11: 167–181.

Coltheart, M., R. Cox, P. Sowman, H. Morgan, A. Barnier, R. Langdon, E. Connaughton, L. Teichmann, N. Williams, and V. Polito. 2018. Belief, delusion, hypnosis, and the right dorsolateral prefrontal cortex: A transcranial magnetic stimulation study. *Cortex* 101: 234-248

Coltheart, M. 2007. Cognitive neuropsychiatry and delusional belief. *Quarterly Journal of Experimental Psychology* 60: 1041–1062.

Coltheart, M., P. Menzies, and J. Sutton. 2010. Abductive inference and delusional belief. *Cognitive Neuropsychiatry* 15(1-3): 261-287.

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

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.

Jackson, M., and K. W. Fulford. 1997. Spiritual experience and psychopathology. *Philosophy, Psychiatry, and Psychology* 4: 41-65.

Lancellotta, E. forthcoming. Is the biological adaptiveness of delusions doomed? The case of predictive coding. *Review of Philosophy and Psychology*.

Lancellotta, E., and L. Bortolotti. 2019. Are clinical delusions adaptive? *WIREs Cognitive Science* 10: e1502.

Jaspers, K. 1963. *General Psychopathology*. Manchester: Manchester University Press.

Gunn, R., and L. Bortolotti. 2018. Can delusions play a protective role? *Phenomenology and the Cognitive Sciences* 17, 813–833.

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

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

McKay R. T., and D. C. Dennett. 2009. The evolution of misbelief. *Behavioral and Brain Sciences* 32(6): 493-561.

Mele, A. 2006. Self-deception and delusions. *European Journal of Analytic Philosophy*, 2(1), 109-124. https://hrcak.srce.hr/91611

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.

Mishara, A. 2010. Klaus Conrad (1905-1961): Delusional mood, psychosis, and beginning schizophrenia. *Schizophrenia Bulletin* 36: 9-13.

Ramachandran, V. S., and S. Blakeslee. 1998. *Phantoms in the Brain: Probing the Mysteries of the Human Mind.* New York: William Morrow.

Sakakibara, E. 2016. Irrationality and pathology of beliefs. *Neuroethics* 9(2): 147–157.

Turnbull, O., A. Fotopoulou, and M. Solms. 2014. Anosognosia as motivated unawareness: The 'defence' hypothesis revisited. *Cortex* 61: 18-29.