博士論文

論文題目　**Delusions as Malfunctioning Beliefs　A Biological Defense of Doxasticism about Delusion**

（信念の機能不全としての妄想　生物学的な観点から妄想の信念主義を擁護する）

氏名　宮園　健吾

Delusions as Malfunctioning Beliefs

A Biological Defense of Doxasticism about Delusion

by

Kengo Miyazono

Acknowledgments

I thank teachers and friends in the two philosophy departments in UT; the department of philosophy at Hongo and the department of philosophy and history of science at Komaba. I especially thank Masaki Ichinose, Richard Dietz, Yukihiro Nobuhara, Koji Ishihara, John O'Dea, Shuhei Shimamura, Shun Tsugita and Ryo Uehara.

I am very grateful, more than I can express in words, to Lisa Bortolotti and Tim Bayne for their help and encouragement. As the authors of wonderful works on delusion and as the speakers of UTCP Lecture Series (Lisa in 2011, Tim in 2012), they convinced me that I should discuss delusion in my dissertation. As the mentors during the actual writing process, they helped me to focus on really important issues, rethink about many of my claims and stay on the right track.

I got insightful comments, questions and criticisms, from the early stage to the final stage, by many philosophers and cognitive scientists including Alex Byrne, Mathew Broome, Greg Currie, Phil Corlett, Daniel Dennett, Tamar Gendler, Sung-il Han, Anna Ichino, Adrian Kwek, Sam Liao, Matteo Mamelli, Ruth Millikan, Agustin Rayo, Marga Reimer, Ema Sullivan-Bissett, Eric Schwitzgebel and Nick Zangwill. My colleagues at Birmingham also gave me important feedbacks at several occasions.

I learned a lot during my visit to Yale (2008-2010) and MIT (2011-2012). I am grateful to George Bealer and Stephen Yablo for accepting me in their departments. I also thank Todai-Yale Initiative and Program for Evolving Humanities and Sociology for financially supporting my visit to Yale and MIT respectively.

I thank all the questions, criticisms and comments at the seminars and conferences where I presented the ideas from this dissertation;

- "Delusion and Harmful Dysfunction Analysis", Philosophy of Medicine Seminar, February 11, 2014, at King's College London
- "Delusion Formation and Bayesianism", iCog Inaugural Conference, December 1, 2013, at University of Sheffield
- "A Theory of Belief for Doxasticism about Delusion", The 2nd PLM Conference, September 13, 2013, at Central European University
- "Bayesian Approaches to Delusion Formation?", Philosophy of Psychiatry Work in Progress Day, June 10, 2013, at University of Lancaster
- "A Theory of Belief for Doxasticism about Delusions", Birmingham Philosophy Society Seminar, March 11, 2013, at University of Birmingham
- "A Theory of Belief for Delusion Doxasticists", The First Conference on Contemporary Philosophy in East Asia, September 7-9, 2012, at Academia Sinica, Taipei
- "Can Doxasticism about Delusion Be Defended?", 2012 Early Career Philosophy Researcher Forum, July 22, 2012, at National Olympics Memorial Youth Center, Tokyo, (in Japanese)
- "Delusions as Malfunctioning Beliefs", The 38th Meeting of Society for Philosophy and Psychology, June 21, 2012, at University of Colorado at Boulder, (poster)
- "The Role of Imagination in Delusion: Two Hypotheses", The 86th Annual Meeting of American Philosophical Association, Pacific Division, April 6, 2012, at Westin Seattle, Seattle, WA, (poster)
- "Delusions as Malfunctioning Beliefs", Harvard/MIT Friends and Eminees Discussion Group, March 6, 2012, at Harvard University

- "Delusions as Malfunctioning Beliefs", The 20th International Meeting of Hongo Metaphysics Club, February 21, 2012, at The University of Tokyo
- "Delusions as Malfunctioning Beliefs", Tim Bayne Lecture Series at UTCP, February 16, 2012, at University of Tokyo
- "Reality-Monitoring Failures and Metacognitive Account of Delusion", 2011 International Neuroethics Society Annual Meeting, November 10, 2011, at Carnegie Institution for Science, Washington DC, (poster)
- "Delusion and Belief", 2011 General Meeting of The Japan Association for the Philosophy of Science, June 4, 2011, at Ehime University, (in Japanese)

The appendix "Two-Factor Theory and Prediction-Error theory" is based on the materials from the book chapter that I wrote with Lisa Bortolotti and Mathew Broome for *Aberrant Beliefs and Reasoning* (edited by Niall Galbraith, Psychology Press, 2014). But, substantial changes and revisions have been made.

My greatest thanks go to Mie, Rikka, Kenichi and Keiko Miyazono.

Table of Contents

Chapter 1: Doxasticism and Causal Role

1.1 Introduction

Doxasticism about delusion (DD) is the claim that delusions are beliefs. There are some *prima facie* reasons for DD (see 1.2), and DD is widely accepted in psychiatry. DSM-5, for instance, defines delusions as an abnormal belief.

> Delusion: A false *belief* based on incorrect inference about external reality that is firmly sustained despite what almost everyone else believes and despite what constitutes incontrovertible and obvious evidence to the contrary. The *belief* is not ordinarily accepted by other members of the person's culture or subculture (i.e., it is not an article of religious faith). When a false *belief* involves a value judgment, it is regarded as a delusion only when the judgment is so extreme as to defy credibility. (emphasis added) (American Psychiatric Association 2013, 819)

Recently, however, a number of psychiatrists and philosophers argued that many delusions do not behave like beliefs. In other words, many delusions do not play the causal roles that are characteristic to beliefs (let us call them "belief-like causal roles"). Egan, for instance, wrote; "delusions, it turns out, display a lot of or that doesn't look terribly belief-like. […] The role that delusions play in their subjects' cognitive economies differs

pretty dramatically from the role that we'd expect beliefs to play." (Egan 2009, 265-266) Let us call this claim, the claim that many delusions fail to play belief-like causal roles, "causal difference thesis" or "CDT" for short.

Now, there is a clear tension between DD and CDT. If CDT is true and many delusions fail to play belief-like causal roles, how can it be the case that those delusions are beliefs? Again, if DD is true and delusions are beliefs, then how can it be the case that many of them fail to play belief-like causal roles? This tension is especially acute given the fact that (broadly) functionalist conception of belief (including functionalism (Armstrong 1968; Lewis 1980), dispositionalism (Schwitzgebel 2001; Ryle 1949), representationalism (Fodor 1987; Nichols & Stich 2003), and, presumably, interpretationism (Davidson 1984; Dennett 1989)) is influential in the philosophical literature on belief. According to functionalist conception, playing a belief-like causal role is necessary for a mental state to be a belief. Thus, there is no such thing as the belief without belief-like causal role.

One might think, on the basis of the functionalist conception of belief, that DD and CDT are incompatible with each other and, hence, at least one of them should be rejected. This "incompatibilism" is dominant in the recent literature on delusion. There are two types of incompatibilists. First, there are those who reject DD and accept CDT (Berrios 1991; Currie 2000; Currie & Jureidini 2001, 2003; Currie & Ravenscroft 2002; Egan 2008; Frankish 2009, 2012; Hohwy & Rajan 2012; Schwitzgebel 2012; Tumulty 2011, 2012). They are "anti-DD

incompatibilists". Anti-DD incompatibilists argue that, at least, many delusions are not beliefs, because they fail to play belief-like causal roles. Second, there are those who reject CDT and accept DD. They are "pro-DD incompatibilists". Pro-DD incompatibilists try to defend the idea that delusions are beliefs by rejecting the claim that delusions fail to play belief-like causal roles (Bayne & Pacherie 2005; Bortolotti 2010, 2011, 2012; Bortolotti & Broome 2012; Reimer 2010).

The main aim of this dissertation is to develop an option that hasn't been explored seriously so far, namely, a compatibilist option. Comatibilists are those who think that DD and CDT are compatible with each other. Certainly, DD and CDT are incompatible with each other according to the functionalist theories of belief. But compatibilists argue that they are compatible with each other under some different theories of belief and those alternatives theories are, at least, as plausible as functionalist ones for independent reasons.

In Chapter 1 and 2, I will introduce DD and CDT with more details and present some reasons for these claims. In addition, I will motivate compatibilist option by showing (1) that there is no good argument against the possibility of mental states without their distinctive causal roles, (2) that there are some examples, other than delusions, of mental states without their distinctive causal roles, and (3) that incompatibilist options (anti-DD and pro-DD) face some difficulties.

The compatibilist theory that I will present in Chapter 3 is called "teleo-attitude functionalism". The core idea of this theory is that beliefs are characterized in the same way that biological organs such as hearts or kidneys are characterized. For the same reason that there can be hearts that fail to pump blood, there can be beliefs that fail to play belief-like causal roles, according to teleo-attitude functionalism. This means that DD and CDT are compatible with each other according to this theory. Hearts that fail to pump blood are "malfunctioning hearts". Analogously, the beliefs that fail to play belief-like causal roles are "malfunctioning beliefs".

Chapter 4 is the discussion on some theoretical issues about teleo-attitude functionalism. I will discuss the nature of teleological function that teleo-attitude functionalism is relying on. I will also reply to some expected objections to the theory, including philosophical objections with imaginary cases and empirical objections involving empirical considerations.

Chapter 5 discusses some implications. First, I will argue that DD is likely to be true, not merely compatible with CDT, according to teleo-attitude functionalism. Second, I will argue that teleo-attitude functionalism, when combined with Wakefield's influential view on the nature of disorder, gives the best account of the pathological nature of delusion.

The appendix, which is coauthored with Lisa Bortolotti and Matthew Broome, discusses the process of delusion formation in detail. In particular, it examines the relationship between two prominent theories of delusion formation, namely, two-factor theory and

prediction-error theory. We will argue that, on the contrary to a popular view, those two views are theoretically compatible with each other, and suggest some particular ways in which two views are actually combined together.

1.2 Doxasticism about Delusion

1.2.1 What is Doxasticism?

Here are some descriptions of delusional subjects.

> **Persecutory Delusion / Grandiose Delusion**
> Lawrence was a 34-year-old history graduate who lived with a supportive partner. He had received a diagnosis of schizo-affective disorder. Since his first admission 11 years previously, he had been admitted to hospital on nine occasions. Despite continuing problems and a poor absence record, he held a full-time job as a clerical officer. At first assessment Lawrence reported that he was being personally threatened by evil forces. He said that he felt that poison was put into his food, that people at work were staring at him, and gathering together to talk about him. The persecutory experiences tended to occur episodically, but when they occurred Lawrence would be very preoccupied for around three or four days. […] Lawrence also described times when he thought he had special powers, including the ability to influence events and to use telepathic contact to communicate with others, and times when his actions were being controlled by others. Unfortunately, episodes of this type often had very serious consequences and had led to compulsory hospitalizations. (Fowler, Garety, & Kuipers 1995, 4)

> **Capgras Delusion**
> DS was a 30-year-old Brazilian man who had been in a coma for three weeks following a head injury (right parietal fracture) sustained in a traffic accident. During the subsequent year, he made remarkable progress in regaining speech, intelligence, and

other cognitive skills. He was brought to us by his parents principally because of his tendency to regard them as imposters. When we first saw him he appeared to be an alert and fairly intelligent young man who was not obviously hysterical, anxious or dysphoric. A 'mini' mental status exam (serial sevens, three objects, writing, orientation in time and place, etc.) revealed no obvious deficits in higher functions, and there was no evidence of dementia. The most striking aspects of his disorder were that he regarded his father as an 'imposter' and he had a similar, although less compelling, delusion about his mother. When asked why he thought his father was an imposter his response was 'He looks exactly like my father but he really isn't. He's a nice guy, but he isn't my father, Doctor'. (Hirstein & Ramachandran 1997, 438)

Anosognosia for Hemiplegia

"Patient L.A.-O (clinical record NA 472, 1980) was a 65-year-old, right-handed woman who was admitted to the emergency department of our hospital on the evening of 2 July 1980. Shortly before admission she had suddenly developed left hemiplegia without loss of consciousness. Alert and cooperative, she claimed that the reason for her hospitalization was sudden weakness and annoying paresthesia of the *right* limbs; her narrative, supplied in a mild state of anxiety, was indeed accompanied by sustained message of the allegedly hyposthenic right inferior limb. She also claimed that the left hand did not belong to her but had been forgotten in the ambulance by another patient. On request, she admitted without hesitation that her left shoulder was part of her body and *inferentially* came to the same conclusion as regards her left arm and elbow, given, as she remarked, the evident continuity of those members. She was elusive about the forearm but insisted in denying ownership of the left hand, even when it had been passively placed on the right side of her trunk. She could not explain why her rings happened to be worn by the fingers of the alien hand." (Bisiach & Geminiani 1991, 32-33)

According to DD, delusions are beliefs; Lawrence *believes* that he is personally threatened by evil forces, DS *believes* that his father was replaced by an imposter, and LA-O *believes* that the left hand doesn't belong to her. It is not the case, for instance, that Lawrence *imagines* that he is personally threatened by evil forces, DS *fears* that his father was

replaced by an imposter, or LA-O *desires* that the left hand doesn't belong to her. (Of course, DD allows for the possibility, for instance, that LA-O believes that the left hand doesn't belong to her *and* desires it at the same time. The claim here is that, according to DD, it is not the case that she *just* desires that the left hand doesn't belong to her.)

Here are some preliminary remarks on DD.

(1) *Clinical and Non-Clinical Delusion*: DD is a claim about delusion in clinical sense. In other words, DD says that the mental states that are referred to as "delusions" in clinical context are beliefs. The case descriptions above give typical examples of delusions in clinical context. Delusion, in this sense, arises as the symptom of varieties of pathological conditions including schizophrenia, brain injury, Alzheimer's disease, etc. Although there is no uncontroversial definition of delusion in this sense, the terminology itself is well established and widely shared. The term "delusion" is often used in English outside clinical context as well such that false or ungrounded beliefs or ideas are delusions. In this non-clinical sense, delusion includes self-deception, daydream, religious belief (e.g. *The God Delusion* by Dawkins), superstition, obsolete scientific theories such as phlogiston theories or Aristotelian physics, etc. Those "delusions" have nothing to do with DD unless they overlap delusions in clinical sense (e.g. clinical delusions with religious content).

(2) *Kinds of Doxasticisms*: There are several different ways to interpret the claim that delusions are beliefs, depending on the interpretation of what the view is actually about.

Concept doxasticism is the view about the concept "belief". According to concept doxasticism, the concept "belief" is applicable to, for instance, Lawrence when he sincerely claims that he is personally threatened by evil forces. *Attribution doxasticism* is about the third-personal attribution of belief (based upon observable behavior). According to attribution doxasticism, beliefs are appropriately third-personally attributed to delusional subjects (based upon observable behavior). For instance, the belief that his father was replaced by an impostor is appropriately third-personally attributed to DS (based upon his observable behavior). *Mental state doxasticism* is about mental state of belief, not about the concept "belief" or the third-personal belief-attribution. According to mental state doxasticism, the state of mind, for instance, LA-O is in when she sincerely claims that the left hand doesn't belong to her is belief. In other words, the kind of the state of mind she is in is the same as the kind of the state of mind I am in when I sincerely claim that there is a bottle of beer in the fridge.

These distinctions are important. But, they have been largely neglected in the previous philosophical discussions on DD. The truth of DD partly depends on the views about what belief is. Thus, we need to fix which type of doxasticism we are interested in in order to determine which type of views about belief is relevant. In general, the views about the nature of the concept "belief" are relevant when we are interested in concept doxasticism. The views about the nature of third-personal belief-attribution are relevant when we are interested in attribution doxasticism. The views about the nature of mental state of belief

are relevant when we are interested in mental state doxasticism.

The distinctions between different doxasticisms are also important when we think about the role of experimental philosophical studies on DD. The study done by Rose et al. (forthcoming) reveals that folk belief-attribution style is quite consistent with DD. In other words, there is a strong tendency among folks to attribute belief to delusional subjects, including the cases where the delusions do not seem to play belief-like causal roles. This study gives a strong and direct support for attribution doxasticism. It also gives a good support for concept doxasticism, assuming that the belief-attribution in this case is not influenced by non-conceptual factors. This does not give, however, a direct support for mental state doxasticism. Indeed, the philosophers who are primarily interested in beliefs in the mental state level (e.g. teleosemanticists) tend to accept the possibility that folk belief-attribution of belief is dissociated from the nature of belief in the mental state level.

In this dissertation, my primary focus is on mental state doxasticism. In other words, I am primary interested in what type of mental state delusional subjects are in. DD is widely accepted in psychiatry, and I take DD in psychiatry to be primarily about what mental state delusional subjects are in. And I want my discussion on DD to be about the same kind of DD that psychiatrists are primarily interested in. It is pretty unlikely that psychiatrists are primarily interested in the concept "belief" or third-personal belief-attribution. In a sense, only philosophers care about those issues. Of course, psychiatrists use mental concepts in

thinking about delusional subjects, and they attribute mental states to delusional subjects. But, this doesn't mean that they are *interested in* concepts or mental state attribution. Hereafter, whenever I use the term "DD" or "doxasticism", I will simply be referring to mental state doxasticism.

(What is the relationship between concept doxasticism and attribution doxasticism? The recent debate on so-called "Knobe effect" gives a good example to illustrate the difference. Knobe (2003) found, in his experimental philosophical studies, that the attribution of intentional action is strongly influenced by the ethical status of the action involved. In particular, it is found that people tend to regard the side-effect of the action of a subject as being intentionally caused when the side-effect is ethically problematic. Experimental philosophers all agree that Knobe effect shows that the attribution of intentional action is sensitive to ethical factors. But, they have debated over whether or not the concept "intentional action" is sensitive to ethical factors. Knobe consistently argues that Knobe effect reveals how the concept "intentional action" works. In other words, the effect is the manifestation of the sensitivity of the concept "intentional action" to ethical factors. Critics deny this. They argue, for instance, that Knobe effect is explained by conversational implicature, that it is explained by the hypothesis that ethical factors switch the interpretation of "intentional action", that it is explained by some implicit beliefs about the nature of intentional action, and so on. For the same reason that those critics can

consistently deny that the concept “intentional action” is sensitive to ethical factors without denying that the attribution of intentional action is sensitive to ethical factors, one can consistently deny, for instance, that the concept “belief” is applicable to delusional subject without denying that belief is appropriately third-personally attributed on delusional subjects.)

(3) *Eliminativism about Belief?*: Some people might reject DD on the ground that the mental state of belief doesn’t exist at all. This is the view held by eliminativists about belief (Churchland 1981). Eliminativist challenge is especially relevant to mental state doxasticism, because eliminativism is the claim about the mental state of belief, not about belief attribution or the concept “belief”. (After all, no one can sensibly deny the existence of the concept “belief” or the practice of third-personal belief attribution.) In her discussion on DD, Bortolotti avoids eliminativism and related issues by simply restrict her discussion on attribution doxasticism. “Belief is a folk-psychological concept that might have correlates at different levels of explanation (scientific psychology, neuroscience, etc.), or turn out to be just a useful construct for the efficient exchange of information and the management of social relationship that characterize contemporary humans. No metaphysical issues about beliefs will be addressed in detail, because I take it that questions about the way in which we ascribe beliefs make sense independently of a complete theory of what beliefs are, over and beyond their role in folk-psychological practices" (Bortolotti 2010, 2).

Although eliminativism is especially relevant to mental state doxasticism, I don't have any interesting things to say about it in this dissertation. Presumably, DD is best regarded as a disguised conditional claim; if something is a belief, then delusions are beliefs too. More precisely, if what we usually take as beliefs (e.g. my belief that there is a bottle of beer in the fridge) are beliefs, then delusions are beliefs too. Eliminativism is perfectly compatible with both positive and negative responses to this conditional, for the same reason that indeterminism is compatible with both positive and negative responses to the conditional claim that if determinism is true, then there is no free will. The conditional version of DD successfully captures the core claim of doxasticists. Doxasticists are, after all, those who claim that delusions are the same type of mental states as what usually take as beliefs (and, thus, if what we usually take as beliefs are beliefs, then delusions are beliefs too).

1.2.2 Some Reasons for DD

According to DD, delusions are beliefs. This view is supported by some *prima facie* reasons.

(1) *Sincere Assenting*: Delusional subjects sincerely assent to their delusions. And sincere assenting to "*p*" gives us a *prima facie* reason to think that the assenter believes that *p*. The assumption used here is very similar to what Kripke calls "disquotational principle".

Disquotational principle

If a normal English speaker, on reflection, sincerely assents to "*p*", then he believes that *p*.

Kripke says that "the principle appears to be self-evident truth" (Kripke 2011, 138). One might wonder, however, if we can really apply this principle, or similar principles, to delusional subjects. One might ask, for instance, "Are the deluded normal speakers of English?" or "Are the assents really based on reflections?" Answers to these questions certainly depend on what are actually meant by "normal speaker" and "on reflection". How should we interpret these phrases? At least, Kripke himself uses these phrases in such a way that disquotational principle may well be applicable to delusional subjects.

> When we suppose that we are dealing with a normal speaker of English, we mean that he uses all words in the sentence in a standard way, combines them according to the appropriate syntax, etc.: in short, he uses the sentence to mean what a normal speaker should mean by it. [...] The qualification "on reflection" guards against the possibility that a speaker may, through careless inattention to the meaning of his words or other momentary conceptual or linguistic confusion, assert something he does not really mean, or assent to a sentence in linguistic error. (Kripke 2011, 137-138)

If this is what "normal speaker" and "on reflection" mean, then disquotational principle may well be applicable to delusional subjects. It doesn't seem to be true that, for instance, when Lawrence assents to the sentence "Poison is put into the food" or "People are staring at me",

he does this only because he is using the sentence in a very different way (and, hence, he is not a normal speaker), or he does this only because of careless inattention or momentary conceptual or linguistic confusion (and, hence, this assent is not on reflection).

However, it is tempting to think that some cases of delusion actually involve conceptual or linguistic problems. For instance, it is tempting to think that a subject with Cotard delusion, LU, who claims that she is dead while admitting that she can move and speak just like living people doesn't really understand the concept of death (McKay & Cipolotti 2007). If she doesn't understand the concept, her assenting to the sentence "I am dead" isn't the expression of her belief that she is dead. She can't have that belief because having the belief requires having the concept of death, which she lacks.

This is certainly a possibility. But, this might not be what is actually going on in Cotard delusion. Referring to the case of LU, Bortolotti argues that the conceptual confusion hypothesis is not very likely to be true (Bortolotti 2010). LU seems to understand the concept of death because she has minimal knowledge about death (e.g. dead people don't move and speak.). What is peculiar about her condition is that even though she knows that dead people don't move and speak, and even though she admits that she can move and speak, she maintains her delusion that she is dead (although she is aware of the tension among her commitments and the recognition of the tension seems to have some impacts on her conviction). This is certainly a remarkable situation, but, as Bortolotti points out, it would

have nothing to do with conceptual or linguistic confusion. Again, on a similar case of Cotard delusion, Campbell argues: "You might propose that what we have here is not, strictly speaking, belief at all, but "empty speech" masquerading as belief. The trouble with this diagnosis is that there are perfectly sincere assertions made by people who seem to understand what they are saying, who may indeed act on the basis of what they are saying" (Campbell 2001, 91).

(2) *Pathological Nature of Delusion*: Delusion is certainly a pathological phenomenon. And the truth of DD seems to be a part of the reason why delusion is pathological. Delusion is pathological partly because, for instance, LA-O seriously believes that the left hand doesn't belong to her. If she doesn't believe it, but rather, for instance, merely imagines it, there might not be anything particularly pathological about it. It is certainly a strange thing to imagine, but we can easily entertain various kinds of strange and unrealistic possibilities in our imagination without losing mental health. As Hume says, "[n]othing is more free than the imagination of man; and though it cannot exceed that original stock of ideas furnished by the internal and external senses, it has unlimited power of mixing, compounding, separating, and dividing these ideas, in all the varieties of fiction and vision." (Hume 1748/2007, 34) Again, there might not be anything particularly pathological if LA-O doesn't believe it, but rather merely desires it. It will certainly be an uncommon desire, but it might be understandable, for example, if there is something about the hand that she really dislikes.

A teenage girl who doesn't like the shape of her nose might desire, without losing mental health, that her nose doesn't actually belong to her, and a better-looking one belongs to her instead.

(3) *Distinguishing Delusion from Other Abnormal Conditions*: DD enables us to distinguish delusion from other kinds of abnormal conditions in principled ways. For example, one can distinguish delusion from hallucination by saying that the former is abnormal belief, while the latter is abnormal perceptual experience. In fact, this is the normal way in which diagnostic manuals and psychiatry textbooks introduce the distinction between them. Again, DD enables us to distinguish delusion from impairments in linguistic ability such as jargon aphasia. Subjects with jargon aphasia make strange utterances because of the impairment in the ability to generate well-formed words, phrases and sentences. On the other hand, Lawrence, for instance, makes strange claims, such as "I can communicate with other people by telepathy", not because he is suffering from linguistic impairments, but because he actually believes strange things. Furthermore, DD might be helpful in distinguishing delusions from obsessive thoughts. It might be that, unlike delusional subjects, the subjects with obsessive thoughts about contamination by gems do not actually believe the contamination. In fact, it has been pointed out that the subjects with obsessive thoughts typically have more acute awareness of the strangeness of the thoughts than typical delusional subjects (Bortolotti 2010, 42). This fact fits well with the hypothesis

that obsessive thought doesn't involve belief, while delusion does.

(4) *Practice in Psychiatry*: There is a widespread practice of describing delusions as beliefs in psychiatry and related fields. We find the descriptions in wide-range of places from introductory textbooks, research papers to diagnostic manuals. And, I think that this practice has to be taken into account seriously, especially because this is the practice among the group of people who are familiar with what delusion is, including those who have direct acquaintances with subjects with delusions. If delusions are not beliefs but, for instance, imaginations, then it means that thousands of psychiatrists have been committed to a serious error in the interpretation of mental states for such a long time. But, how can that happen?

(5) *Self-Knowledge/Self-Understanding*: It is not just psychiatrists who describe delusions as beliefs. Delusional subjects themselves do so too. For instance, if I ask Lawrence "Do you believe that poison was put in your food?", he will say "Yes, I do." He would never answer "No, I don't. I am just imagining it" or "No, I don't. I am just desiring it." In fact, Lawrence's case description includes; "[…] he described himself as unable to stop *believing* that the telepathic contact and his mission were real." (emphasis added) (Fowler, Garety & Kuipers 1995, 4) DD seems to be a part of the self-understanding of the delusional subjects.

Of course, self-understanding or self-knowledge is fallible. So, I am not arguing that,

since delusional subjects regard their delusions as beliefs, delusions are in fact beliefs. It is not very difficult to find everyday cases (e.g. too positive self-image) and empirical studies (e.g. various kinds of confabulation experiments) where one has inaccurate self-knowledge. At the same time, though, it is too extreme to think that self-knowledge doesn't give any evidence whatsoever about the first order target states. For instance, we don't usually conclude that perception doesn't give any evidence whatsoever about external environment on the ground that there are many everyday cases (e.g. everyday hallucination) and psychological illusions (e.g. Ponzo illusion) where one has inaccurate perceptual experiences. In many cases, self-knowledge is a good indicator of first order target states, in a similar way that perception is a good indicator of external environment in many cases. Presumably, the right thing to say here would be that the fact that the delusional subjects regard their delusions as beliefs gives a *prima facie* reason for DD.

1.3 Causal Difference Thesis

CDT says that many delusions do not play belief-like causal roles. To say that a mental state plays a belief-like causal role is to say that the state causes something and is caused by something in one of the distinctive belief-like ways. Describing what belief-like causal roles actually are wouldn't be very easy, and different philosophers may have different

understanding of belief-like causal roles. Still, there are, at least, two constraints on belief-like causal role that are widely accepted implicitly or explicitly.

Belief-like causal roles have to form a natural collection such that we can reasonably say that they are the same kind of causal roles. Otherwise, it wouldn't be reasonable to group them together and give them a single name "belief-like causal role". And, of course, the natural collection always has to include the paradigmatic ones. So, when a mental state plays a certain causal role, R, but there is no natural collection of causal roles including R and the paradigmatic belief-like causal roles, then R is not a belief-like causal role.

Naturalness

A mental state plays a belief like causal role only if there is a natural collection of causal roles that includes the causal role played by the state and paradigmatic belief-like causal roles.

There are various kinds of propositional attitudes, such as belief, imagination, desire, etc. And, correspondingly, there are paradigmatic belief-like causal roles, paradigmatic imagination-like causal roles, paradigmatic desire-like causal roles, etc. Presumably, a belief-like causal role doesn't have to be a paradigmatic one for the same reasons that a bird doesn't have to be a paradigmatic bird. (A penguin is a bird, but not a paradigmatic one.) But, a belief-like causal role has to be, at least, relatively similar to the paradigmatic belief-like causal roles than to paradigmatic imagination-like causal roles, paradigmatic

desire-like causal roles, etc. Otherwise, there wouldn't be any interesting sense in which the causal role is belief-like, rather than imagination-like or desire-like. So, when a mental state plays a certain causal role, R, but R is not more similar to paradigmatic belief-like causal roles than to paradigmatic imagination-like or desire-like causal roles, then R is not a belief-like causal role.

Comparative Similarity

A mental state plays a belief like causal role only if the causal role played by the state is more similar to paradigmatic belief-like causal roles than to other kinds of paradigmatic causal roles.

Now, CDT seems to be supported by clinical observations.

(1) *Delusion – Evidential Inputs*: Typically, delusions lack belief-like sensitivity to evidential inputs. Especially, delusion's insensitivity to counterevidence is widely acknowledged, and this is even sometimes regarded as one of the essential, definitional properties of delusion. According to DSM-5, delusion "is firmly sustained despite what almost everyone else believes and despite what constitutes incontrovertible and obvious evidence to the contrary." (American Psychiatric Association 2013, 819) So, for instance, it is very difficult to remove Lawrence's delusion that poison was put into the food by showing him some counterevidences such as the result of chemical analysis. He might, for example, explain it away by not very plausible ideas such as the idea that the chemical analysis is

indeed a fake. Similarly, in a case reported in (Davies et al. 2001, 136), a 33-year-old male subject with chronic schizophrenia seriously claims that he doesn't have any internal organs. "Although his doctors had told him that this was a physiological impossibility, and despite some acknowledgement on the part of the patient that he could not quite understand how such thing was possible, the patient said that he could not rid himself of the belief." The subject maintains his delusion about the internal organ despite of the overwhelming counterevidence (e.g. testimony, physiology), and despite of the recognition of the counterevidence as counterevidence on his part.

(But, this doesn't mean that rational persuasion doesn't have any impact on delusion. At least some delusions can be influenced by rational persuasion at least in some degree. The best example is cognitive behavioral therapy for delusions whose effectiveness has been empirically supported (e.g. Brakoulias et al. 2008; Drury et al. 2000). The effectiveness of cognitive behavioral therapy shows that, in the appropriate settings and appropriate relationships between clinicians and patients, clinicians can influence, by rational persuasion, the preoccupation and conviction of delusional subjects.)

(2) *Delusion - Behavioral Outputs*: Delusions often lack belief-like impact on non-verbal behavior. This is not always the case, but it happens frequently enough. A classic observation of this is given by Bleuler,

> Other patients imagine themselves transformed into animals and even into things, and yet they usually do not adhere to one idea. Just as a patient can be the Pope, the Emperor, the Sultan, and eventually God in one person, he can also be a pig and a horse. Nevertheless, the patients rarely follow up the logic to act accordingly, as, for instance, to bark like a dog when they profess to be a dog. Although they refuse to admit the truth, they behave as if the expression is only to be taken symbolically, in the same way perhaps as when a man is insultingly called a pig. (Bleuler 1924, 140)

Again, Stone and Young wrote, about Capgras delusion;

> […] although in some cases of Capgras delusion patients act in ways that seem appropriate to their beliefs, in many other cases one finds a curious asynchrony between the firmly stated delusional belief and actions one might reasonably expect to have followed from it. […] This failure to maintain a close co-ordination of beliefs and actions may be typical of the delusions that can follow brain injury. (Stone & Young 1997, 334)

For example, when one believes that his wife is replaced by an imposter, the belief is expected have some impact on his non-verbal behavior. For example, we expect that he refuses to live with the "imposter", call police, go out in search for the real wife, and so on. In many cases of Capgras delusion, however, the patients do not show these expected non-verbal behavior.

(Some delusions do have significant impact on non-verbal behavior. For instance, Lawrence, when preoccupied with his persecutory thoughts, "would sometimes act on this beliefs, occasionally becoming aggressive and confrontational with work colleagues." And,

again, “[w]hen preoccupied with feelings of special powers, Lawrence would write letters to political and religious leaders. More seriously, he would also persistently pester people whom he thought to be good, sometimes even complete strangers.” (Fowler, Garety, & Kuipers 1995, 4)

(3) *Delusion - Other Mental States*: Delusions often lack belief-like coherence with other mental states.

(3.1) *Other Beliefs*: Delusions are often incoherent with other beliefs of the subjects. LU, a subject with Cotard delusion whom I mentioned earlier, maintains her commitment to the delusion that she is dead. At the same time, she recognizes that she can move and speak and also that dead people can’t move and speak. In general, so-called “circumscribed delusions” have this sort of features. Circumscribed delusions are not elaborated and do not have global impact on other beliefs of the subjects. Typically, monothematic delusions (i.e. delusions with single themes) are circumscribed. What is peculiar about circumscribed delusions is that, even though the delusions are not very sensible against the background of other beliefs they have, delusions and other beliefs somehow live together without competing with each other. In other words, the delusions do not cause the revision of other beliefs such that the delusions are sensible against the background of the revised beliefs, nor other beliefs cause the revision of delusion such that insensible thoughts are eliminated.

(On the other hand, elaborated delusions do cause the revision of other beliefs such that

the delusions are sensible against the background of revised beliefs .For instance, the subject, whom I mentioned above, with the delusion that he doesn't have any internal organs "expressed the belief that spirit doctors had come to his room one night to perform a magical operation in order to remove his internal organs. This happened, he believed, because he was being punished by God for some evil or sin that he had committed, although he was uncertain about the nature of the sin" (Davies et al. 2001, 136). What is going on here seems to be that the original delusion that he doesn't have internal organs has significal impacts on other beliefs, and his entire worldview is infected by delusional ideas.)

(3.2) *Affective States*: Delusions often fail to have belief-like impact on affective responses. For instance, Bovet and Parnas report the following case; "One of our patients, a 50-year-old female with paranoid schizophrenia and delusional ideas, which she in no way enacted, lived peacefully with her mother in a small Swiss town which she had apparently never left, helping with house- and garden-keeping. She expressed her paranoid ideas about her sister, which she maintained for years *quietly and without anger*" (emphasis added) (Bovet & Parnas 1993, 588). More impressive examples are found in Capgras patients. If I seriously believe, for instance, that my wife is missing and replaced by an imposter, I will be anxious, scared, worried, etc. And, presumably, I will be uneasy about the presence of imposter. But, in many cases, Capgras patients are happy with their situations. Some of them are friendly enough to "imposters", and others even express strong positive affective

feelings toward them (Christodoulou 1977; Wallis 1986). Lucchelli and Spinnler report a case where a Capgras patient, who believed that his wife Wilma had been replaced by a "double", "never became angry or aggressive. Even in the presence of the "false" Wilma, his behavior did not differ in any significant way from his usual (for instance, he did not show any difference or hesitancy to share his everyday life with her). However, he was adamant that she could not possibly be his wife, although he was never able to explain his conviction." (Lucchelli & Spinnler 2007, 189)

(Again, this is not always that case. For instance, Christodoulou reports a case where a housewife, JD, with Capgras delusion about her daughters "refused to talk to her daughters and expressed fears that the 'doubles' would poison her." (Christodoulou 1977, 557) He also reports another case where another housewife, PK, with Capgras delusion about her husband "reported to the police that her husband had died and that an identical-looking man had taken his place. She put on black dress in mourning of her 'late' husband, refused to sleep with his 'double' and angrily ordered him out of the house, shouting 'go to your own wife'." (ibid., 558))

Those features above are present in some delusions but not in others (insensitivity to evidence is, however, a universal feature of all delusions). Let us call these features "delusional features" and the delusions with some of delusional features "problematic delusions". The problematic-ness is a matter of degree. The degree of problematic-ness of a

delusion depends on the number of delusional feature it has (e.g. a delusion is insensitive to evidence and it doesn't have impact on non-verbal behavior) as well as the degree to which it has delusional features (e.g. a delusion is extremely insensitive to counterevidence).

Now, the observations above seem to support CDT to a significant degree. Look at the following two cases.

Hannah's Belief

Hannah believes that there is a bottle of beer in the fridge. The belief was caused by the perceptual experience she got few minutes ago when she was checking the fridge. The belief makes her happy. In other words, the belief causes a positive affective response. The belief, in conjunction with the desire for beer, causes her to go to the fridge to get the beer. And the belief, in conjunction with the belief that if the beer is in the fridge then there is no beer in the storage, causes another belief that there is no beer in the storage.

Martin's Imagination

Martin imagines that there is a bottle of beer in the fridge without believing it. He imagines this even though he saw few minutes ago that there is NO beer in the fridge. Unlike corresponding belief, this imagination doesn't make him very happy. The imagination doesn't cause as strong affective response as the one caused by corresponding belief. The imagination doesn't cause him to go to the fridge to get the beer despite of his desire for beer. The imagination doesn't cause him to believe that there is no beer in the storage despite of his belief that if the beer is in the fridge, then there is no beer in the storage. In other words, he doesn't use the imagination as the premise of a *modus ponens* inference.

Arguably, Hannah's belief plays a paradigmatic belief-like causal role, and Martin's imagination plays a paradigmatic imagination-like causal role.

Now, according to the naturalness condition, there has to be a natural collection of causal roles that includes the causal roles of problematic delusions and the causal role of Hannah's belief, if the problematic delusions play belief-like causal roles. But, are there any such collections? Can we really group them together to form a natural collection? This would be a matter of degree. But, to the extent that the collection of causal roles is unnatural, we will lose the confidence in the idea that problematic delusions play belief-like causal roles.

Again, according to comparative similarity condition, the causal roles of problematic delusions have to be more similar to the causal role of Hannah's belief than the causal role of Martin's imagination, if these delusions play belief-like causal roles. But, is that really true? Hannah's belief is sensitive to evidence (i.e. she believes that there is a bottle of beer in the fridge because she saw that there is a bottle of beer in the fridge), while Martin's imagination is not (i.e. he imagines that there is a bottle of beer in the fridge although he saw that there is NO bottles of beer in the fridge). Hannah's belief has impact on her non-verbal behavior (i.e. she believes that there is a bottle of beer in the fridge, and she goes to the fridge to get beer when she wants beer.), while Martin's imagination doesn't (i.e. he imagines that there is a bottle of beer in the fridge, but he doesn't go to the fridge to get beer when he wants beer). Again, Hannah's belief causes strong affective response (i.e. she is very happy), while Martin's imagination doesn't (i.e. he is not very happy). But, then, it is not clear at all that the causal roles of problematic delusions are more similar to the causal role

of Hannah's belief than the causal role of Martin's imagination (Currie 2000; Currie & Jureidini 2001; Currie & Ravenscroft 2002; Egan 2008). In other words, it is not clear at all that comparative similarity condition is satisfied by problematic delusions.

In short, we have two arguments for CDT here.

Argument from Naturalness

(1) A mental state plays a belief-like causal role only if there is a natural collection of causal roles that includes the causal role played by the state and paradigmatic belief-like causal roles. (Naturalness Condition)
(2) It is not the case that there is a natural collection of causal roles that includes the causal role played by problematic delusions and paradigmatic belief-like causal roles.
(3) Therefore, problematic delusions fail to play belief-like causal roles.

Argument from Comparative Similarity

(1) A mental state plays a belief like causal role only if the causal role played by the state is more similar to paradigmatic belief-like causal roles than to other kinds of paradigmatic causal roles. (Comparative Similarity Condition)
(2) It is not the case that the causal roles played by problematic delusions are more similar to paradigmatic belief-like causal roles than to other kinds of paradigmatic causal roles.
(3) Therefore, problematic delusions fail to play belief-like causal roles.

Chapter 2: Motivating Compatibilism

2.1 Multiple-Functionability of Mental States

There are some reasons for DD and CDT. However, there is a clear tension between DD and CDT. If CDT is true and many delusions fail to play belief-like causal roles, how can it be the case that those delusions are beliefs? Again, if DD is true and delusions are beliefs, then how can it be the case that many of them fail to play belief-like causal roles? In this dissertation, I will develop a compatibilist view according to which DD and CDT are in fact compatible with each other. This section gives some motivations for exploring this option.

The first motivation comes from the fact that there is no good argument against the possibility of mental states without their distinctive causal roles. Let us use the term "multiple-functionability" as referring to the possibility that some mental states fail to play their distinctive causal roles.

According to functionalist theories, mental states are not multiply-functionable; there is no belief without belief-like causal roles, no desire without desire-like causal roles, no fear without fear-like causal roles, and so on. But, I don't find any good argument against multiple-functionability of mental states. One might think that the multiple-realizability

argument, which is the standard argument for functionalism, rules out the multiple-realizability of mental states. But, this is not true. The multiple-realizability argument doesn't necessarily rule out multiple-realizability, because there might be some theories of mental states that allow for multiple-realizability and multiple-functionablity at the same time. For instance, Lewis (1980) proposed such a theory. Lewis believes that the following two famous cases are possible. The first case, Martian pain, is about multiple-realizability of pain, and the second case, mad pain, is about the multiple functionability of pain.

Martian Pain

there might be a Martian who sometimes feels pain, just as we do, but whose pain differs greatly from ours in its physical realization. His hydraulic mind contains nothing like our neurons. Rather, there are varying amounts of fluid in many inflatable cavities, and the inflation of any one of these cavities opens some valves and close others. His mental plumbing pervades most of his body—in fact, all but the heat exchanger inside his head. When you pinch his skin you cause no firing of C-fibers—he has none—but, rather, you cause the inflation of many smallish cavities in his feet. When these cavities are inflated, he is in pain. And, the effects of this pain are fitting: his thought and activities are disrupted, he groans and writhes, he is strongly motivated to stop you from pinching him and to see to it that you never do again. In short, he feels pain but lacks the bodily states that either are pain or else accompany it in us.

Mad Pain

There might be a strange man who sometimes feels pain, just as we do, but whose pain differs greatly from ours in its causes and effects. Our pain is typically caused by cuts, burns, pressure, and the like; his is caused by moderate exercise on an empty stomach. Our pain is generally distracting; his turns his mind to mathematics,

facilitating concentration on that but distracting him from anything else. Intense pain has no tendency whatever to cause him to groan or writhe, but does cause him to cross his legs and snap his fingers. He is not in the least motivated to prevent pain or to get rid of it. In short, he feels pain but his pain does not at all occupy the typical causal role of pain.

Since both of these cases are possible, according to Lewis, a successfully theory of pain needs to allow these possibilities. Lewis's proposal is this.

Statistical Functionalism: Pain
x feels pain iff *x* is in a physical state of the type whose tokens statisitically-normally play pain-like causal roles in the appropriate population to which *x* belongs.

The theory certainly allow for those possibilities. First, Martian Pain is possible because it is perfectly possible that the Martian is in a physical state of the type whose tokens statistically-normally play pain-like causal roles in the appropriate population to which she belongs (the population of Martians). For instance, maybe, the states of having some specific cavities inflated statistically-normally play pain-like causal roles in the population of Martians. Second, Mad Pain case is possible because it is perfectly possible that the mad guy is in a physical state of the type whose tokens statistically-normally play pain-like causal roles in the appropriate population to which she belongs (human population). For instance, maybe, the mad guy is in the state of C-fiber-firing when he is in pain, and the states of C-fiber-firing statistically-normally play pain-like causal roles in human population.

I am not going to examine this theory of pain here (I will come back to this in Chapter 3). My point is just that the multiple-realizability is, theoretically speaking, perfectly compatible with the multiple-functionability and this theory perfectly demonstrates it.

Another possible argument against multiple-functionability would be something like this. If there is such thing as belief without belief-like causal roles, then how can we know that it is a belief? After all, we usually figure out what mental states people are in by looking at its causes and effects. For instance, I come to think that *x* believes that there is a bottle of beer in the fridge on the basis of the fact that *x* just saw that there is a bottle of beer in the fridge. Or, I come to think that *y* believes that it is raining outside on the basis of the fact that *y* is behaving as if it is that case that it is raining outside. If there are some beliefs without belief-like causal roles, then we will never know that these are beliefs.

I don't find the argument very compelling. The argument seems to assume that it is necessary for a mental state to be a belief that some person can figure out that it is a belief from a third-personal point of view. I don't see any reasons to accept this assumption. First of all, I don't see any reason to accept the idea that it is necessary for a mental state to be a belief that anyone can figure out that it is a belief. Why can there be beliefs that no one can ever find out that it is a belief? Second, I don't see any reason to accept the idea that a belief needs to be such that someone can figure out that it is a belief from a third personal point of view. Why, for instance, isn't it sufficient that the subject of the belief herself can figure out

that it is a belief? Certainly, it would be difficult to identify, from outside, a belief as a belief if it fails to play a belief-like causal role. But, it might not be difficult for the subject herself to identify it as a belief from inside. (Indeed, some philosophers argue that introspective self-knowledge of propositional attitudes is not relying on the observation of the causal roles they play (e.g. Goldman 2006).)

Against a similar argument for the claim that believers need to be minimally rational (i.e. believers need to be minimally rational because otherwise we can't figure out what they belief from outside), Sober wrote;

> The idea that there is some necessary connection between belief and rationality obtains its plausibility from the fact that we seem to have no other way of figuring out what someone believes, other than by assuming that the person is at least minimally rational. But if this is the basis of our conviction that belief implies core rationality, we ought to be suspicious of it, since it is really a form of disguised operationalism. We would have no truck with the claim that temperature is necessarily connected with thermometer readings, or intelligence with IQ tests. We assume that any scientifically respectable property or magnitude must be multiply accessible; there should be no such thing as the only way of finding out if it applies. So we might conjecture that belief, if it is to be a scientifically respectable property, ought to be decoupled somewhat from the postulate of core rationality. There should be other methods of attributing beliefs which do not assume that the subject is minimally rational. These other methods might then yield belief attributions which could conflict with those obtained via the assumption of core rationality, and it would then be a matter of the overall cogency of theory which attributions to take seriously. (Sober 1985, 168)

If Sober is correct, all scientifically respectable properties and states are multiply accessible

in principle. If belief is a scientifically respectable state, then, it is also multiply assessable in principle. There is no such thing as the only way to attribute belief. So, there must be some other ways of attributing beliefs than observing their causes and effects.

2.2 Similar Phenomena

The second motivation for the compatibilist option comes from the fact that the multiply-functioning mental states might not be very uncommon after all. In other words, there are some other cases where mental states fail to play their distinctive causal roles. Here are some examples.

Unlike normal beliefs, some delusion lack belief-like impact on non-verbal behavior and affective responses. Pain asymbolia is quite similar to delusions in this respect. Pain asymbolia is a rare condition, caused by lesions in the posterior insula. Here is a description of 6 patients with pain asymbolia.

> Although all 6 patients could adequately recognize stimuli and distinguish sharp from dull, all of them showed a lack of response to painful stimuli applied over the entire body. Neither superficial nor deep pain stimulation elicited a motor withdrawal, grimacing, or an appropriate emotional response. One patient not only failed to show a withdrawal response but also exhibit a reaction of "approach" to the painful stimuli (i.e., he directed his limb toward the noxious stimuli). Inappropriate emotional responses were common: 4 patients smiled or laughed during the pain testing procedure. This abnormal behavior ceased abruptly on discontinuing stimulation. All

> patients appeared quite unaware of their abnormal reactions and seemed unable to learn appropriate escape of avoidance responses. None of them became anxious or angry during the pain testing procedure; in fact, while all could recognize pain, none of them reported any unpleasant feeling. Patient showed normal autonomic reactions (tachycardia, hypertension, sweating, mydriasis) during the painful stimulation, but failed to react with a flinch, blind, or adequate emotional responses to threatening gestures presented to both hemispaces. Five patients also failed to react to verbal menaces. (Berthier, Starkstein, & Leiguarda 1988, 43)

Given the fact that all of the patients recognize that they are in pain and they show normal autonomic responses, it would be difficult to deny that they are in pain. Nonetheless, their pains don't seem to have pain-like impact on behavior as well as affective responses. Presumably, pain asymbolia involves pains without pain-like causal roles.

Unlike normal beliefs, many delusions can coexist with other beliefs that are incoherent with them. Addictive desires are pretty similar to delusions in this respect. As Holton and Berridge pointed out, the causal roles of addictive desire significantly deviate from the causal roles of normal desires.

> The desire becomes insulated from factors that, in normal intentional behavior, would undermine it, and so persists even when the addict knows that acting on it would be highly damaging. The addict may recognize that taking the drug again will incur the loss of family, friends, job, and most that makes life worth living, and yet still continue to take it. […] There is another way in which an addictive desire doesn't not typically function like a desire to see the Pyramids or to get a paper finished before the weekend. It does not serve as an input to deliberation, something to be weighed, along with other competing desires, in deciding what to do. Instead addictive desire functions as something more like an intention: as something that, unless checked, will lead, in a rather direct way, to action. This combination of features – the insulation of

> addictive desires from factors that should undermine them, and their tendency to lead directly to action – means that addictive behavior is very different from ordinary behavior that results from deliberation. (Holton & Berridge 2013, 245)

Addictive desires, according to Holton and Berridge, coexist with other factors that undermine normal desires, in a quite similar way in which delusions coexist with other beliefs that are incoherent with them. In addition, addictive desires do not serve as the input to deliberation in the way that normal desires do. Holton and Berrige argue that addictive desires behave more like intention, which do not serve as the input for deliberation and directly linked to action. Presumably, addictive desires are the desires without desire-like causal roles.

Unlike normal beliefs, delusions are not very sensitive to evidence. Anxiety disorders are quite similar to delusion in this respect. Normally, anxiety is activated by the signs of dangers or threats. In anxiety disorders, the anxiety (or other negative emotions) ceases to be sensitive to the signs. At least, anxiety ceases to be a proportionate response to anticipated danger or threats. Presumably, anxiety disorders involve anxieties without anxiety-like causal roles.

> A 36-year-old teacher is referred to a Community Mental Health Team by her GP. She is worried about her physical health, but physical examination and other tests by her GP have found no abnormalities. She describes episodes where her heart pounds, she feels hot and faint, and has an overwhelming need to escape. This first happened during a staff meeting at school, and again whilst in a large supermarket. Now she is

apprehensive about going out in case she experiences another attack. (Clark et al. 2005)

A 19-year-old girl presented after her cleaning rituals had so exhausted her that she had given up and could now enter only two of the five rooms in her flat. For more than a year she has worried that if her house is not sufficiently clean, her young son will become ill and could die. Having touched a surface she has to disinfect it repeatedly – a procedure performed in a particular way and taking several hours. In addition, she repetitively washes her hands and sterilizes all the crockery and cutlery before eating. She realizes that she is 'going over the top', but she cannot stop thinking that items may have germs on them. This leads to disabling anxiety and fear for her son's health, which she can only resolve by cleaning. This helps temporarily, but soon the thoughts return again. (ibid.)

More examples of this sort will be found from psychopathological cases. But, I stop here. In the following, I will not say anything about those phenomena explicitly. But, I will be implicitly assuming that what I will be saying on delusion can potentially be applicable to those phenomena too.

2.3 Incompatibilisms

2.3.1 Anti-DD Incompatibilism

The third motivation for compatibilism comes from the fact that incompabilist options might not be fully satisfactory. In the following, I will discuss anti-DD incompatibilism and pro-DD incompatibilism in turn, and point out some worries.

Anti-DD incompatibilists resolve the tension between DD and CDT by rejecting DD. In other words, they deny that delusions are beliefs. Now, if delusions are not beliefs, what are they?

Anti-DD incompatibilist presented a number of alternatives to DD.

Currie and corroborators (Currie 2000; Currie & Jureidini 2001; Currie & Ravenscroft 2002) argue that delusions are not beliefs but imaginations. At the same time, those imaginations are misidentified as beliefs by delusional subjects. In other words, delusional subjects are imagining, not believing, the content of their delusions and, at the same time, they have a false meta-belief that they believe the content of their delusions. This metacognitive failure is hypothesized as being caused by the failure of the recognition of the self-generatedness of imagining.

> What is it for an imagining to be treated as if it were a belief, when in fact it is not one? What we suggest happens is this: The deluded subject fails to monitor the self-generatedness of her imagining that P. Because of this, her idea that P presents itself as something generated by the world beyond the self. In the kinds cases we are currently considering, it would be natural for it to seem that the idea that P is directly a response to the precipitating experience, as something peculiarly connected with and, hence, validated by the experience. (Currie & Jureidini 2001, 160)

Egan (2009) agrees that many delusions do not fail to play belief-like causal roles, but he also thinks that many of them fail to play imagination-like causal roles. According to Egan,

delusions are not beliefs nor imaginations, but "bimaginations". Bimagination is the intermediate mental state in-between belief and imagination.

> Classifying delusions as straightforward, paradigmatic cases of belief is problematic because it predicts that delusions ought not to display the sorts of circumscription and evidence independence that they in fact display. Classifying them as straightforward, paradigmatic cases of imagination is problematic because it predicts that they should display *more* circumscription and evidence-independence than they in fact display. What would be nice would be to be able to say that the attitude is something inbetween paradigmatic belief and paradigmatic imagination – that delusional subjects are in states that play a role in their cognitive economies that is in some respects like of a standard-issue, stereotypical belief that P, and in other respects like that of a standard-issue stereotypical imagining that P. (Egan 2009, 268)

Currie and Jones (2006) express a quite similar view.

> […] delusions are considered as a class of states do not fit easily into rigid categories of either belief or imagination. While delusions generally have a significant power to command attention and generate affect, they vary a great deal in the extent to which they are acted upon and given credence by their possessors. In that case it may be that cognitive state do not sort themselves neatly into categorically distinct classes we should label 'beliefs' and 'imaginings', but that these categories represent vague clustering in a space that encompasses a continuum of states for some of which we have no commonly accepted labels. (Currie & Jones 2006, 312)

Schwitzgebel proposes another in-between proposal. According to Schwitzgebel, delusions are what he calls "in-between beliefs". In other words, it is not the case that delusions are beliefs, but it is not the case either that they are non-beliefs.

> Is it possible, then, that cases of delusions are, at least sometimes (when the functional role or dispositional profile is weird enough), cases in an in-betweenish gray zone – not quite belief and not quite failure to believe? (Schwitzgebel 2012, 15)

Against the background of process 1/ process 2 distinction in dual-processing theory, Frankish (2009, 2012) introduces the distinction between level 1 belief and level 2 belief. "To have the level 1 belief that *p* is simply to be disposed to have in ways that would be rational on the assumption that *p* is true, given one's other beliefs and desires (level 1 belief ascription is holistic). […] By contrast, to have a level 2 belief that *p* is to be committed to a policy of taking *p* as a premise in one's conscious reasoning and decision making" (Frankish 2012, 24). According to Frankish, delusions are level-2 beliefs. This explains why problematic delusions do not have strong impact on non-verbal behavior. In general, level-2 beliefs do not have strong impact on non-verbal behavior. (Frankish's view can be understood as a version of DD. But, it is, at least, different from the simple version of DD.)

> […] if delusions are beliefs, which type are they, level 1 or level2? The obvious answer is that they are level 2, at least initially—a view that accords with Brotolotti's focus on conscious belief. Delusions are typically ascribed on the basis of the patient's avowals, rather than inference from their unreflective, nonverbal behaviour. Indeed, it is hard to see what sort of nonverbal behaviour could warrant the ascription of some highly detailed or bizarre delusions, such as the delusion that one is dead (as opposed to, say, the belief that life isn't worth living). (Frankish 2012, 25)

According to Howhy and Rajan (2012), delusions are not doxastic states, but perceptual states. They support this idea by pointing out some interesting commonalities between delusions and perceptual illusions. For instance, many perceptual illusions, such as Müller-Lyer illusion, are not revisable. Similarly, delusions are not revised in the face of overwhelming counterevidence. Again, illusions such as rubber hand illusion do not have impact on our beliefs about the nature of animate and inanimate objects. Similarly, many delusions do not have impact on other things subjects believe, and so on.

> There may thus be a more than passing analogy between delusions and illusions. This would remove delusions from the domain of beliefs and alight it closer with perceptions. […] This alignment of delusion with illusions means treating delusions as a kind of perceptual state. (Hohwy & Rajan 2012, 8)

(There are more anti-doxastic proposals. But I stop here. This list is not intended to be an exhaustive one.)

Now, I am not going to examine these anti-doxastic proposals one by one. Rather, I will point out some general worries, which come from what I have already mentioned as the *prima facie* reasons for DD.

The first *prima facie* reason for DD is that delusional subjects sincerely assent to their delusions. Now, if delusions are not beliefs but something else, then why do delusional subjects sincerely assent to the their delusions? For instance, if a delusional subject merely

imagines, not believes, that his wife was replaced by an imposter, why does he sincerely assent to the sentence "My wife was replaced by an imposter"? When I believe that there is a bottle of beer in the fridge, I will sincerely assent to the sentence "A bottle of beer is in the fridge." On the other hand, when I merely imagine that there is a bottle of beer in the fridge without believing it, I will not sincerely assent to the sentence. I might pretend to assent to it in a pretense play, but it is not a sincere assenting.

The second *prima facie* reason for DD is that DD seems to be a part of the reason why delusions are pathological. Now, if delusions are not beliefs but something else, why are they pathological? As I already mentioned, the pathological nature of delusion is somehow mysterious if delusions are just imaginations. It looks as though there is nothing wrong in imagining, for instance, that wife is replaced by an imposter. Again, the pathological nature of delusion might be mysterious if delusions are bimaginations or in-between beliefs. On bimagination proposal, Bayne and Fernández (2008) writes,

> […] theorizing about delusion (and, to a lesser extent, self-deception) typically begins with the thought that these states are pathological beliefs – they violate certain norms of belief formation. It is unclear how Egan's account might accommodate this thought, for nothing can be pathological belief unless it is also a belief. (Bayne & Fernández 2008, 17)

Here, Bayne and Fernández are making the same point. If delusions are not beliefs but

bimaginations, then what explains the pathological nature of delusions? What is wrong about LA-O's bimagination that left hand doesn't belong to her? The same challenge goes also to in-between belief proposal. If delusions are not beliefs but in-between beliefs, then what explains the pathological nature of delusions? What is wrong with Lawrence's in-between belief that he is threatened by evil forces?

It wouldn't be a good idea to respond to this by claiming that bimagination is an intrinsically pathological such that anyone who is in the state of bimagination is in a pathological condition. At least, this response doesn't seem to be available to Egan. He argues that self-deception involves "besire", which is the intermediate state between belief and desire. Clearly, it can't be the case that besire is an intrinsically pathological mental state. After all, it is not the case that self-deception is generally pathological. But, then how is it that bimagination, the intermediate state between belief and imagination, is intrinsically pathological, while besire, the intermediate state between belief and desire, is not? What explains this difference? The same thing can be said about Schwitzgebel's in-between belief proposal. It can't be the case that in-between belief is intrinsically pathological state, because many of the examples of in-between beliefs he gives are not pathological at all (Schwitzgebel 2001, 2002). For instance, he thinks that a woman who sincerely claims that all races are intellectually equal but shows some implicit racist behavior in-between believes that all races are intellectually equal. This example is clearly a

non-pathological one.

The third *prima facie* reason for DD is that DD helps us to distinguish delusion from other conditions. Now, if delusions are not beliefs but something else, then how can we distinguish delusion from other conditions? Hohwy and Rajan argue that delusions are perceptual states. But, then, how can we distinguish delusion from perceptual illusion or hallucination? In the case of Müller-Lyer illusion, we have the experience where it seems as though one of two lines is longer than the other, but we do not usually believe that one is actually longer than the other. Again, in the case of auditory verbal hallucination, patients have the experience as if they hear some spoken words. But it is not the case that all subjects with auditory verbal hallucination come to have the delusions associated with their hallucinatory experience.

The forth *prima facie* reason for DD comes from the practice among psychiatrists to describe delusions as beliefs. Now, if delusions are not beliefs but something else, how can it be the case that thousand of psychiatrists describe delusions as beliefs for such a long time?

One possible reply is something like this. Maybe, psychiatrists are using the term "belief" in a special sense, and, in that sense, it is true that delusions are beliefs. But, in philosophical sense of "belief", it is not the case that delusions are beliefs and, thus, DD is false. (In this reply, it is assumed that DD is supposed to be about "belief" in philosophical sense, not "belief" in psychiatric sense.)

But, as far as I can tell, there is no such difference between "belief" in philosophical sense and "belief" in psychiatric sense. The core meaning of "belief" in philosophical sense is, I think, "taking something to be true", and this seems to be shared by "belief" in psychiatric sense. When a psychiatrist say, for instance, that Lawrence "believes" that he is personally threatened by evil forces, it basically means that he takes it to be true that he is personally threatened by evil forces. Philosophical "belief" is distinguished from religious "belief", and the same thing seems to be true about "belief" in psychiatric sense. When psychiatrists say, for instance, that LA-O "believes" that the left hand doesn't belong to her, this "belief" is distinguished from religious "belief". Of course, some delusions have religious contents, and these delusions are also called "beliefs" by psychiatrists. It shows that there are some overlaps between psychiatric "belief" and religious "belief". It doesn't show, however, that psychiatric "belief" is significantly different from philosophical "belief". For, the same thing is true about philosophical "belief". Whenever religious believers take their religious "beliefs" to be true, they are also "beliefs" in philosophical sense. Thus, there are some overlaps between philosophical "belief" and religious "belief" as well.

The fifth *prima facie* reason for DD comes from the fact that delusional subjects regard their delusions as beliefs. Now, if delusions are not beliefs but something else, how can it be the case that delusional subjects regard their delusions as beliefs?

Currie and collaborators assume that a delusional subject's metacognitive capacity with

regard to propositional attitude is impaired, and especially unreliable. According to them, delusional subject misidentify their imaginations (=delusions) as beliefs. Certainly, there are many studies suggesting the existence of metacognitive impairment in delusional subjects. Importantly, some of the studies do support their idea that delusions involve the loss of the capacity to identify imagining from something else (Anselmetti et al. 2007; Brébion et al. 2000; Jenkinson et al. 2009). However, we shouldn't take those studies as the empirical support for the assumption by Currie and collaborators. What is found in those studies is, strictly speaking, different from what Currie and collaborators expected. Basically, these studies are about perceptual level misidentification, while Currie and collaborators expected propositional attitude level misidentification. More precisely, these studies suggest the tendency of delusional subjects to make perception/mental imagery misidentification, while Currie and collaborators expected the tendency to make belief/propositional imagination misidentification. Thus, those studies do not support their claim that the self-knowledge of delusional subjects with regard to propositional attitudes is especially unreliable (although they do support the idea that their self-knowledge of experiential states are especially unreliable).

Before moving on, I want to point out another worry, which is about the moral and legal responsibility of delusional offenders. The current practice of attributing responsibility to subjects seems to rely on the assumption that people in general act on the basis of what

they believe and what they desire. For example, one of the core assumptions in attributing responsibility to a main, *A*, for his killing another man, *B*, might be that *A* kills *B* because he believes that *B* is having an affair with his wife, and he wants to punish *B*. This is also true when we discuss the responsibility of delusional offenders. It might turn out that *A*'s thought about the affair is in fact delusional (i.e. delusion of jealousy). Still, in examining *A*'s responsibility for the act, we assume that *A* kills *B* because *A* (delusionally) believes that *B* is having an affair with his wife and he wants to punish *B*. If anti-DD incompatibilists are correct and thus DD is false, then this assumption behind the practice of attributing responsibility collapses. Some anti-DD theorists argue that delusions are not beliefs but some kinds of in-between states. But, if this is true, then what should we say about responsibility? If *A*'s act is not based upon the belief that *B* is having an affair with his wife but, instead, on the bimagination that *B* is having an affair with his wife, then how should we assess *A*'s responsibility? As Bortolotti points out, the anti-DD proposals seem to make it difficult to determine whether or not *A*'s act is intentional (assuming that intentional action is based upon belief and desire), which causes serious complications in the current practice of attributing responsibility (Bortolotti 2010, 21). It looks as though in-between state accounts do not just invite theoretical complications in the classification of mental states by introducing various kinds of in-between states, but also the practical complications in the attribution of responsibility by making it unclear whether or not the actions at issue are

intentional. Are these complications really worthwhile?

2.3.2 Pro-DD Incompatibilism

Pro-DD incompatibilists resolve the tension between DD and CDT by rejecting CDT. In doing so, pro-DD incompatibilists need to give reasons to be skeptical about CDT.

Pro-DD incompatibilists need to show that the causal difference between delusions and normal beliefs is not significant. Here, I will discuss Bortolotti's (2010, 2012) attempt, which is the best example of pro-DD incompatibilist strategy.

Bortolotti's argumentative strategy is this. For each delusional feature, she gives some concrete examples, from empirical studies and everyday observations, where the feature is also found, in some degree, in the functional roles of what we uncontroversially call "belief". In other words, Bortolotti argues that any of the delusional features are, in some degree, also had by some uncontroversial beliefs too. She says,

> The delusion that I am dead is very different from the belief that the supermarket will be closed on Sunday, but this does not show that there is a categorical difference between delusions and beliefs. Here is a challenge. For each delusion, I'll give you a belief that matches the type if not the degree of irrationality of the delusion. (Bortolotti 2010, 259)

> The psychological literature invites us, instead, to consider that beliefs are often badly integrated with other beliefs, unsupported by evidence, resistant to change, and behaviourally inefficacious. Once we accept that everyday beliefs can be irrational in

> these ways, it is a short step to maintain that there is a continuity between everyday beliefs and clinical delusions. (Bortolotti 2012, 39)

She gives a lot of interesting examples to support her claim. Here are some of them. Many uncontroversial beliefs, such as racist beliefs and religious beliefs, are insensitive to counterevidence. It is often very difficult to remove, by giving counterevidence, someone's belief that black people are lazy and intellectually incompetent, that the miscarriage was the punishment of sin, and so on. Some uncontroversial beliefs are not coherent with other beliefs. For example, many people maintain superstitious beliefs, beliefs about magic or beliefs about supernatural phenomena, even though they believe in scientific worldview and, furthermore, they recognize that their superstitious beliefs etc. are highly unlikely under the scientific worldview. Again, some uncontroversial beliefs do not have impact on non-verbal behavior. For instance, people might believe that using condoms is important because unprotected sex is dangerous, but they do not use condoms actually.

Bortolotti's argument, however, is not very persuasive, and I will explain the reasons in the following. (For the sake of fairness, I need to mention the fact that Bortolotti is not actually discussing CDT, but something slightly different from it. The aim of her argument is to refute, not the claim that delusions fail to play belief-like causal roles, but the claim that delusions are significantly more irrational (with respect to the sensitivity to evidence and the coherence with other beliefs and non-verbal behavior) than uncontroversial beliefs.)

In evaluating her argument, we can ask two kinds of questions.

(1) How appropriate are these examples? Do they really show that uncontroversial beliefs have some delusional features? (e.g. Someone might think that religious beliefs are sensitive to some kinds of evidence, such as testimonial evidence.) Are they really uncontroversial examples of beliefs? (e.g. Some one might deny that the subjects in the condom example really believe that using condoms is important.)

(2) Assuming that they are appropriate examples, do they undermine CDT? In other words, assuming that delusional features are shared by some uncontroversial beliefs in some degree, does this undermines that idea that problematic delusions fail to play belief-like causal roles?

In the following, I will put my focus on the question (2), and I will leave the question (1) open. I will not go into the question (1), firstly, because she gives a lot of examples and I just don't have enough space to examine those examples one by one and, secondly, the most serious difficulties of Bortolotti's argument would in fact be related to question (2).

In my view, the answer to (2) is likely to be "NO". In other words, even if Bortolotti's examples are really appropriate, they do not undermine CDT.

(1) *Degree Matters*: Bortolltti main point is, simply put, that causal difference between delusions and normal beliefs is just a matter of degree. One simple worry to this is that degree might matter. In other words, the causal difference in degree between *A* and *B* might be significant enough to establish that As are not playing B-like causal roles. Indeed, in the normal folk-psychological classification of mental states, there seem to be some pairs of

different states where the one state differs causally from the other only in degree. This means that causal difference in degree is sometimes significant enough to draw the distinction between two kinds of mental states in the folk-psychological classification.

For instance, is there anything more than the causal difference in degree between beliefs and acceptances? About the distinction between belief and acceptance, Schwitzgebel writes,

> Generally speaking, acceptance is held to be more under voluntary control of the subject than belief and more directly tied to a particular practical action in a context. For example, a scientist, faced with evidence supporting a theory, evidence acknowledged not to be completely decisive, may choose to accept the theory or not to accept it. If the theory is accepted, the scientist ceases inquiring into its truth and becomes willing to ground her own research and interpretations in that theory; the contrary if the theory is not accepted. (Schwitzgebel 2010)

Schwitzgebel's phrase "acceptance is held to be *more* under voluntary control of the subject than belief and *more* directly tied to a particular practical action in a context" (emphasis added) seems to suggest that even though voluntary control and connection to practical actions in contexts are more evident in acceptances, this is a matte of degree. And, probably, he is right. Perhaps, beliefs are not perfectly outside of our voluntary control. It is uncontroversial that belief can indirectly be influenced by voluntary control. For instance, it would be possible to change our belief by voluntarily manipulating evidence, change the focus of attention, change our everyday habit and so on. Even direct voluntary control over

belief might be possible (e.g. Ginet 2001; Ryan 2003). For instance, one might be able to voluntarily choose to believe something in some cases. Again, it is not the case that acceptance is purely practical and totally insensitive to evidence or truth. The scientist in Schwitzgebel's example might accept the theory for practical purposes. But, it is important to note that she has, at least, some evidence for the theory. It would be hard for her to accept the theory if she has no evidence whatsoever for the theory, and even harder if she has strong counterevidence against it, no matter how practically useful accepting it is. This shows that, not just beliefs, but also acceptances are sensitive to evidence and truth in some degree.

Similarly, it is not clear that there is anything more than the causal difference in degree between belief/imagination (Currie & Ravenscroft 2002; Nichols & Stich 2000; Nichols 2004) or imagination/supposition (Gendler 2000; Currie & Ravenscroft 2003; Egan 2007).

(2) *Jointly Significant Difference*: Here is how Bortolotti actually argues. First, she discusses what she calls "procedural irrationality" (= incoherence with other beliefs) and argues that it doesn't make create significant (enough to deny DD) difference between delusions and uncontroversial beliefs and then, moving on, she discusses what she calls "epistemic irrationality" (= insensitivity to evidence) and argues that it doesn't make significant difference either. And, she also discusses what she calls "agential irrationality" (= having no impact on non-verbal behavior) and argues, again, that it doesn't make

significant difference either. And, she concludes that, since none of these irrationalities make significant difference, there is no significant difference between delusions and uncontroversial beliefs after all.

This line of thought, however, is problematic. This neglects the possibility that even if none of the delusional features make significant causal difference *individually*, they make significant difference *jointly*. Presumably, having a slight fever is, individually, not a very strong indication that I am ill. After all, there are some healthy situations where I can have a slight fever (e.g. immediately after exercise). Heavy sneezing, again, is not, individually, a very strong indication of illness. Hey fever can cause heavy sneezing without making me ill. Sore throat is not, individually, indicating illness very strongly either. Shouting at a Karaoke bar can cause it without making me ill. Nonetheless, slight fever, heavy sneezing and sore throat strongly indicate that I am ill jointly. If I have a slight fever, heavy sneezing and sore throat at the same time, then there should be something wrong with me. I must be ill. Similarly, the fact that problematic delusions are incoherent with other beliefs might not give a strong support for CDT individually. As Bortolotti pointed out, some uncontroversial beliefs are incoherent with other beliefs too. The fact that problematic delusions are insensitive to counterevidence might not support CDT very strongly individually either. Some uncontroversial beliefs are insensitive to counterevidence too. Again, the fact that problematic delusions do not have impact on non-verbal behavior might not give a strong

support for CDT individually either. Some uncontroversial beliefs do not have impact on non-verbal behavior either. Nonetheless, these delusional features might support CDT strongly jointly.

This worry is serious given the fact that the delusional features arise often concurrently. In other words, it is often the case that one delusional state has several delusional features at the same time. Furthermore, as Bortolotti admits herself, delusions are typically "irrational across more dimensions than non-delusional beliefs." (Bortolotti 2012, 39) In our terminology, delusions typically have more delusional features than uncontroversial beliefs.

(3) *Same Argument from the Opposite Direction*: Suppose that Bortolotti is right, and it is true that there is nothing more than the causal difference in degree between problematic delusions and uncontroversial beliefs. But, it might still turn out that there is nothing more than the causal difference in degree between problematic delusions and other kinds of states, such as imaginations. Given comparative similarity requirement, in this case perhaps we are justified in saying that problematic delusions play belief-like causal roles only if the degree to which problematic delusions differ from beliefs is smaller than the degree to which they differ from imaginations. But, the main point of the clinical observations that I mentioned is that it is not clear that this condition is satisfied, as I suggested with the example of Hannah's belief and Martin's imagination. In other words, those clinical observations make it unclear that the degree to which problematic delusions differ from

beliefs is smaller than the degree to which they differ from imaginations.

Perhaps, there is nothing more than the causal difference in degree between beliefs and imaginations in the first place, in which case there is no surprising that there is nothing more than the causal difference in degree between problematic delusions and uncontroversial beliefs. Currie and Jones (2006) seem to accept this view. They argue that beliefs and imaginations are on a continuum and delusions are located somewhere between them. Perhaps they will agree with Bortolotti's claim that the causal difference between delusions and beliefs is just the one of degree. But, they will also claim that the same thing is true about the causal difference between delusions and imaginations.

Perhaps the idea that beliefs and imaginings are on a continuum is correct. Bayne and Pacherie (2005) carefully distinguish different kinds of (propositional) imaginations. Simple imagination is the state we are in when we entertain a certain proposition without having any attitude toward it (e.g. entertaining the thought that the population of Nepal has doubled in the last 25 years without having any attitude toward it). Counterfactual imagination is the state we are in when we entertain a scenario in which a proposition holds while believing of this scenario that it is not actual (e.g. enjoying fictional stories). Indicative imagination is the state we are in when we imagine that P with the inclination to think that P is true (e.g. imagining that JFK was killed by Mafia). Presumably, we should regard imaginations as on the continuum with beliefs where simple imaginations are at the end of

the continuum, indicative imaginations are much closer to beliefs and counterfactual imaginations are somewhere in the middle.

(4) *What Makes Delusion Pathological?*: Bortolotti's argument leaves a question. Bortolotti emphasizes the similarity between problematic delusions and uncontroversial irrational beliefs. But, then, why is it that delusions are pathological while uncontroversial irrational beliefs are not? She tentatively suggests that delusions are pathological because they have significant negative impact on well-being.

> One possibility is that delusions are pathological in that they negatively affect the well-being and the health of the subjects who report them (as many have already argued). Irrational beliefs that are not delusions seem less distressing, and don't seem to exhaust the cognitive resourses of the subjects in the same way delusions do. (Bortolotti 2010, 260)

Delusions certainly negatively affect the well-being of delusional subjects in all sorts of ways. They cause psychological distress, prevent subjects from being fully involved in social life and personal relationship, prevent them from fulfilling their abilities and talents, increase the risk of suicidal acts, and so on. Presumably, there is an important link between negative impact on well-being and pathology. For instance, it is plausible to think that negative impact on well-being is necessary for pathology. As Wakefield pointed out, "disorder is in certain respects a practical concept that is supposed to pick out only conditions that are

undesirable and grounds for social concern" (Wakefield 1992b, 237). But, the problem with this answer is that negative impact on well-being would not be sufficient for pathology, even if it is necessary. For instance, drinking too much alcohol, smoking, pain, psychological stress, SNS addiction, grief, etc. have negative impact on well-being, but they are not pathological. Thus, just saying that delusions have negative impact on well-being is not enough for a full explanation of the pathological nature of delusions (see Chapter 4 for more on the pathological nature of delusion).

Chapter 3: Teleo-Attitude Functionalism

In the last chapter, I motivated compatibilist by showing that (1) there is no good argument against multiple-functionalibility of mental states, and (2) there might be some other examples of mental states that are multiply-functionable, and (3) incompatibilist options (Anti-DD and Pro-DD) have some problems. Given the fact that mainstream, functionalist theories are incompatibilist theories, the main task for compatibilist is to present an alternative compatibilist theory of belief. Presenting such a theory is the main aim of this chapter. A compatibilst theory of belief is, in short, the one that allows for the possibility of beliefs without belief-like causal roles. In section 3.1, I will examine and reject some potential compatibilist theories. In section 3.2, I will present what I take to be the best compatibilist theory. The theory is called "teleo-attitude functionalism".

3.1 Compatibilist Theories of Belief

3.1.1 Identity Theory and Phenomenalism

An example of compatibilist theory is the identity theory of belief that identifies belief with a certain kind of physical state on type-type basis. The theory clearly allows for the

possibility of beliefs without belief-like causal roles. Even if a mental state fails to play a belief-like causal role, it is perfectly possible that the state is identical with a right kind of physical state. The problem of this view is well known; it fails to allow for multiple-realizability of belief. In other words, the view certainly allows for the possibility of beliefs with multiple kinds of causal roles, but it fails to allow for the possibility of beliefs with multiple kinds of physical basis.

Another example of compatibilist theory is phenomenalism, according to which to believe something is to be in a certain type of states with a certain kind of phenomenology. Hume is the best known defender of this view. According to Hume, belief is defined in terms of its phenomenal character that he calls "force", "viviacity" or "liveliness" (although he is not fully satisfied with these terms in describing the phenomenology of belief).

> It follows, therefore, that the difference between *fiction* and *belief* lies in some sentiment or feeling, which is annexed to the latter, not to the former, and which depends not on the will, nor can be commanded at pleasure. […] I say then, that belief is nothing but a more vivid, lively, forcible, firm steady conception of an object, than what the imagination alone is ever able to attain (Hume 1748/2007, 35)

The theory also allows for the possibility of beliefs without belief-like causal roles (assuming that phenomenology of belief can be dissociated from the causal role of it). Even if a mental state fails to play a belief-like causal role, it doesn't rule out that possibility that the state

has force, vivacity and liveliness. Is this a plausible theory of belief? The main objection to Hume's view is that this theory is not applicable to non-occurrent, dispositional beliefs (Price 1970). Even if it is granted that occurrent beliefs have distinctive phenomenal character, still dispositional beliefs do not have any phenomenal character. One way of avoiding this objection is to say that believing something dispositionally is to be disposed to be in the occurrent state with force, vivacity and liveliness. However, this move blurs the important distinction between dispositionally believe something and being disposed to believe something. Dispositionally believing something implies believing it. On the other hand, being disposed to believe something doesn't. For example, it is possible that new students at a college are strongly disposed to believe that the college is better than other colleges in the country, but they do not believe it yet. Another way of avoiding the objection is to say that phenomenalism is the theory of occurrent belief, not belief in general. Presumably, this is a plausible interpretation of Hume from a historical point of view (Marušić 2010). But, this makes phenomenalism irrelevant to our discussion. Delusion is our main focus. Delusions can be occurrent, but they can also be dispositional. It is not the case, for instance, that the subjects with persecutory delusion are always occurrently having persecutory thoughts. Some have persecutory delusions for many years. In those years, persecutory thoughts are sometimes occurrent and sometimes dispositional. So, if phenomenalism is just a theory of occurrent belief, it is not relevant to the discussion of

delusions. When we want to know if delusions, including dispositional delusions, are beliefs or not, phenomenalism, thus understood, doesn't help us.

3.1.2 Statistical Functionalism

Lewis's (1980) discussion of pain is relevant here. He believes that pain without pain-like causal role is conceivable and possible. In particular, he believes that mad pain case that I already mentioned is possible. He thinks that, given the possibility of pain without pain-like causal role, any successful theory of pain needs to allow for the possibility. On this ground, he rejected the standard functionalism about pain according to which pains are, simply, the states that play pain-like causal roles. Here is his alternative theory;

Statistical Functionalism: Pain
x feels pain iff *x* is in a physical state of the type whose tokens statistically-normally play pain-like causal roles in the appropriate population to which *x* belongs.

Statistical functionalism does allow for the possibility of pains without pain-like causal roles. According to this theory, the mad pain guy feels pain, even though his pain doesn't play a pain-like causal roles, as long as he is in a physical state of the type whose tokens statistically-normally play pain-like causal roles in the appropriate population to which he belongs. For instance, he might be in the physical state of the type "C-fiber-firing" whose

tokens statistically-normally play pain-like causal roles in the appropriate population to which he belongs.

Now, the theory can easily be generalized to other kinds of mental states, including belief;

Statistical Functionalism: Belief

x believes something iff *x* is in a physical state of the type whose tokens statistically-normally play belief-like causal roles in the appropriate population to which *x* belongs.

According to statistical functionalism about belief, there can be beliefs without belief-like causal roles for the same reason that there can be pains without pain-like causal roles according to statistical functionalism about pain. According to this theory, I believe something, even when my belief doesn't play a belief-like causal role, as long as I am in a physical state of the type whose tokens statistically-normally play belief-like causal roles in the appropriate population to which I belong. This means that statistical functionalism is a compatibilist theory of belief.

Unfortunately, however, statistical functionalism is not a *plausible* compatibilist theory of belief. Here are some problems.

First, statistical functionalism doesn't get off the ground without some criteria about what "appropriate population" is. The right hand side of statistical functionalism about pain

is actually false about the mad pain guy if we take "appropriate population" to be the population comprising of himself and fellow mad pain guys. In that population, it is not the case that the mad pain guy is in a physical state of the type whose tokens statistically-normally play pain-like causal roles. Instead, he is in a physical state of the type whose tokens statistically-normally play mad-pain-like causal roles in that population. Thus, Lewis needs to provide some criteria of determining "appropriate population" according to which the "appropriate population" for the mad pain guy is not the population of mad pain guys but rather, for instance, the population of mankind. But, as Lewis admits, providing such criteria is not an easy job.

Second, statistical functionalism leads to absurd consequences. Think about the following case. The mad pain guy belongs to the population of mankind. In the population of mankind, most of the C-fiber-firing-tokens play pain-like causal roles. The mad pain guy has been in the state of C-fiber-firing for a long time, although his C-fiber-firing-token doesn't play a pain-like causal role. In this case, mad pain guy genuinely feels pain according to statistical functionalism about pain, because the right-hand side of it is true about him. One day, however, a massive nuclear war occurs and, for some reasons, the mad pain guy is the only survivor on earth. Now, it is no longer the case that most of the C-fiber-firing-tokens play pain-like causal roles in the population of mankind. Mad pain guy is the only member of the population and his C-fiber-firing token doesn't play a pain-like causal role.

Consequently, right-hand side of statistical functionalism about pain is false about him and, hence, he is not in pain anymore! But, of course, this is absurd. Certainly, nuclear weapons are capable of killing many people, but they are not capable of removing pain. They are not painkillers.

Think, again, about the following parallel case. In human population, let us suppose, most of the "B"-tokens play belief-like causal roles. Problematic delusional subjects are in the state of B when they are deluded, although their B-tokens don't play belief-like causal roles. In this case, problematic delusional subjects genuinely believe their delusions according to statistical functionalism about belief because the right-hand side of it is true about them. One day, however, a massive nuclear war occurs and, for some reasons, problematic delusional subjects are the only survivors on earth. Now, it is no longer the case that most of the B-tokens play belief-like causal roles in the population of mankind. Problematic delusional subjects are the only members of the population and their B-tokens don't play belief-like causal roles. Consequently, the right-hand side of statistical functionalism about belief is false about them and, hence, they do not believe their delusions anymore! But, of course, this is absurd. Nuclear weapons are not capable of removing delusional beliefs.

3.1.3 Normativism

Bayne (2010), in his discussion on delusion, proposed another theory of belief, which looks like a compatibilist one.

Normativism
A mental state is a belief iff the state is subject to the norms of belief.

Norms of belief include, among others, the norm of truth (corresponding to the world) and consistency (being consistent with other beliefs). It is important to distinguish between a mental state's being *subject to* the norms of belief and its *conforming* these norms. Generally speaking, *x*'s being subject to norm *N* is distinguished from *x*'s conforming *N* in such a way that it is possible that *x* is subject to *N* without actually conforming *N*. For instance, we can fail to conform ethical norms to which we are subject. Normative functionalism regards being subject to the norms of belief, not conforming the norms, to be crucial for a mental state to be a belief. When Bayne said, "If delusions are beliefs, then we can quite properly hold those who are delusional to account for failing to live up the norms of belief. For example, we can criticize the delusional individual for holding to be true something that is inconsistent with other things that he or she believes," (Bayne 2010, 334) he is clearly committed to the view that, for something to be a belief, it doesn't have to conform (or live up to) the norms of belief.

Normativism seems to be a compatibilist theory of belief. The theory seems to allow for, for instance, the possibility of beliefs that are pretty inconsistent with other beliefs. These states are not conforming the norm of consistency, but they could nonetheless be subject to the norm. And, according to normativism, being subject to the norms of belief is sufficient for them to be beliefs. Conforming them is not necessary.

But, again, normativism would not be a plausible compatibilist theory. Here is a worry. How can we know whether a given mental state is a belief or not? According to normativism, we need to see whether the state is subject to the norms of belief or not. Then, the question is, how can we know that the state is subject to the norms of belief? It is not very difficult to see that a given mental state conforms the norms of belief. What we need to do is to find out that the state is true, consistent with other beliefs and so on. But, how can we know that a state is subject to the norms of truth, consistency, and so on, especially when the state fails to conform these norms? For instance, are problematic delusions subject to these norms? One might think that they are subject to these norms because, for instance, problematic delusional subjects retain the minimal capacity to follow the norms (e.g. capacity to produce true beliefs, maintain a consistent set of beliefs). But, this answer is unsatisfactory, even if it is really true that, in an interesting sense of "minimal", problematic delusional subjects retain the minimal capacity to follow the norms of belief. Just showing that a person has the minimal capacity to follow the norms of belief is not sufficient to show that his delusions,

the particular states he is in, are subject to the norms of belief. It is not the case, for instance, when he has the minimal capacity to follow the norms of belief, that all of his mental states are subject to the norms of beliefs. His desires, his intentions, his imaginations, his perceptual states, all of these are not subject to the norms of belief. Thus, to show that his delusion is subject to the norms of belief, we need to show something more than that he has minimal capacity to follow the norms of belief. Presumably, we need to show that his delusions, unlike his desires or intentions, have some intrinsic features in virtue of which they are subject to the norms of belief. But, what are the intrinsic features? How can we find that?

Bayne himself admits that it is very difficult to determine, according to normativism, whether a given mental state is a belief or not.

> Assuming this normative approach, do delusions qualify as beliefs or not? I doubt that delusions have the kind of unitary nature that would be needed in order for this question to have a determinate answer. Some delusions might be best understood as commitment involving, in which case we can and should evaluate them with respect to the norms of belief. (And when so evaluated the patient will invariably come up short, for delusions are 'by definition' pathologies of belief.) Other delusions might be best thought of as a kind of imaginative charade, and not legitimately evaluated with respect to the norms of truth and rationality. The case for regarding a delusion as a doxastic state may differ from patient to patient and may even fluctuate for particular patients from one occasion to another. Not only might it be difficult to tell whether a delusion involves the requisite kind of commitment on the part of its subject to qualify as a belief, in some cases there may simply be no fact of the matter about this. In short, the normative view of things does not make it easier to answer the question of whether delusions are beliefs; indeed, it might even make it harder to answer such

questions than the functional role approach does. (Bayne 2010, 334-335)

This indeterminacy might be a procedural one. In other words, the indeterminacy might come from the lack of procedure to determine whether a given mental state is a belief or not (it is "difficult to tell whether a delusion involves the requisite kind of commitment on the part of its subject to qualify as a belief"). Alternatively, as Bayne suggested when he says that "in some cases there may simply be no fact of the matter about this", the indeterminacy might be a metaphysical one. In other words, the indeterminacy comes from the lack of the fact of matter as to whether a given mental state is a belief or not. In any case, this indeterminacy causes a trouble for normativism. Presumably, there is something wrong with the theory of belief according to which the question about whether a given mental state is a belief or not is indeterminate in these ways.

3.2 Teleo-Attitude Functionalism

3.2.1 Teleo-Attitude Functionalism

Now, I turn to what I take to be the best compatibilist theory of belief. The key notion is *teleology* or, more precisely, the *teleological notion of function*. Here is the basic idea. Filtering metabolic wastes from blood is an essential, defining feature of kidney. Still, there

can certainly be kidneys without filtering metabolic wastes from blood. For instance, it is possible that I get serious renal failure and my kidney stops filtering metabolic wastes from blood. The teleological theory of kidney explains why there can be kidneys without filtering metabolic wastes from blood despite of the fact that filtering metabolic wastes from blood is the essential feature of kidney.

Teleological Theory of Kidney

Something is a kidney iff it has the function of filtering metabolic wastes from blood.

(In philosophy of biology, this idea is often expressed by the claim that kidney-type is functionally defined.) There can be kidneys without filtering metabolic wastes from blood because, according to teleological theory, first, it is possible that an organ has the function of filtering metabolic wastes from blood but fails to do it actually and, second, having the function of filtering metabolic wastes from blood is sufficient for the organ to be a kidney. These kidneys are often called "malfunctioning kidneys". Now, moving on to belief, the basic idea here is that, for the same reason that the teleological theory of kidney allows for the possibility of kidneys without filtering metabolic wastes from blood, the teleological theory of belief would allow for the possibility of beliefs without belief-like causal roles.

Now, I tentatively propose the following theory of belief ("tentative" because this will be revised soon);

Teleo-Attitude Functionalism (TAF): Belief

A mental state is a belief iff it has the function of playing a belief-like causal role.

There can be beliefs without belief-like causal roles according to TAF because, first, it is possible that a mental state has the function of playing a belief-like causal role but fails to play it actually and, second, having the function of playing a belief-like causal role is sufficient for the state to be a belief. These beliefs are, so to speak, "malfunctioning beliefs".

Just like other philosophers advocating teleological theories of mental states, I accept etiological function or etiological analysis of function (Millikan 1984, 1989a; Neander 1991a, 1991b). Hereafter, when I use the term "function" or "teleological function" without any specifications, I will be talking about etiological function. Etiological function of something is, details aside, the effects for which it is selected. In other words, to say that something has the etiological function of doing *F* is to say that it has a right kind of history where the ancestors of it were selected for doing *F*. A kidney has the function of filtering metabolic wastes from blood because kidneys have been selected for filtering metabolic wastes from blood. I accept etiological function as opposed to it rivals such as systemic function (Cummins 1975) because of its theoretical feature. Malfunctionability is a crucial feature of TAF. Etiological function clearly allows for malfunctionability. In other words, it is possible that something has the etiological function of doing *F* without actually doing *F*. This is

possible because having the etiological function of doing *F* is having a right kind of history, and something can have a right kind of history without actually doing *F*. On the other hand, it is not clear that other kinds of functions allow for malfunctionability. (I will come back to this issue in the next chapter.)

It is notable that TAF is free from the problems for earlier compatibilist theories. Statistical functionalism leads to the absurd consequence that delusional beliefs can be removed by radical statistical changes caused by, for instance, a nuclear war. TAF doesn't lead to such a consequence, because statistical-normality has nothing to do with etiological function. Etiological functions are not statistically-normal performances. Borrowing Millikan's famous example, sperms have the function of fertilizing an egg despite of the fact that, statistically normally, sperms do not fertilize an egg. They have the function of fertilizing an egg in virtue of having a right kind of history. Normativism invites the problem of procedural and metaphysical indeterminacy with regard to whether a given mental state is a belief or not. TAF doesn't. There is no metaphysical indeterminacy there because there is no metaphysical indeterminacy with regard to whether the state has a right kind of history or not. Procedurally, whether a mental state has a right kind of history or not can be studied by looking at its current performance (experimental psychology, neuroscience, psychiatry) and examining evolutionary hypotheses about the state (evolutionary psychology, evolutionary psychiatry).

Here are some clarificatory remarks on TAF.

(1) *Teleological Functionalism and Standard Functionalism*: TAF is a version of teleological functionalism. Godfrey-Smith nicely summarizes the difference between standard functionalism and teleological functionalism.

> Most recent philosophy of mind has been "functionalist" in some sense or other. We can distinguish two basic forms of functionalism in philosophy of mind. First, there is the more orthodox view which I will call "dry functionalism." This view understands function in terms of causal role, and it identifies mental states in terms of their typical causal relation to sensory inputs, other mental states, and behavioral outputs. Second there is "teleo-functionalism." The view makes use of a richer, biological concept of function more closely allied to traditional teleological notions, a concept often analyzed with the aid of evolutionary history. For the dry functionalists, one essential property of any mental state is the pattern of behavioral outputs which the state, in conjunction with the rest of the system, tends to cause in various circumstances. For the teleo-functionalist, what is essential to the mental state is no what it *tends* to do but what I is *supposed* to do. (Godfrey-Smith 1996, 13)

A number of philosophers (Dennett 1991; Lycan 1982, 1995; Sober 1985; Sterelny 1990) developed teleological functionalist theories for various reasons; (1) teleological functionalism can deal with some counterexamples to standard functionalism (e.g. nation of China case), (2) teleological functionalism gives much better explanation of consciousness than standard functionalism, (3) teleological functionalism is more consistent, than standard functionalism, with function-analytic explanatory strategy which is widely used in psychology, and so on.

(2) *Teleosemantics and TAF*: Teleosematicists believe that teleological function is the key in explaining, naturalistically, the representational content of mental states (Dretske 1986, 1991; Millikan 1984, 1989b; Neander 1995; Papineau 1984). More precisely, teleological function is the key in solving a serious problem for naturalistic theories of the content of mental states, namely, the problem of explaining how misrepresentation is possible. Millikan nicely summarizes the main idea of this. "False representations are representations, yet they fail to represent. How can that be? In can be in the same way that something can be a can opener but be too dull and hence fail to open cans, or something can be a coffee maker yet fail to make coffee because the right ingredients were not put in or it was not turned on. They are 'representations' in the sense that the biological function of the cognitive systems that made them was to make them represent things. Falsefood is thus explained by the fact that that purpose often go unfulfilled" (Millikan 2004, 64-65).

One might think, however, that teleosemantics has nothing to do with TAF. Teleosemantics is a view about the content of mental states, while TAF is a view about the attitude of mental states. Teleosemantics tells us something about, for instance, in virtue of what the belief that there is a bottle of beer in the fridge has the content "there is a bottle of beer in the fridge" as opposed to the content "Obama is the president of the US", while TAF tells us something about, for instance, in virtue of what the belief is a belief as opposed to a desire.

This is based upon a misleading, if not incorrect, understanding of teleosemantics. In teleosemantics, content and attitude are, in fact, pretty closely related to each other and, hence, it would be misleading to say that teleosemantics has nothing to do with attitude. This is especially evident in Millikan's version of teleosemantics. Millikan strongly opposes Fregean view of thought, according to which content (or sense) can be detached from attitude.

> On the theory proposed, intentional representations always come with propositional attitudes attached. It is essential to them that they have some kind of function, that they are designed for a particular kind of use. Frege's notion of sense, which implied that you can first represent a proposition and then add an intentional attitude to it, has done a lot of damage, I believe. There are not and could not be intentional representations that lacked attitude. There are no intentional representations without purposes, and having a purpose guarantees attitude. (Millikan 2004, 81)

> […] content of the representation turns out to be an abstraction from a fuller affair intrinsically involving an imbedding mood or propositional attitude. Put simply, there is no such thing as content without mood or attitude; content is an aspect of attitude. (Millikan 1995, 155)

"There is no such thing as content without mood or attitude" because, in Millikan's theory, the determination of content of a state is relative to the attitude of it. According to Millikan, there are three basic categories of attitude; descriptive attitude, directive attitude and pushmi-pullyu attitude. Descriptive states are the ones that, intuitively, describe the states

of affairs (e.g. beliefs), while directive states are the ones that, intuitively, direct the action of organisms (e.g. desires). Pushmi-pullyu states are the ones that are descriptive and directive at the same time (e.g. perception of affordance). The way in which the content of descriptive states is determined is quite different from the way from the way in which the content of directive states is determined. The content of a descriptive state is determined by what the state needs to correspond to if the consumer of the state (the mechanism that use the state for its purposes) is to perform its function in its normal way. On the other hand, the content of a directive state is determined by what its consumer is supposed to produce in response to the state. Pushmi-pullyu states have both descriptive content and directive content, and these contents are determined in the ways in which the contents of descriptive states and directive states are determined.

Thus, Millikan's teleosemantics is incomplete without a view about what makes descriptive states descriptive, directive states directive and pushmi-pullyu state pushmi-pullyu. I do not go into the details of her view on attitude here (see Millikan 2004, chapter 6). Crucial point is that the view about attitude tends to be a crucial of teleosemantics theories. And, the view about attitude tends to be deeply teleological, as we can expect when Millikan says that "having a purpose guarantees an attitude".

(3) *Teleological Approaches to Emotion*: Teleological approaches to mental states are quite popular in emotion literature. TAF can be understood as an application of the idea

behind these approaches. The idea that emotions serve some functions or purposes is not very new. For instance, the part III of Descartes' *The Passions of the Soul* is dedicated to, among others, the investigations of the functions of emotions. Descartes was strongly committed to the idea that all emotions have purposes. Because of this commitment, he was puzzled by timidity and fear because it is not very easy to find their purposes.

> Although I cannot believe that nature has given to mankind any passion which is always vicious and has no good or praiseworthy function, I still find it very difficult to guess what purpose these two passions might serve. It seems to me that timidity has some use only when it frees us from making efforts which plausible reasons might move us to make if this passion had not been aroused by other, more certain reasons, which made us judge the efforts to be useless. Besides free the soul from such efforts, it is also useful for the body in that it slows the movement of the spirits and thereby prevents us from wasting our energy. But usually it is very harmful, because it diverts the will from useful actions. And because it results simply from our having insufficient hope or desire, we need only increase these two passions within us in order to correct it. In the case of fear or terror, I do not see that it can ever be praiseworthy or useful. It, too, is not a specific passion, but merely an excess of timidity, wonder and anxiety – an excess which is always bad, just as boldness in an excess of courage which is always good (provided the end proposed is good). (Descartes 1649/1985, 392)

The argument here is very interesting. Descartes admits timidity as a kind of emotion, on the basis of his observation that it helps people to avoid unnecessary efforts. On the other hand, he denies that fear as an independent emotion for the reason that it doesn't serve any purposes. He suggested that fear is not an independent emotion but rather an excess of other emotions such as timidity, wonder or anxiety. Those emotions serve some purposes in

normal, non-excessive cases. But, they do not serve any purposes when they are excessive. Here, we can see how strongly Descartes is committed to the idea that all emotions have purposes. The emotions that do not serve any functions are denied the status as independent emotions. They are rather regarded as the excessive instances of other emotions that serve some purposes in normal, non-excessive circumstances.

Most, if not all, contemporary emotion researchers in psychology accept the Cartesian idea that emotions have functions. Many of them think that emotions have their functions essentially. Their definitions of emotion, thus, are deeply teleological, referring to etiological function, adaptation, natural selection, etc.

> Emotions are part of the biological solution to the problem of how to plan and to carry out action aimed at satisfying multiple goals in environments which are not perfectly predictable. Examples of the multiple goals simultaneously pursued by mammals include: to find supplies of food and water, to hoard such supplies, to maintain oneself in proper climatic conditions, to avoid predators, to maintain territory, to find and court mating partners, to care for young, to guard one's position in the dominance hierarchy. (Oatley & Johnson-Laird 1987, 36)

> To behave functionally according to evolutionary standards, the mind's many subprograms need to be orchestrated so that their joint product at any given time is functionally coordinated, rather than cacophonous and self-defeating. This coordination is accomplished by a set of superordinate programs – the emotions. They are adaptations that have arisen in response to the adaptive problem of mechanism orchestration. (Tooby & Cosmides 2000, 92)

> The emotions are specialized modes of operation shaped by natural selection to adjust

> the physiological, psychological, and behavioral parameters of organism in ways that increase its capacity and tendency to respond adaptively to the threats and opportunities characteristic of specific kinds of situations. (Nesse 1990, 268)

TAF defines belief in a quite similar way that these definitions define emotions. These definitions define emotions in terms of their functions. TAF defines belief in terms of its function. These definitions allow for "malfunctioning emotions". For instance, according to Nesse, pathological anxieties in anxiety disorder do not increase the capacity to respond adaptively to the threats. Nonetheless, these anxieties are emotions, because they do have the function of increasing the capacity to respond adaptively to the threats. TAF allows for "malfunctioning beliefs". Some beliefs might fail to be sensitive to evidence or causing appropriate non-verbal behavior. Nonetheless, these beliefs are beliefs, because they do have the function of being sensitive to evidence or causing appropriate non-verbal behavior.

3.2.2 From Mental States to Mechanisms

I said that TAF is just a tentative proposal. It is just tentative because it is problematic at a basic issue about etiological function. Godfrey-Smith nicely explains the problem.

> It is one thing to say that a mechanism has a biological function. Eyes have evolved through natural selection, and have the function of picking up information in the form of light waves. Are particular states of this device, ipso facto, functionally characterizable? Does the etiology of this apparatus bestow a biological function on my current visual experience of 'Blue Poles'? An application of Wright Line would seem to

> deny functions to states. Structural features of the visual apparatus are products of an evolutionary history, a history of heritable variation in fitness. But, states of visual system are not the right sort of things to have such a history. There exists no way their properties and powers can lead to there being future states of the same type (as oppose to states of different type, not as opposed to no states at all). There exists no way success can beget success, in the relevant fashion. No features of an experience of 'Blue Poles' are likely to bring it about that, through natural selection, there are more visual experiences of 'Blue Poles' sort in the future. A state may or may not profit an organism, but its nature is not the product of the success of previous states of the same type, and has no propensity to lead to the future survival and proliferation of the type. (Godfrey-Smith 1989, 542)

TAF assumes that beliefs are the bearers of functions (of playing belief-like causal roles). But, beliefs can't be the bearers of any kinds of functions because, as Godfrey-Smith stresses, mental states such as beliefs are not the products of natural selection. My belief that Obama is the president of USA is, for instance, not the product of natural selection. I have this belief not because this belief has been selected, but because I learned from someone that Obama was reelected for presidency in 2012. Thus, we need to revise TAF in such a way that the resulting theory doesn't assume that beliefs are the bearers of functions. The theory I propose rather assumes that the mechanisms that produce and consume (or use) beliefs are the bearers of functions. (For a similar idea, see Millikan (1984, 1989a, 2002) on "derived function")

To say that a mental state with content "*p*" plays a belief-like causal role is to say that the state is caused by certain distinctive inputs (e.g. distinctive perceptual inputs, other

beliefs), and causing certain distinctive outputs (e.g. distinctive other beliefs, emotions, bodily behavior). I call these inputs and outputs 'B"*p*"-appropriate inputs' and 'B"*p*"-appropriate outputs'. In this terminology, to say that a mental state with content "*p*" has the function of playing a belief-like causal role is to say that the state has the function of being caused by B"*p*"-appropriate inputs and causing B"*p*"-appropriate outputs. Now, instead of attributing this function to the mental states, we could attribute the function to the mechanisms that produce and consume the states. In other words, the mechanism that produce the state has the function of producing it in response to B"*p*"-appropriate inputs, and the mechanism that consume the state has the function of producing B"*p*"-appropriate outputs in response to the state. Based on this idea, we can revise TAF in the following way.

TAF2: Belief

A mental state, with content "p", is a belief iff (1) its producer has the function of producing the state in response to B"p"-appropriate inputs, and (2) its consumer has the function of producing B"p"-appropriate outputs in response to the state.

According to TAF2, A mental state with content "p" is a belief just in case it has a right kind of producer and a right kind of consumer. The right kind of producer is the one with the function of producing the state at issue in response to B"p"-appropriate inputs. The right kind of consumer is the one with the function of producing B"p"-appropriate outputs in response to the state at issue.

Here are some sample cases.

Case 1: Suppose that there is a mental state of mine, *S1,* with the content "there is a bottle of beer in the fridge". *S1* is produced by a mechanism, *P1*, in response to the perceptual experience of the bottle of beer in the fridge. *P1* does this because it is a part of the function of *P1* to produce *S1* in response to the perceptual experience of the bottle of beer in the fridge. Suppose, also, that another mechanism, *C1*, makes me behave as if there is a bottle of beer in the fridge in response to *S1. C1* does this because it is a part of the function of *C1* to make me behave as if there is a bottle of beer in the fridge in response to *S1*. According to TAF2, *S1* is a belief. It has a right kind of producer and a right kind of consumer. The producer of *S1*, *P1*, has the function of producing *S1* in response to B"there is a bottle of beer in the fridge"-appropriate inputs (i.e. the perceptual experience of the bottle of beer in the fridge), and the consumer of *S1*, *C1*, has the function of producing B"there is a bottle of beer in the fridge"-appropriate outputs (i.e. my behaving as if there is a bottle of beer in the fridge) in response to *S1*.

Case 2: Suppose that there is a mental state of mine, *S2,* with the content "the world is coming to an end". *S2* is produced by a mechanism, *P2*, in response to the perceptual experience of some marble tables in a café. *P2* does this not because it a part of its function to produce *S2* in response to the perceptual experience of marble tables. Rather, *P2* does this because of its malfunctioning. *P2* is, when it is well-functioning, supposed to produce *S2,* not

in response to the perceptual experience of marble tables, but in response to good evidence for believing that the world is coming to an end, such as NASA's announcement that a 70-mile-wide asteroid is going to hit the earth in few weeks. Suppose, also, that another mechanism, *C2*, makes, in response to *S2*, me behave as if the world is coming to an end. *C2* does this because it is a part of the function of *C2* to make, in response to *S2*, me behave as if the world is coming to an end. According to TAF2, *S2* is a belief. It has a right kind of producer and a right kind of consumer. The producer of *S2*, *P2*, has the function of producing *S2* in response to B"the world is coming to an end"-appropriate inputs (e.g. NASA's announcement), and the consumer of *S2*, *C2*, has the function of producing B"the world is coming to an end"-appropriate outputs (i.e. my behaving as if the world is coming to an end) in response to *S2*.

Case 3: Suppose that there is a mental state of mine, *S3,* with the content "there is a bottle of beer in the fridge". *S3* is produced by a mechanism, *P3*, in response to the perceptual experience of the bottle of beer in the fridge. *P3* does this because it is a part of the function of *P3* to produce *S3* in response to the perceptual experience of the bottle of beer in the fridge. Suppose, also, that another mechanism, *C3*, makes me behave as if there is NO bottle of beer in the fridge in response to *S3*. *C3* does this not because it is a part of the function of *C3* to make me behave as if there is NO bottle of beer in the fridge in response to *S3*, but because it is malfunctioning. Rather, *C3* is, when it is well-functioning, supposed to

make me behave as if there is a bottle of beer in the fridge in response to *S3*. According to TAF2, *S3* is a belief. It has a right kind of producer and a right kind of consumer. The producer of *S3*, *P3*, has the function of producing *S3* in response to B"there is a bottle of beer in the fridge"-appropriate inputs (i.e. the perceptual experience of the bottle of beer in the fridge), and the consumer of *S3*, *C3*, has the function of producing B"there is a bottle of beer in the fridge"-appropriate outputs (i.e. my behaving as if there is a bottle of beer in the fridge) in response to *S3*.

Case 4: Suppose that there is a mental state of mine, *S4,* with the content "the world is coming to an end". *S4* is produced by a mechanism, *P4*, in response to the perceptual experience of some marble tables in a café. *P4* does this not because it a part of its function to produce *S4* in response to the perceptual experience of marble tables. Rather, *P4* does this because of its malfunctioning. *P4* is, when it is well-functioning, supposed to produce *S4,* not in response to the perceptual experience of marble tables, but in response to good evidence for believing that the world is coming to an end, such as NASA's announcement. Suppose, also, that another mechanism, *C4*, makes, in response to *S4*, me behave as if the world is NOT coming to an end. *C4* does this not because it is a part of its function to make me behave as if the world is NOT coming to an end in response to *S4*, but because it is malfunctioning. Rather, *C4* is, when it is well-functioning, supposed to make me behave as if the world is coming to an end in response to *S4*. According to TAF2, *S4* is a belief. It has a

right kind of producer and a right kind of consumer. The producer of *S4*, *P4*, has the function of producing *S4* in response to B"the world is coming to an end"-appropriate inputs (e.g. NASA's announcement), and the consumer of *S4*, *C4*, has the function of producing B"the world is coming to an end"-appropriate outputs (i.e. my behaving as if the world is coming to an end) in response to *S4*.

Case 5: Suppose that another mental state of mine, *S5*, with the content "I drink water", is produced by a mechanism, *P5*, in response to some states carrying information about the insufficient water intake. *P5* produces *S5* this because it is a part of its function to produce *S5* in response to those informational states. Suppose, also, that another mechanism, *C5*, makes me behave so that I drink water (go to the kitchen, get a glass, turn the faucet on, etc.) in response to *S5*. *C5* does this because it is a part of its function to make me behave so that I drink water in response to *S5*. According to TAF2, *S5* is not a belief. It doesn't have a right kind of producer and a right kind of consumer. The producer of *S5*, *P5*, doesn't have the function of producing *S5* in response to B"I drink water"-appropriate inputs, and the consumer of *S5*, *C5*, doesn't have the function of producing B"I drink water"-appropriate outputs.

Rather, we should say that *S5* is a desire. Here is the desire-version of TAF2.

TAF2: Desire

A mental state, with content "p", is a desire iff (1) its producer has the function of

producing the state in response to D“p”-appropriate inputs, and (2) its consumer has the function of producing D“p”-appropriate outputs in response to the state.

According to TAF-Desire, *S5* would be a desire. The producer of *S5*, *P5*, has the function of producing *S5* in response to D”I drink water”-appropriate inputs (i.e. the states carrying information about the insufficient water intake), and the consumer of *S5*, *C5*, has the function of producing D”I drink water”-appropriate outputs (i.e. making me behave so that I drink water) in response to *S5*.

TAF2 can be generalized as a general theory of propositional attitudes. Here is the general, schematic version of TAF2.

TAF2: General

A mental state, with content “p”, is an *X* iff (1) its producer has the function of producing the state in response to *X*“p”-appropriate inputs, and (2) its consumer has the function of producing *X*“p”-appropriate outputs in response to the state.

I close this chapter with a brief note on “normativity.” Is TAF2, like Bayne’s account, a normativist account of belief? The answer is “Yes and No.” It is a normativist account in so far as it regards biological function as essential for belief and biological function is a normative notion in the sense that biological function of something is the standard according to which it is supposed to work. For example, when we say that the function of kidney is to filter metabolic wastes from blood, we imply that it is the standard according to which

kidneys are supposed to work. Kidneys are supposed to filter metabolic wastes from blood. According to TAF2, thus, belief is a normative state in the same sense that kidneys, hearts, prefrontal cortex, etc. are normative entities. But, of course, this notion of "normative" is thin. Unlike other normative accounts of belief in the literature (e.g. Bayne, Davidson, Zangwill, Wedgwood), TAF2 doesn't say that the mental state of belief essentially involves the prescriptive norms of truth or rationality. In TAF2, truth or rationality does not play any significant roles in describing the essential features of belief.

Chapter 4: Theoretical Issues on TAF2

This chapter discusses two theoretical issues related TAF2. First, I will explain and justify my commitment to etiological function for TAF2 (3.1). Second, I will examine some challenges to TAF2 and show that these challenges can be met (3.2).

4.1 Etiological Function

4.1.1 Etiological Function and Malfunction

Many supporter of etiological analysis claim that etiological analysis gives the analysis of the concept of function in biology (e.g. Buller 1998; Godfrey-Smith 1994; Neander 1991a, 1991b). Let us call it "biological etiologism". The claim is controversial. An obvious problem is that, for instance, when Harvey discovered that the function of heart in 1616, he didn't know anything about Darwinian evolutionary biology. Another problem is that, according to the widely accepted distinctions introduced in "Tinbergen's Four Questions", the question about function (of behavior) is distinguished from the question about evolutionary history (of behavior). As Godfrey-Smith pointed out, "[t]his is clearly an embarrassment for any historical theory of function which seeks to capture biological usage: on the historical view

there should be three questions, not four as the functional question *is* a question about evolutionary history" (Godfrey-Smith 1994, 351). I am not particularly interested in defending biological etiologism here. My attitude toward etiological analysis is purely pragmatic. In other words, I am interested in etiological analysis because of its utility in philosophical theorizing. This attitude can be called "pragmatic etiologism". Millikan, another pragmatic etiologist, expresses this attitude in the following way.

> The point of the notion "proper function" [which is Millikan's own version of etiological function] was/is mainly to gather together certain phenomena under a heading or category that can be used productively in the construction of various explanatory theories. The ultimate defense of such a definition can only be a series of illustration of its usefulness… (Millikan 1989a, 289)

Now, the utility of etiological function mainly comes from the fact that it allows for malfunction. In the last chapter, I said that etiological function clearly allows for malfunction, but other kinds of functions might not. I explain this in the following.

Three major alternatives to etiological function are,

Systemic Function

x has the systemic function of doing *F* iff (1) *x* is a part of a complex system (complex enough to allow for functional analysis) and (2) *F* is *x*'s contribution to the activity or capacity of the system. (Cummins 1975, 1983)

Goal-Contribution Function

x has the goal-contribution function of doing *F* iff (1) *x* is a part of a goal-directed system and (2) *F* is *x*'s contribution to the achievement of the goal. (Boorse 1976, 2002)

Life-Chance Function

x has the life-chance function of doing *F* iff *x*'s performance of *F* contributes to its bearer's having better life chances than the imaginary duplicate of the bearer without *x*. (Bigelow & Pargetter 1987)

My kidney has the etiological function of filtering metabolic wastes from blood. It also has the systemic, goal-contribution, and life-chance functions of doing it. The kidney has the systemic function of doing it because (1) the kidney is a part of my body, which is a complex system, and (2) filtering metabolic wastes from blood is the kidney's contribution to the activity or capacity of the body. The kidney has the goal-contribution function of doing it because (1) the kidney is a part of my body, which is a goal-directed system, and (2) filtering metabolic wastes from blood is the kidney's contribution to the achievement of the goals such as survival and reproduction. The kidney has the life-chance function of doing it because the kidney's filtering metabolic wastes from blood contributes to my having better life chances than the imaginary duplicate of mine without kidney.

At the first glance, these alternatives do not seem to allow for malfunctionability. In other words, they do not seem to allow for the possibility that *x* it has function of doing *F* and it fails to do *F*. This is because the failure of actually doing *F* rules out the possibility that *x* has the function of doing *F*, according to the alternatives. Suppose that John's kidney fails to

filter metabolic wastes from blood due to renal failure. The kidney doesn't have the systemic function of filtering metabolic wastes from blood because the kidney doesn't actually filter metabolic wastes from blood and, hence, the filtering can't be the kidney's contribution to the activity or capacity of John's body. The kidney doesn't have the goal-contribution function of doing it because the kidney doesn't filter metabolic wastes from blood and, hence, the filtering can't be the kidney's contribution to John's survival and reproduction. The kidney doesn't have the life-chance function of doing it because it doesn't filter metabolic wastes from blood and, hence, the filtering can't contribute John's having better life chances than the imaginary duplicate without kidney.

Some argue, however, that we could say that these functions allow for malfunction by simply changing the interpretation of the term "malfunction". One possible suggestion is that, instead of interpreting "*x* is malfunctioning" as "*x* has the function of doing *F* and fails to do *F*", we can interpret it as "*x* lacks the function of doing *F* and most of the other tokens of the same type have the function of doing *F*. (e.g. Prior 1985; Godfrey-Smith 1993; Roe & Murphy 2011) In doing so, other functions would allow for malfunctionability. For instance, John's kidney is malfunctioning because it lacks the systemic function of filtering metabolic wastes from blood and most of the other kidneys have the systemic function of doing so. But, this proposal is not very attractive for two reasons. First, according to this proposal, some kidney-tokens, such as John's, don't have the function of filtering metabolic wastes from

blood, while other well-functioning tokens do have the function. This means that kidney-type can't be defined in terms of the function of filtering metabolic wastes from blood. We can't say that kidneys are the organs with the function of filtering metabolic wastes from blood. But, this goes against the popular view that organ types such as kidney-type are defined in terms of their functions. Second, if, for instance, renal failure becomes much common for some reasons and it turns out that it is not statistically rare anymore, then we have to say, according to this proposal, that John's kidney is not malfunctioning anymore, which is absurd. As Neander noted, "we can't cure diseases just by spreading them around" (Neander 1995, 111).

Another possible suggestion would be to, first, distinguish the function of tokens from the function of types and, then, interpret "*x* is malfunctioning" as "*x* lacks the function of doing *F* and the type to which *x* belongs has the function of doing *F*" (Boorse 2002; Cummins & Roth 2010). In doing so, other functions might allow for malfunctionability. For instance, John's kidney is malfunctioning because it lacks the goal-contribution function of filtering metabolic wastes from blood and kidney-type has the goal-contribution function of doing so. But, this proposal is not very attractive unless some non-statistical ways of characterizing the functions of types are available. If "kidney-type has the goal-contribution function of filtering metabolic wastes from blood" just means "most of the tokens of the kidney-type have the goal-contribution function of filtering metabolic wastes from blood", then this

suggestion is basically the same as the first one and, thus, it faces exactly the same problem.

4.1.2 Counterfactual Function?

Nanay (2010) proposed a new kind of function, which is explicitly designed to allow for malfunctionability. I call it "counterfactual function".

> **Counterfactual Function**
> *x* has the counterfactual function of doing *F* iff (1) *x* performs *F* in some nearby possible worlds and (2), in the closest possible world where *x* performs *F*, the performance positively contributes to the inclusive fitness of its bearer.

Counterfactual function allows for malfunctionability. Although John's kidney doesn't filter metabolic wastes from blood, it has the counterfactual function of doing it, because (1) there are some nearby non-actual possible worlds where it does filter metabolic waste from blood (e.g. the world where John overcomes renal failure after appropriate medical treatments) and (2), in the closest possible world where it does filtering, the filtering positively contributes to John's inclusive fitness.

I am not, however, very attracted to counterfactual function. Here are the reasons.

Counterfactual analysis of function simply regards function as something evolutionary useful (in a counterfactual situation). This causes some potential problems. First, there might be useless functions. Second, there might be some non-functions that are accidentally

useful.

(1) *Useless Functions*: It would be plausible to say that a polar bear's fur has the function of reducing heat loss even if the bear resides in a zoo in the tropics, or that the wing muscles of birds have the function of enabling flight, even in small birds on stormy islands where flight is detrimental to survival. But, counterfactual analysis wouldn't allow these function attributions. For instance, in the closest possible world where the fur of the polar bear in a zoo in the tropics reduces heat loss (which is actual world), it doesn't positively contribute to inclusive fitness (and, hence, the second clause is not satisfied). So, the fur doesn't have the function of reducing heat loss. On the other hand, the fur has the etiological function of reducing heat loss because the fur of polar bears was selected for reducing heat loss.

(2) *Useful Non-Functions*: My nose has the counterfactual function of supporting eyeglasses, my heart has the counterfactual function of making irregular beat when the irregular beat helped diagnosis, the legs of Michael Phelps have the counterfactual function of serving as paddle, and so on. For instance, my nose supports eyeglasses in some nearby possible worlds including actual world, (and, thus, the first clause is satisfied) and in the closest possible world where my nose does it (which is actual world), it positively contributes to my inclusive fitness (and, thus, the second clause is also satisfied). All of these are somewhat strange function attribution. On the other hand, etiological function doesn't allow for these strange function attributions. For instance, my nose doesn't have the etiological

function of supporting eyeglasses simply because human nose wasn't selected for supporting eyeglasses.

Another worry is related to the first clause of counterfactual analysis. The clause says that *x* performs *F* in some nearby possible worlds. This clause is needed to rule out the consequence, for instance, that the eyes of an impala have the counterfactual function of killing predators by emitting laser beam. There is no interesting sense of "function" according to which an impala's eyes have the function of killing predators by emitting laser beam. The first clause is needed to rule this out. (The second clause of the analysis can't rule this out. In the closest possible world in which its eyes kill predators by emitting laser beam, the killing positively contributes to its inclusive fitness.) The eyes of an impala doesn't have the counterfactual function of killing predators by emitting laser beam, because there is no nearby possible word where its eyes kill predators by emitting laser beam (and, hence, the first clause is not satisfied).

But, depending on how we define "nearby" here, counterfactual analysis might give some problematic function attributions.

First, suppose that John is suffering from a very serious (imaginary) incurable kidney disorder. It is incurable and, thus, he can't overcome it without some medical miracles. (We could, if it gives a better case, assume that this disorder is not just incurable, but also unavoidable due to, for instance, the fact that it is caused by some chromosome

abnormalities.) As long as John's kidney is a human kidney, we will say that it has function of filtering metabolic wastes from blood and it is malfunctioning in failing to do so. Etiological analysis of function, on one hand, enables us to say this. Human kidney was selected for filtering metabolic wastes from blood. On the other hand, it is not clear that counterfactual analysis enables us to say that. For, it is not clear that there are some nearby possible worlds where his kidney filters metabolic wastes from blood. A medical miracle is needed for him to overcome his condition. Accordingly, it is not clear that the first clause is satisfied. But, if there is no nearby possible world where his kidney filters metabolic wastes from blood, then we should conclude that the kidney doesn't have the counterfactual function of filtering metabolic wastes from blood and, hence, it is not counterfactually malfunctioning in failing to do so.

Second, suppose that, in near future, various kinds of powerful enhancement techniques are developed and, with the help of these techniques, various kinds of enhanced capacities become available. For instance, human eyes can detect ultraviolet, working memory can store over 30 random combinations of letters and numbers for few hours, people can run 100m in 5 seconds, and so on. Then, there would be some nearby possible words where a man's eyes can detect ultraviolet, his working memory stores over 30 random combinations of letters and numbers for few hours, his legs enables him to run 100m in 5 seconds. (Actual world is one of those worlds if he actually has those enhanced capacities.) Assuming that

these enhanced capacities positively contribute to inclusive fitness, we have to say that his eyes have counterfactual function of detecting ultraviolet, his working memory has function of storing over 30 random combinations of letters and numbers for hew hours, his legs have the function of enabling running 100m in 5 seconds, and so on. But, these are certainly strange function attributions. On the other hand, etiological function doesn't allow for these function attributions because, for instance, human eyes weren't selected for detecting ultraviolet. Ancestral human eyes didn't detect ultraviolet in the first place.

4.1.3 Does Etiological Function Fail to Allow for Malfunction?

Etiological function allows for malfunctionability, while many other alternatives don't. Davies (2000, 2001), however, argued that etiological function doesn't in fact allow for malfunctionability. If his argument is successful, TAF2, which relies on etiological function, doesn't allow for malfunction, which causes a problem.

His argument rests upon four premises. (The following discussion includes my own reconstruction of the argument.)

(1) A defective trait token loses the property selected for.
(2) Loss of the property selected for disqualifies the token from "selected functional type" that is defined in terms of the property.
(3) The tokens that do not qualify as the members of the "selected functional type" cannot possess the etiological function associated with the "selected functional type".

(4) A token of a trait etiologically malfunctions iff it fails to do *x* and it has the etiological function of doing *x*.

A clarification. Davies defined the term "selected functional type" in a special way. For example, kidney-type is not a selected functional type. Kidney-type is, rather, a "generic type" according to his terminology. A generic type has defective tokens such as defective kidney-tokens. On the other hand, selected functional types don't have defective tokens. Selected functional types are, therefore, things like non-defective-kidney-type or non-defective-heart-type.

The following are the kidney-version of these premises.

(1) A defective kidney-token loses a property selected for (i.e. fails to filter metabolic wastes from blood).
(2) Failing to filter metabolic wastes from blood disqualifies the token from non-defective-kidney-type.
(3) The tokens that do not qualify as members of non-defective-kidney-type cannot possess the etiological function of filtering metabolic wastes from blood.
(4) A kidney-token etiologically malfunctions iff it fails to filter metabolic wastes from blood and it has the etiological function of doing it.

The argument goes like this. Suppose that *x* is a defective kidney. From (1), it follows that *x* fails to filter metabolic wastes from blood. From (2), then, it follows that *x* doesn't qualify as a member of well-functioning-kidney-type. From (3), then, it follows that *x* cannot possess the etiological function of filtering metabolic wastes from blood. And, from (4), *x* doesn't

etiologically malfunction.

The problem of this argument is about the premise (3). Indeed, this premise is inconsistent with the very central idea of etiological function. The premise (3) is committed to the idea that that the membership of non-defective-kidney-type is necessary for something to have the etiological function of filtering metabolic wastes from blood. But, nothing about etiological function implies that it is necessary. All that is required for something to have the etiological function of filtering metabolic wastes from blood is to have a right kind of history. The membership of non-defective-kidney-type is not required.

Still, Davies seems to think that etiological theorists should accept the premise (3). He is working on the following definition of etiological function,

> *x* of the type *T* in organism *O* in selective environment *E* has the function of doing *F* iff
> I. Past instances of *T* in *O* performed *F* in *E*,
> II. *T* was heritable,
> III. Past performances of *F* caused an increase in O's relative ability to satisfy demands of *E* (relative to other organisms in the population lacking *T*),
> IV. This increase in *O*'s ability to satisfy selective demands of *E* resulted in an increase in *O*'s long-term relative rate of reproduction,
> V. This increase in relative reproduction resulted in the persistence or proliferation of *O* and hence tokens of *T*.

Davies argues that, in this definition, *T* should be understood as the variable for selected functional types, such as non-defective-kidney-type, not generic types, such as kidney-type.

This is because, according to Davies, the variable *T* in the right hand side, especially in I and II, is for selected functional types, not for generic types. (He thinks that *T* in V is ambiguous but it should be interpreted as for selected functional types for the sake of avoiding equivocation.) But, why? Why can't we think of I and II as talking about kidney-type? Why, for instance, can't we say that the past tokens of kidney-type performed the filtering metabolic wastes from blood? Maybe, Davies thinks takes I to be saying that *all* past instances of *T* in *O* performed *F* in *E*. In this case, certainly *T* can't be for kidney-type, because it is not the case that all past instances of kidney-type filter metabolic wastes from blood. But, etiological theorists do not accept this idea, and there is no reason that they should accept it. There is no reason to accept it because there is no reason to accept the idea, for instance, that in order for kidney to be selected for filtering metabolic wastes from blood, all past tokens of the type need to do the job of filtering.

4.2 Objections to TAF2

4.2.1 Philosophical Objections

I classify potential objections to TAF2 into two broad categories. First, there are some philosophical objections that invoke imaginary philosophical cases. Second, there are some empirical objections that invoke empirical, evolutionary issues. I discuss these objections in

tern in the following.

There are two sub-categories within philosophical objections. First, one might think that the right hand side of TAF2 is not necessary for someone to believe something. In other words, one might think that there are some cases where someone believes something but the right hand side of TAF2 is not satisfied. These objections are "necessity objections". Second, one might think that the right hand side of TAF2 is not sufficient for someone to believe something. In other words, one might think that there are some cases where someone doesn't believe something but the right hand side of TAF2 is satisfied. These objections are "sufficiency objections".

The representative necessity objection is famous Swampman objection.

> ***Swampman***
> Suppose lightning strikes a dead tree in a swamp; I am standing nearby. By body is reduced to its elements, while entirely by coincidence (and out of different molecules) the tree is turned into my physical replica. My replica, The Swampman, moves exactly as I did; according to its nature it departs the swamp, encounters and seems to recognize my friends, and apprears to return their greetings in English. It moves into my house and seems to write articles on radical interpretation. No one can tell the difference (Davidson 1987, 443).

Swampman objection goes like this. First, intuitively, Swampman has beliefs and desires. After all, Swampman behaves exactly the same way as Davidson does and, hence, we can explain and predict its behavior by attributing beliefs and desires for the same reason that

we can explain and predict Davidson's behavior by attributing beliefs and desires to him. Second, Swampman has no cognitive mechanisms with any etiological functions, which means that the right hand side of TAF2 is not true about him. Hence, this case shows the right hand side of TAF2 is not necessary for someone to belief something. Swampman doesn't satisfy the right hand side of TAF2, but he is a believer.

The representative sufficiency objection is what I call "mad belief objection". "Mad belief" case is an imaginary case, a parody of Lewis's mad pain case, by Schwitzgebel (2012).

> ***Mad Belief***
> Daiyu, let's suppose or at least let's try to suppose, believes that most pearls are white. However, this belief was not caused in the normal way. It was not caused, for example, by having seen white pearls nor by hearing testimony to the effect that pearls are white nor by inferring that peals are white from some other facts about pearls and whiteness. It was caused, let's say, by having spent 4 s watching the sun set over the Pacific Ocean. And, for her, that is just the sort of event that would cause that belief: Daiyu would never have formed that believe by any normal means like those described above; rather the kinds of events that cause that belief in her in all "nearby possible worlds", or across the relevant range of counterfactual circumstances, are perceptions of setting-sun events of a certain sort, and maybe also eating spicy radish salad on a Wednesday. The kinds of events that would cause her to cease believing that most pearls are white are also atypical; watching the sun rise over the Atlantic, perhaps, or putting daisies in her hair. Furthermore, Daiyu's belief that most pearls are white has entirely atypical effects. It does not cause her to say anything like "most pearls are white" (which she would like to deny; she'd say instead that most pearls are black) or to think to herself in inner speech that most pearls are white. She would not feel surprise were she to see a translucent purple pearl. If a friend were to say to Daiyu that she was looking for white jewelry to accompany a dress, Daiyu would not at all be inclined to recommend a pearl necklace. Nor is he disposed to infer from her believe that most pearls are white that there is a type of precious object used in jewelry that is white. Daiyu's belief that pearls are white, instead, causes her to flush

on the left side of her body when talking on the telephone and to say "17" whenever someone asks her to pick a number between one and 20 (Schwitzgebel 2012, 14).

Mad belief objection goes as follows. First, intuitively, Daiyu fails to believe that most pearls are white. The mental state Daiyu is in in the described situation is not a belief. Indeed, Schwitzgebel says that "mad belief of this radical sort, I hope you will agree, is inconceivable, or at least inconceivable in any sense that can serve as a guide to possibility." (Schwitzgebel 2012, 14) Schwitzgebel's point is that it is impossible that Daiyu's mental state, playing the strange causal role, is a belief. Second, it might turn out that the right hand side of TAF2 is true about Daiyu. In other words, it might turn out that the producer of his mental state at issue has the function of producing the state in response to B"most pearls are white"-appropriate inputs, and its consumer has the function of producing B"most pearls are white"-appropriate outputs in response to the state. Certainly, the producer is not actually producing the state in response to B"most pearls are white"-appropriate inputs and the consumer is not producing B"most pearls are white"-appropriate outputs in response to the state. But, this doesn't rule out the possibility that the producer and the consumer have the functions of doing these things. Hence, there can be some cases where Daiyu doesn't believe that most pearls are white, even though the mental state at issue satisfies the right hand side of TAF2. This means that the right hand side of TAF2 is not sufficient for someone to believe something.

Before presenting my responses to these objections, let me add some clarificatory comments on Swampman case. Swampman case is used in varieties of philosophical discussions, and it is important to distinguish different kinds of Swampman arguments, making sure which type of Swampman argument we are talking about here.

(1) *Swampman Objection to Etiological Function*: Some people take Swampman case to be giving a problem for etiological analysis of function (Boorse 1976; Bigelow & Pargetter 1987). Swampman is not a bearer of etiological functions. For instance, his heart (or the entity which looks exactly like a heart) doesn't have the etiological function of pumping blood, his kidney (or the entity which looks exactly like a kidney) doesn't have the function of filtering metabolic wastes from blood, and so on. Some people take this to be counterintuitive. They argue that, intuitively, his heart has function of pumping blood, his kidney has the function of filtering metabolic wastes from blood, and so on.

This challenge is, however, not a very serious one. For the sake of argument, I grant that, intuitively, Swampman is a function bearer (although I don't take it to be intuitively obvious). Now, as I already said, I am a pragmatic etiologist. In other word, I accept etiological analysis of function because of its theoretical features. It is not my claim that etiological function captures the intuition that is associated with folk or biological concept of function. Thus, the fact that etiological analysis has some minor counterintuitive consequences is not very problematic as long as the analysis is useful enough. Second, this

objection shows, at best, that there is a concept of function according to which Swampman is a function-bearer. Still, it doesn't rule out the possibility that there is another concept of function according to which he is not a function-bearer, and the concept is captured by etiological analysis. Indeed, many supporters of etiological analysis are pluralists, and they happily accept different concepts of function that are not captured by etiological function (Buller 1998; Millikan 1989a, 2002).

So, what is at issue here is not that Swampman is not a function-bearer (which is not very problematic), but that *Swampman has no beliefs according to TAF2*.

(2) *Swampman Objection to Representational Theory of Consciousness*: Often, Swampman case is discussed as the counterexample for the representational theory of consciousness (e.g. Block 1998; Dretske 1997; Tye 1998). More precisely, Swampman case is intended to be the counterexample to the conjunction of etiological analysis, teleosemantics, and representational theory of consciousness. The objection goes as follows. According to etiological analysis, Swampman is not a function-bearer. Thus, he doesn't have any mental states with etiological functions. According to teleosemantics, this implies that Swampman has no contentful mental states. This implies, in turn, according to representational theory of consciousness, that Swampman has no conscious mental states. But, this is counterintuitive. Since Swampman is exact physical duplicate of Davidson, and since consciousness supervenes physical states, and since Davidson is conscious, it has to be the

case that Swampman is also conscious. If this is really a good objection, then we should reject, at least, one of three commitments involved here. This issue, however, has little to do with my discussion. In my discussion, Swampman case is intended to be the counterexample to the conjunction of etiological analysis and TAF2. Even if Swampman case turns out to be a good counterexample to the conjunction of etiological analysis, teleosemantics and representational theory of consciousness, it might not be a good counterexample to the conjunction of etiological analysis and TAF2. For instance, the idea that consciousness supervenes physical state plays a big role in the Swampman objection to the former. The idea, even if it is plausible, doesn't play any role in the Swampman objection to the latter.

So much for clarifications. Now, my responses to Swampman objection, the representative necessity objection, and mad belief objection, the representative sufficiency objection, are essentially the same. The idea comes from Papineau's (2001) response to Swampman objection to teleosemantics, which is essentially the same as Swampman objection to TAF2. The core idea behind Papineau's response is that teleosemantics is a scientific claim about how the actual world is like. According to Papineau, the main commitment of teleosemantics is the idea that the core characteristics of belief are, as a matter of scientific fact in the actual world, displayed by a certain kind of selectional states with a certain kind of etiological functions. Swampman objection can't be the counterexample to this. The case only shows that there are some non-actual possible words

where something other than selectional states displays the core characteristics of beliefs. This is perfectly compatible with the claim, which is the main claim of teleosemantics, that those characteristics are displayed by selectional states in the actual world.

Please note that this response is different from a more popular response deployed by Millikan (1996) and Neander (1996). They take Swampman case to be analogous to the case where the core characteristics of water (e.g. transparency, falling from sky in rainy days, etc.) are displayed by something other than H2O, say, XYZ. It would certainly be possible that XYZ, not H2O, displays the core characteristic of water. But, this doesn't imply that it is possible that water is not H2O. Water, according to the standard Kripkean picture, rigidly refers to H2O in every single possible world given the fact that it refers to H2O in the actual world. Similarly, according to Millikan and Neander, it is possible that the core characteristics of belief is displayed by something other than selectional states. Swampman case shows that possibility. But, it doesn't imply that it is possible that beliefs are not selectional states for the same reason in the case of water. This response, obviously, assumes that the term "belief" is a rigid designator. But this assumption is controversial. Papineau's response, on the other hand, doesn't rest on this assumption. Papineau pointed out that the semantics of the term "belief" shouldn't be the main concern of teleosemantics, because teleosemantics is a scientific claim about the nature of states that display a certain kind of characteristics in the actual world. How the term "belief" refers to "may raise issues

of some interest, but if so, they are surely orthogonal to the concerns which motivate teleosemanticists" (Papineau 2001, 286).

Basically, I repeat the same response to both Swampman objection and mad belief objection to TAF2. Papineau regards teleosemantics as a scientific claim about the nature of states that display a certain kind of characteristics in the actual world. I also regard TAF2 as a scientific claim about the nature of states that display a certain kind of characteristics in the actual world. Swampman case shows there are some non-actual possible words where something other than selectional states displays the core characteristics of beliefs. Mad belief case shows (or can be used to show) that there are some non-actual possible words where selectional states do not display any of the core characteristics of beliefs. But, these possibilities are perfectly compatible with the claim, which is the main claim of TAF2, that those characteristics are, as a matter of scientific fact in the actual world, displayed by the selectional states.

In effect, I take TAF2 to be the same kind of theory as the evolutionary theories of emotion by Oatley & Johnson-Laird, Tooby & Cosmides and Nesse. The idea that emotions are shaped by natural selection is a crucial part of those theories of emotions. But, it is not a very good objection to them that Swampman has emotions even though their mental states (or mechanisms) weren't selected in the past. Swampman objection completely misses the point here because those theories of emotions are making scientific claims about the nature

of states that display a certain kind of characteristics in the actual world. Swampman objection and Mad Belief objection to TAF2 miss the point for exactly the same reason.

4.2.2 Empirical Objection

The most crucial empirical objection is adaptationism objection. Adaptationism objection goes as follows. There is a real chance (not mere logical or metaphysical possibility) that our mind is not adaptations. Especially, there is a real chance that belief-producers and belief-consumers are not adaptations. In that case, we are not believers according to TAF2. The problem, accordingly, is that there is a real chance that we are not believers according to TAF2. But, this is absurd, because it is obvious that we are believers.

Here is a preliminary remark on this objection. According to this objection, TAF2 allows for the real chance that we are not believers. And this is said to be a problem. If it is a real problem, however, other popular theories of belief would face similar problems. Think, for instance, about standard functionalism. According to standard functionalism, believers are those who have some states that play belief-like causal roles. But, doesn't it make some empirically falsifiable predictions about, for instance, functioning and structure of brain? What if it turns out that we don't have any occupiers of belief-like causal role in our heads? It looks as though standard functionalism too allows for the real chance that we are not believers. Standard functionalist might respond by saying that standard functionalism is an

"abstract" theory in the sense that it is free from any commitment about functioning and structure of brain. For instance, standard functionalism is designed to allow for multiple-relizability of mental states, which means that the theory doesn't give any falsifiable predictions about the physical make-up of believers. But, as Churchland (1981) pointed out, this "abstract" strategy wouldn't be very satisfactory. His objection is that the "abstract" strategy proves too much. For instance, the exactly the same strategy can be used to make alchemy an empirically unfalsifiable theory. He wrote,

> Being 'ensouled by mercury' or 'sulphur,' or either of the other two so-called spirits, is actually a functional state. The first, for example, is defined by the disposition to reflect light, to liquefy under heat, to unite with other matter in the same state, and so forth. And each of these four states is related to the others, in that the syndrome for each varies as a function of which of the other three states is also instantiated in the same substrate. Thus, the level of description comprehended by the alchemical vocabulary is abstract: various material substances, suitably 'ensouled' can display the features of a metal, for example, or even of gold specifically. For it is the total syndrome of occurrent and causal properties which matters, not the corpuscularian details of the substrate. Alchemy, it is concluded, comprehends a level of organization in reality distinct from and irreducible to the organization found at the level of corpuscularian chemistry (Churchland 1981, 80-81).

So, if it is problematic for a theory of belief to allow for the real chance that we are not believers, the problem is shared by standard functionalism (and other similar theories) as well.

Now, let us look at the objection more closely. In the following, I will critically examine

the premise of the objection that there is a real chance that our mind is not an adaptation, referring to two main figures who would strongly support this premise; Fodor and Gould.

Let us begin by Fodor. Here is what he says about adaptationism about human psychology.

> Since psychological structure (presumably) supervenes on neurological structure, genotypic variation affects the architecture of the mind only via its effect on the organization of the brain. And, since nothing at all is known about *how* the architecture of our cognition supervenes on our brain's structure, it's entirely possible that quite small neurological reorganizations could have effected wild psychological discontinuities between our minds and the ancestral ape's. This is really *entirely* possible; we know nothing about the mind/brain relation with which it's incompatible. In fact, the little we do know points in the other direction: Our brains are, at least by any gross measure, very similar to those of apes; but our minds are, at least by any gross measure, very different. So it looks as though relatively small alternations to the neurology must have produced very large discontinuities ("saltations," as one says) in cognitive capacities in the transition from the ancestral apes to us. If that's right, then there is no reason at all to believe that our cognition was shaped by the gradual action of Darwinian selection on prehuman behavioral phenotypes (Fodor 2001, 87-88).

Here, Fodor is arguing against the argument for adaptationism about human psychology from complexity, according to which human psychology is an adaptation because it is extremely complex and sophisticated, and the adaptationist hypothesis gives the best explanation of the complexity and sophistication. Fodor argues, in response to this argument, that something other than natural selection can also produce complex and sophisticated things. In particular, he claims that we can't rule out the possibility that our

mind, complex and sophisticated, is the product of the small accidental change in neural structure. Nothing we know about mind-brain relation rules out the possibility that small accidental neural change brings pretty complex and sophisticated psychological capacities. If Fodor is correct, then, there is probably a real chance that our belief-producers and belief-consumers are the products of small accidental change in neural structure as opposed to natural selection. But, then, it follows from this that there is a real chance that those mechanisms do not have any etiological functions. I call this "Fodorian objection".

It is, however, not very difficult to see the problem of Fodorian objection. First, Fodor just assumes that the small accidental change hypothesis is a rival, competing hypothesis to adaptationist hypothesis. This is simply false. Even if small accidental change hypothesis turns out to be true, it is perfectly compatible with the truth of adaptationist hypothesis. For instance, when the new neural structure, produced by a small accidental change, proliferated in the population over generations because of its positive contribution to inclusive fitness, we just say that the neural structure was selected. Millikan wrote in response to Fodor, "how does Fodor suppose that the very small genetic change in one of our lucky ancestors just happened to get handed down to all the rest of us?" (Millikan 2004, 8) Millikan's point is that, when Fodor tries fill in the story about how the new neural structure proliferated in the population over generations, the story will just turn out to be an adaptationist story (see also (Okasha 2003)).

Second problem of Fodorian objection is that even if it turns out that something is not the product of natural selection in its evolutionary origin, it doesn't rule out the possibility that it has etiological function. For, there is no reason to think that etiological function is wholly determined by the natural selection working at the origins of traits. Indeed, many philosophers, such as Godfrey-Smith (1994) or Millikan (2002), think that the etiological function is not determined by the natural selection working at origins of traits but rather by the relatively recent natural selection working for the maintenance of the traits. Those philosophers, "modern" etiological theorists, would say that, even if it turns out that belief-producers and belief-consumers are not the products of natural selection in their evolutionary origins, they still have etiological functions as long as they have been maintained relatively recently by natural selection.

Let us turn to Gould. He introduced, in the famous paper co-authored with Lewontin (1979), the term "spandrel" as referring to the traits that are the by-products of selected traits, and are not selected themselves.

Here is a quick terminological note. "Spandrel" shouldn't be confused with another famous term "exaptation" (Gould & Vrba 1982). Spandrels are by-product traits, while exaptations are "co-opted" traits. They are different, although they partially overlap. Some exaptation were initially selected for a purpose, and later co-opted for other purposes. For instance, feather of birds were initially selected for the purpose of thermal regulation, and

later co-opted for flight. Other exaptations initially proliferated as the by-products of other traits, and later they were co-opted for some purposes. This kind of exaptations are spandrels at the same time. For instance, the architectural spandrel, the spaces left over between structural elements of a building, proliferated as the by-product of core elements of buildings, and later co-opted for various purposes, including aesthetic ones.

Gould argues that there are so many psychological spandrels, much more than psychological adaptations.

> The human brain, as nature's post complex and flexible organ, throws up spandrels by the thousands for each conceivable adaptation in its initial evolutionary restructuring. What, then, by the criterion of relative frequency, is the best strategy for a useful evolutionary psychology - the sociobiology of strict Darwinism (which can only access the tiny proportion of adaptive traits), or a structural and correlational analysis that tries to map the spandrels of the brain's evolved capacity? (Gould 1991, 58)

As the examples of psychological spandrels, he mentions religion, fine and practical arts, the norms of commerce, the practice of war, and so on. Now, if it is true that there are much more psychological spandrels than psychological adaptations, then there is a real chance that the belief-producers and belief-consumers are spandrels too. But, if these mechanisms are psychological spandrels, they do not have any etiological functions. I call this "Gouldian objection".

Gouldian objection, I think, is misguided. First, Gouldian objection shares the same

mistake with Fodorian objection in assuming that *x*'s being a by-product of something in its evolutionary origin rules out that *x* has some etiological functions. Gould's usage of the term "function" supports this assumption. For him, *x* has the function of doing *F* just in case *x* was selected for doing *F* in its evolutionary origin (Gould & Vrba 1982). But, modern etiological theorists do not accept this usage. It might turn out that *x*, a by-product in its evolutionary origin, has been maintained for some purposes by relatively recent selection. In that case, *x* has those purposes as its etiological functions according to modern theorists.

Second problem is more important. The term "spandrel" blurs the distinction, which is extremely important, between the mechanisms that are by-products, and the performances that are by-products. Think, first of all, about a heart. The heart is, undoubtedly, the product of natural selection, not a by-product of something else. As an organ, the heart is not a by-product. But, some of its performances can be called "by-products" (or, more plausibly, "side-effects"). For instance, a heart produces some noise, which is the by-product of its functional performance; pumping blood. Crucial point here is that the term "by-product" can be used in two different levels; the level of organ and the level of the performance of organ. When I say that an organ is a by-product, I imply that the organ is not the product of natural selection. When I said that a performance of an organ is a by-product, I imply that the performance is not the function of the organ (i.e. the organ wasn't selected for the performance). Creating noise, for instance, is not the function of a heart (i.e. hearts weren't

selected for creating noise). The same distinction is crucial when it comes to psychological by-products. It is one thing to say that a certain psychological mechanism is the by-product of something else. It is another thing to say that some of its performances are the by-product of its other performances (that are the functions of the mechanism). When I say that a mechanism is a by-product, I imply that the mechanism is not the product of natural selection. When I said that a performance of a mechanism is a by-product, I imply that the performance is not the function of the mechanism (i.e. the mechanism wasn't selected for the performance).

With this distinction at hand, we clearly see the place where Gouldian objection goes wrong. The claim at issue here is that the belief-producing mechanisms and belief consuming mechanisms are by-products of something else. If they are by-products, they are by-product mechanisms. In supporting the claim that there is a real chance that these mechanisms are by-product mechanisms, Grouldian needs to provide the reason to believe that there are many by-product mechanisms, presumably much more than adaptation mechanisms. But, all the examples Gould gives in support of this claim that there are much more psychological spandrels than psychological adaptations are, in fact, the examples of by-product performances of some mechanisms. The problem is that those examples do not make it probable that there are many by-product mechanisms any more than the examples of the by-product performances of heart such as producing noise, adding some weight to body,

etc. make it probable that heart is a by-product organ.

Religious belief, which is Gould's primary example, is a clear example of by-product performance. In cognitive science of religion, by-product view of religious belief is currently quite popular (Bloom 2012; Boyer 2001, 2003; Guthrie 1993; Pinker 1997). Here, the hypothesis is that production of religious beliefs is the by-product performances of some mechanisms that evolved for other purposes, such as mindreading, agent detection, understanding physical objects, and so forth.

> […] religion emerges out of capacities, traits, and inclinations that have evolved for other purposes. It is an evolutionary accident. More specifically, the notion is that certain universal religious beliefs—such as belief in supernatural beings, creationism, miracles, and body-soul dualism—emerge as by-products of certain cognitive systems that have evolved for understanding the physical and social world. (Bloom 2012, 185)

> According to this view, religious beliefs are the by-products of evolved cognitive mechanisms. These cognitive mechanisms enable us to reason about the intentional states of others and to recursively embed intentional states within other intentional states, and make it possible for us to think what others think, including absent or even dead persons, fictional characters, and also supernatural agents. There is no need to invoke a set of dedicated, input-restricted mechanisms for religion, or for representing God. (Pyysiäinen & Hauser 2009, 105)

So, for instance, one specific hypothesis of this type, suggested in the second quote, is that the belief in supernatural agents is the by-product performance of mindreading mechanism. Mindreading mechanism was selected for producing the beliefs about mental states of

human agents. But, as the by-product, the mechanism also produces the beliefs about the mental states of supernatural agents. For some reasons, the mechanism is designed to be too sensitive to the cues of mindful agents. Presumably, this is an adaptive design because, in the ancient environment, failing to detect existing mindful agents was much more devastating than mistakenly detect non-existing mindful agents. The production of the beliefs about the mental states of supernatural agents, according to this hypothesis, is the by-product performance of mindreading mechanism for the same reason that producing some noise is the by-product performance of heart. Heart is not selected for producing noise, but for pumping blood. Producing noise is just the by-product of pumping blood. Mindreading mechanism was not selected for producing beliefs about the mental states of supernatural agents, but for producing beliefs about the mental states of human agents. Producing beliefs about the mental states of supernatural agents is just the by-product of producing beliefs about the mental states of human agents. Please note that it is not the part of the hypotheses of this sort that the mechanisms underlying religious beliefs are by-product mechanisms. For instance, mindreading mechanism is not considered as a by-product mechanism. On the contrary, mindreading mechanism is assumed to be an adaptation for producing beliefs about the mental states of human agents.

The same thing is true for all of other examples Gould gives. Buller points out the same issue in discussing Grould's view (though he talked about exaptations, not spandrel),

> [...] the exaptations Gould sites are all examples of specific behaviors, mental acts, beliefs, attitude and preferences. Such phenomena are the *outputs* of proximate mechanisms, generated in response to the inputs from experience. But, evolutionary psychologists claim that our psychological adaptations are (some of) the proximate mechanisms that generate such outputs, not the outputs themselves. (Buller 2005, 85)

So, the upshot is that Gouldian objection fails to give any reasons to think that belief-producers and belief-consumers are by-product mechanisms. He correctly pointed out that there are many psychological by-product performances. But, as I already said, this does not make it probable that belief-producers and belief-consumers are by-product mechanisms any more than the examples of the by-product performances of heart make it probable that heart is a by-product organ.

Chapter 5 TAF2 and Delusions

In the last two chapters, I proposed, illustrated, and defended TAF2. This chapter discusses the implications of TAF2 on the nature of delusion. An obvious implication is that, since the theory is a compatibilist one, it makes DD and CDT compatible with each other. I have already talked about this. In this chapter, I will discuss two further implications. First implication is related the question as to whether DD is actually true (not merely compatible with CDT) according to TAF2 (4.1). I will conclude that there is a real chance (not mere logical or metaphysical possibility) that DD is actually true according to TAF2. Second implication is related to the pathological nature of delusion (4.2). I will argue that TAF2, coupled with a certain plausible view of the nature of disorder, gives a nice explanation of the reason why delusions are pathological mental states.

5.1 Doxasticism Reconsidered

5.1.1 Producer of Delusion

TAF2 allows for the possibility of beliefs without belief-like causal roles. Nothing in TAF2 requires that beliefs actually play belief-like causal roles. Rather, it only requires that

beliefs have a right kind of producer and a right kind of consumer. Thus, according to TAF2, DD is perfectly compatible with CDT. Now, it is one thing to say that DD is compatible with CDT, and it is another to say that DD is actually true. It is perfectly possible that DD is compatible with CDT, but it is false after all. In this section, I will examine if DD is actually true according to TAF2. In other words, I will examine if it is actually true that delusions are beliefs according to TAF2.

According to TAF2, a delusion is a belief just in case it has a right kind of producer and a right kind of consumer. Then, the question is, does a delusion have a right kind of producer and a right kind of consumer?

Let us begin by producer. What is the producer of delusions? In order to answer the question, we need to look at delusion formation process. Although delusion formation process is not perfectly understood yet, there have been significant progresses recently in "cognitive neuropsychiatry," the research field in which mental disorders are studied with cognitive neuroscientific methodology. My view is that if the current mainstream theories of delusion formation in cognitive neuropsychiatry are on the right track, then it is likely that delusions have a right kind of producer.

The main commitment of the standard picture in cognitive neuropsychiatry is the idea that delusions, at least monothematic delusions, are the responses to some abnormal experiential states. Following Campbell (2001), let us call this commitment "empiricism"

about delusion formation.

Empiricist account of Capgras delusion is quite popular. Ellis and Young (1990) propose that Capgras delusion arises from a deficit in face processing where, on one hand, the face recognition system is intact but, on the other hand, there is a loss of affective responses to familiar faces. This deficit is hypothesized as being caused by the disrupted connection between face recognition system and autonomic nervous system. Because of this deficit, the Capgras patient has an abnormal experience of seeing a face that looks just like a close relative, but without the affective response that would normally be an integral part of the experience. This hypothesis is supported by the finding that Capgras subjects have a reduced galvanic skin response to faces and in particular do not respond more to familiar faces than to unfamiliar faces (Elli et al., 1997).

Similarly, other kinds of monothematic delusions are hypothesized as the responses to some distinctive kinds of abnormal experiences (see the table below).

Candidates for Abnormal Experiences in Monothematic Delusions (Davies et al, 2001)

- *Capgras delusion*: Unusual experience of faces or a sense that "something is different" as a result of flattened affective responses
- *Cotard delusion*: Loss of strong emotional experiences and a feeling of emptiness or a sense that "everything is different" as a result of global affective flattening
- *Frégoli delusion*: Unusual experience of people as a result of heightened affective responses
- *Reduplicative paramnesia*: Unusual experience as a result of heightened affective responses or a heightened sense of personal significance attached to remembered

events

- *Alien control, thought insertion*: Loss of experience of self-initiation of action or thought
- *Unilateral neglect*: Loss of kinesthetic and proprioceptive experience of the arm and a feeling of the arm as being alien
- *Mirrored-self misidentification*: Unusual experience of one's own face seen in the mirror or experience of reflected objects as if they were on the other side of the glass with loss of the ability to interact fluently with mirrors

Now, let us look again at two hypothetical cases that I gave earlier for illustration of TAF2 (omitting consumer part).

Case 1: Suppose that there is a mental state of mine, *S1,* with the content "there is a bottle of beer in the fridge". *S1* is produced by a mechanism, *P1*, in response to the perceptual experience of the bottle of beer in the fridge. *P1* does this because it is a part of the function of *P1* to produce *S1* in response to the perceptual experience of the bottle of beer in the fridge.

Case 2: Suppose that there is a mental state of mine, *S2,* with the content "the world is coming to an end". *S2* is produced by a mechanism, *P2*, in response to the perceptual experience of some marble tables in a café. *P2* does this not because it a part of its function to produce *S2* in response to the perceptual experience of marble tables. Rather, *P2* does this because of its malfunctioning. *P2* is, when it is well-functioning, supposed to produce *S2,* not in response to the perceptual experience of marble tables, but in response to good evidence for believing that the world is coming to an end, such as NASA's announcement that a 70-mile-wide asteroid is going to hit the earth in few weeks.

As I already said, *S1* and *S2* have a right kind of producers. In other words, both *S1* and *S2* satisfy the first clause of the right hand side of TAF2.

Some views in cognitive neuropsychiatry fit very well with idea that delusions are like *S1* with regard to production process. For instance, Maher claims that delusions are normal and rational response to abnormal perceptual experience.

> It is the core of the present hypothesis that the explanations (i.e. the delusions) of the patient are derived by cognitive activity that is essentially indistinguishable from that employed by non-patients, by scientists, and by people generally. The structural coherence and internal consistency of the explanation will be a reflection of the intelligence of the individual patient. The content of the explanation will reflect the cultural experience of the patient with general explanatory systems (scientific, religious, political, etc.). In brief, then, a delusion is a hypothesis designed to explain unusual perceptual phenomena and developed through the operation of normal cognitive processes. (Maher 1974, 103)

This claim is recently supported and sophisticated by Coltheart and colleagues (2010). They support Maher's claim by showing that delusion formation process can be modeled as a rational Bayesian updating of credence. For instance, given the abnormal experience of seeing a face that looks just like the subject's wife without appropriate affective component (strictly speaking, Coltheart and colleagues talk about "abnormal unconscious data", not "abnormal experience"), it is Bayesian-rational for the subject to adopt the hypothesis that the women in front of him is a stranger rather than the hypothesis that the woman is his wife. Although the imposter-hypothesis has lower prior probability than wife-hypothesis (i.e. imposter-hypothesis is less plausible than wife-hypothesis independently from the

experiential data), the former gets higher likelihood than the latter (i.e. the former predicts the experience better than the latter). And, they argue, it would be rational, overall, to adopt imposter-hypothesis rather than wife-hypothesis.

> The general point here is that if the stranger hypothesis explains the observed data much better than the wife hypothesis, the fact that the stranger hypothesis has a lower prior probability than the wife hypothesis can be offset in the calculation of posterior probabilities. And indeed it seems reasonable to suppose that this is precisely the situation with the subject suffering from Capgras delusion. The delusional hypothesis provides a much more convincing explanation of the highly unusual data than the nondelusional hypothesis; and this fact swamps the general implausibility of the delusional hypothesis. So if the subject with Capgras delusion unconsciously reasons in this way, he has up to this point committed no mistake of rationality on the Bayesian model. (Coltheart et al. 2000, 278)

The proposal by Coltheart and colleagues is more sophisticated than Maher's. In particular, Coltheart and colleagues posit a bias of ignoring counterevidence to the delusions at the stage of delusion maintenance. This bias is supposed to explain the difference between, for instance, the subjects with Capgras delusions and the subjects with the damage in ventromedial prefrontal cortex who are hypothesized as sharing the same abnormal experience with the former without having Capras delusion (Tranal et al. 1995). While ventromedial patients are able to remove imposter-hypothesis in the face of overwhelming counterevidence, including the testimony of friends and clinicians, Capras subjects, because of the bias, are unable to do so.

The view suggested by Maher and Colthert with colleagues, let us call it "MC view", fits very well with the idea that delusions are like *S1* with regard to production process. *S1* is produced in response to the perceptual experience of a bottle of beer in the fridge. Capgras delusions, similarly, are produced in response to the abnormal perceptual-affective experience of familiar faces. It is perfectly normal and rational to hold that there is a bottle of beer in the fridge when one perceives a bottle of beer in the fridge. It is also perfectly normal and rational, according to MC view, to hold that the woman is not his wife when he has the abnormal perceptual-affective experience of familiar faces.

MC view, however, is quite controversial. Against MC view, some argue that delusions are abnormal and irrational response to the abnormal experience. Stone and Young (1997), for instance, suggest that healthy belief formation involves a balance between two principles; the principle of doxastic conservatism, according to which we need to modify our beliefs as little as possible, and the principle of observational adequacy, according to which we need to accommodate new observation into our beliefs. Stone and Young propose the idea that delusions are produced as the result of abnormal bias toward observational adequacy. In other words, delusional subjects put too much emphasis on new observation, forgetting that the changes in beliefs system need to be minimal.

> The deficit to the perceptual system of those who suffer from Capgras delusion leads to an anomalous perceptual experience. In the face of this experience, there is the

> challenge of balancing observational adequacy with conservativism. In people who resolve this by forming the Capgras delusion, the balance goes too far in the direction of observational adequacy. (Stone & Young 1997, 350)

This claim was recently supported and sophisticated by McKay (2012). McKay argues, in response to Coltheart and colleagues, that delusion formation process should be modeled as a Bayesian-irrational updating of credence. According to McKay, Coltheart and colleagues fail to show that delusion formation process is Bayesian-rational because their assignment of prior probability to imposter hypothesis is unrealistically high. Given a more realistic assignment of prior probability, delusion formation process is likely to be a Bayesian-irrational process. McKay accepts the idea of Stone and Young, and mathematically models the bias for observational adequacy as the Bayesian-irrational bias of discounting prior probability.

> Following Stone and Young, I suggest that the second factor in delusion formation comprises a bias towards explanatory adequacy. Capgras delusion, on this story, results when brain damage or disruption causes the face recognition system to become disconnected from the autonomic nervous system, generating the anomalous data *o*, (Factor One). This disconnection occurs in conjunction with a bias towards explanatory adequacy (Factor Two), such that the individual updates beliefs as if ignoring the relevant prior probabilities. He thus adopts as a belief the hypothesis that the best explains the abnormal perceptual data available to him – the strange hypothesis h_s. (McKay 2012, 345-346)

Let us call this "SYM view". SYM view fits very well with the idea that delusions are like *S2*

in terms of production process. *S2* is produced by a malfunctioning mechanism, *P2*, in response to the perceptual experience (of marble tables) that doesn't make the content of *S2* ("the world is coming to an end") very probable. Similarily, according to SYM view, Capras delusion is produced by a malfunctioning mechanism in response to the abnormal perceptual-affective experience of faces that doesn't make the content of the delusion very probable. Stone and Young say that the healthy balance between the principle of doxastic conservatism and observational adequacy is compromised, implying that the mechanism involved in belief formation is malfunctioning. McKay hypothesizes that the bias of discounting prior probability is peculiar to delusional subjects, and healthy subjects and non-delusional subjects with ventromedial prefrontal damage do not have it, suggesting again that the mechanism involved in belief formation is malfunctioning.

So, if MC view is correct, delusions are like *S1*. If, on the other hand, SYM view is correct, delusions are like *S2*. Either way, it is likely that delusions have a right kind of producers, given the fact that both *S1* and *S2* have a right kind of producers. Certainly, there are many other views on delusion formation process, but most of them fit well either with *S1* model or *S2* model.

Here are some worries and my responses to them.

(1) *The Limitation of Empiricism?*: In the discussion above, I am relying on empiricist theories of delusion formation. Empiricist theories, however, might not be applicable to all

kinds of delusions. Especially, it is not clear how the empiricist theories are applicable to polythematic delusions (i.e. delusions with multiple themes). Indeed, the formation process of polythematic delusions is not very well understood in comparison to that of monothematic delusions.

This is a fair worry. Here are my answers. First, it is not the case, strictly speaking, that the above discussion needs to rely on the truth of empiricist theory of delusion formation. What the discussion needs is, basically, the truth of the clause (1) of the right hand side of TAF2 with regard to delusions. This doesn't require the truth of empiricism, in fact. For instance, the clause (1) would also be true about delusions if the delusions are produced not experientially out of experience, but inferentially out of other beliefs. Second, it is not impossible that polythematic delusions are also explained by empiricist approach (Coltheart et al. 2011; Coltheart 2013). First, it might be the case that polythematic delusions are the responses to multiple abnormal experience with multiple themes. Second, it might be the case that polythematic delusions are very abnormal responses, due to severe malfunction of belief-forming mechanism, to abnormal experiences. Third, it might be the case that polythematic delusions are the responses to ambiguous abnormal experience. The idea is that the ambiguity of the abnormal experience cause multiple responses, many of them end up being adopted by subjects due to malfunctioning belief-forming mechanism.

(2) *Abnormal Unconscious Data?*: Coltheart and colleagues argue that the "abnormal

experience" to which delusional subject respond to is, in fact, totally unconscious. They argue that "the abnormality in Capgras delusion, which prompts the exercise of abductive inference in an effort to generate a hypothesis to explain this abnormality, is not an abnormality of which the patient is aware" (Coltheart et al. 2010, 264). Since "unconscious abnormal experience" seems to be contradictory, they use the term "abnormal data" instead of "abnormal experience". It wouldn't be appropriate to call this proposal "empiricism" anymore. One might think that this causes a trouble for my proposal, because it is not clear that the unconscious data in response to which the belief that *p* is produced count as B"*p*"-appropriate inputs.

I don't think that this worry is very serious, because the argument by Coltheart and colleagues for their claim is not very strong.

> The problem we see here is that it is common to treat experiences as by definition conscious, so that nothing of which a person is not conscious can be called an "experience", and the term "conscious experience" is a tautology. If that is how the term "experience" is being used when one proposes that delusions are rational responses to abnormal experiences, then that theory of delusion is false for many kinds of delusions – the Capgras delusion, for example. The reason is that *people are not conscious of the activities of their autonomic nervous systems, and so a man would not be conscious of a failure of his autonomic nervous system to respond when he encountered his wife*. Hence, what happens here is not an abnormal experience, because it is not an experience. (Coltheart et al. 2010, 264)

The argument is that, in general, people are not conscious of the activities of autonomic

nervous system. Hence, the reduced activity of the system is not something Capgras subjects can be consciously aware of. However, the fact that Capgras subjects are not consciously aware of the reduced activities of autonomic nervous system is perfectly compatible with the possibility that the reduced activities lead to abnormal conscious experiences that are causally relevant in Capgras delusion formation process, for the same reason that we are not consciously aware of the activity of visual cortex is perfectly compatible with the possibility that the activity of visual cortex lead to visual experience of which we are aware (see Davies & Egan 2013).

5.1.2 Consumer of Delusion

While delusion production process has been extensively studied recently, delusion consumption process hasn't. So, the only thing we can do here is to speculate about it.

The most remarkable feature in the delusion consumption process is that delusions sometimes fail to cause appropriate non-verbal behavior. Followng Sass (1994), let us call this phenomenon "double-bookkeeping". For instance, a delusion with content "this is not my wife" fails to cause appropriate non-verbal behavior such as searching for "true wife", treating the woman as if she is not the real wife, and so on. The question that we are interested in here is whether or not this double-bookkeeping delusion has, nonetheless, a right kind of consumer. In other words, the question is whether or not the consumer of the

double-bookkeeping delusion has, nevertheless, the function of producing appropriate B"this woman is not my wife"-appropriate outputs. In my view, it is perfectly possible that the consumer has that function.

Bortolotti and Broome (Bortolotti 2011; Bortolotti & Broome 2012) hypothesize that double-bookkeeping is caused by the lack of relevant motivational factors. They presented a number of motivational explanations that are worth exploring. Double-bookkeeping happens because the relevant delusional subjects (1) think that they have no genuine control over their own behavior, which undermines the motivation to act, (2) fail to find the goal of action desirable, which undermines the motivation to act, (3) have general problem in producing self-willed action (Frith 1992), (4) the delusional subject do not find the goal of action attractive due to flattening of affect (Bleuler 1950), (5) find it hard to act in such a way as to enjoy the pleasant emotions associated with the goal of action (Foussias & Remington 2010), (6) feel hopeless and pessimistic about the probability of achieving their goals due to emotional disturbances. Those explanations suggest the following picture on double-bookkeeping. Double-bookkeeping is something analogous to the case where a coffee machine, having the function of producing coffee, fails to produce coffee due to the lack of water. There is nothing wrong with the coffee machine. Also, coffee beans have been put in. Nonetheless, the machine fails to produce coffee because water hasn't been supplied. Analogously, in double-bookkeeping, there is nothing wrong with the consumer mechanism

itself. Also, belief that the woman is not his wife has been “put in” to the mechanism. But, the mechanism fails to produce B”the woman is not my wife”-appropriate behavioral outputs because right kind of motivational states are not supplied. Let us call this hypothesis “missing motivation hypothesis.” Crucial point for our discussion is that, according to missing motivation hypothesis, the consumer mechanism has the function of producing B”this woman is not my wife”-appropriate behavioral outputs in response to the Capgras belief, although the mechanism fails to do this actually, for the same reason that the coffee machine has the function of producing coffee, although the machine fails to do it actually.

Here is another possible scenario. Another coffee machine might fail to produce coffee, even though coffee beans and water have been put in, due to its own malfunctioning. The machine is broken. Analogously, it might be the case that, in double-bookkeeping, the consumer mechanism fails to produce B”the woman is not his wife”-appropriate behavioral outputs, even though relevant beliefs and desires have been “put in”, due to its own malfunctioning. This is an alternative hypothesis about double-bookkeeping. Let us call this “malfunctioning consumer hypothesis.” Again, crucial point for our discussion is that, according to malfunctioning consumer hypothesis, the consumer mechanism has the function of producing B”the woman is not my wife”-appropriate behavioral outputs in response to the Capgras belief for the same reason that the coffee machine that fails to produce coffee due to its own malfunctioning still has the function of producing coffee.

Both missing motivation hypothesis and malfunctioning consumer hypothesis imply that the consumer mechanism of delusions has the function of producing appropriate behavioral outputs in response to delusions. But, there would be some alternative hypotheses too. A fruit juice maker doesn't produce coffee. It doesn't produce coffee not because the right kind of ingredients are not supplied, nor because it is broken. It is just not supposed to produce coffee in the first place. Presumably, double-bookkeeping is similar to this. In that case, the consumer mechanism doesn't have the function of producing B"the woman is not my wife"-appropriate behavioral outputs in response to the delusion that the woman is not his wife for the same reason that the fruit juice maker that fails to produce coffee doesn't have the function of producing it.

The view of Currie and collaborators (Currie 2000; Currie & Jureidini 2001; Currie & Ravenscroft 2002) might be understood in this way. According to their view, double-bookkeeping occurs because delusions are imaginations, and imaginations, in general, do not have belief-like causal impact on non-verbal behavior. Presumably, it would be their view that imaginations do not have belief-like causal impact on non-verbal behavior because they are not consumed by the mechanism whose function is to produce belief-like behavioral outputs. For instance, the imagination that this woman is not my wife is not consumed by the mechanism whose function is to produce B"this woman is not his wife"-appropriate behavioral outputs in response to the imagination. Again, Frankish's view

(2009, 2012) can be read in a similar way. According to Frankish, double-bookkeeping occurs because delusions are "level-2 beliefs", and "level-2 beliefs", in general, have only weak impact on non-verbal behavior. According to Frankish, "level-2 beliefs" are conscious, controlled, binary, functional states, while "level-1 beliefs" are unconscious, passive, graded, dispositional states. It might be Frankish's view that "level-2 beliefs", in general, have only weak impact on non-verbal behavior because they are not consumed by the mechanism whose function is to produce appropriate non-verbal behavior.

I don't rule out these hypotheses. After all, all of these are empirical issues about which we don't know so much yet. My point is just hat missing motivational hypothesis and malfunctioning consumer hypothesis are among the available options and they imply that double-bookkeeping delusions have a right kind of consumers.

Summing up, I have argued, in this section, that there is a real chance (not mere logical possibility) that delusions have a right kind of producers and a right kind of consumers. In other words, there is a real chance that the delusion with content "p" is produced by the mechanism with function of producing the state in response to B"*p*"-appropriate inputs, and consumed by the mechanism with the function of producing B"*p*"-appropriate outputs in response to the state. This means that there is a real chance that the delusion is a belief according to TAF2. Future empirical researches on delusion production and delusion consumption processes will improve our understanding of this issue.

5.1.3 Delusions as In-Between States?

I close this chapter by briefly discussing the view that delusions are some kinds of in-between states. Presumably, this is the most popular view on the nature of delusion among philosophers. The reason why it is popular is quite straightforward. The causal roles of delusions are significantly different from paradigmatic belief-like causal roles, but they are also significantly different from other sorts of paradigmatic causal roles, such as paradigmatic imagination-like causal roles.

> Classifying delusions as straightforward, paradigmatic cases of belief is problematic because it predicts that delusions ought not to display the sorts of circumscription and evidence independence that they in fact display. Classifying them as straightforward, paradigmatic cases of imagination is problematic because it predicts that they should display *more* circumscription and evidence-independence than they in fact display. What would be nice would be to be able to say that the attitude is something inbetween paradigmatic belief and paradigmatic imagination – that delusional subjects are in states that play a role in their cognitive economies that is in some respects like of a standard-issue, stereotypical belief that P, and in other respects like that of a standard-issue stereotypical imagining that P. (Egan 2009, 268)

> […] delusions are considered as a class of states do not fit easily into rigid categories of either belief or imagination. While delusions generally have a significant power to command attention and generate affect, they vary a great deal in the extent to which they are acted upon and given credence by their possessors. In that case it may be that cognitive state do not sort themselves neatly into categorically distinct classes we should label 'beliefs' and 'imaginings', but that these categories represent vague clustering in a space that encompasses a continuum of states for some of which we have no commonly accepted labels. (Currie & Jones 2006, 312)

> Is it possible, then, that cases of delusions are, at least sometimes (when the functional role or dispositional profile is weird enough), cases in an in-betweenish gray zone – not quite belief and not quite failure to believe? Bortolotti raises this possibility in her discussion on "sliding scale" approach on pp. 20-21 [Bortolotti 2010], but offers only slender reason to dismiss it. She suggests that the sliding scale approach makes it not straightforward how to answer questions about whether an action is intentional or not, which complicates ethical and policy applications. In reply to this, of course, the friend of the sliding scale might suggest that in many cases of delusion it *shouldn't* be straightforward to assess intentionality, and the ethical and policy applications *are* complicated, so that a philosophical approach that renders these matters straightforward is misleadingly simplistic. (Schwitzgebel 2012, 15)

According to Egan, delusions are the states that are in-between beliefs and imaginations. Egan calls them "bimaginations". Currie and Jones hold a pretty similar view. According to Schwitzgebel, (problematic) delusions are not beliefs but not non-beliefs. They are borderline cases between belief and non-beliefs. Schwitzgebel calls them "in-between beliefs."

Bayne is a consistent critic of bimagination proposal and other in-between proposals (Bayne 2010, Bayne & Hattiangadi 2013). His main criticism is based upon his normativism. According to normativism, beliefs are the states that are subject to the norms of belief, such as, such as norms of truth or consistency. Imaginations are, on the other hand, the states that are subject to the norms of imagination, such as the norm of internal consistency among imaginations. If there are such things as bimaginations, according to normativism, then they have to be the states that are subject to the norms of bimagination. But, it is not clear

what “the norms of bimagination” amount to.

> Functional roles seem to be continuous, in the sense that it seems possible that between any two of the propositional attitudes recognized by folk psychology there may be other functional roles that could—at least in principle—demarcate intermediate propositional attitude types, such as bimagination. By contrast, it is rather less clear whether normative conceptions of the mind have quite the same room for intermediate propositional attitude kinds, for norms may not be ‘continuous’ in the way that functional roles perhaps are. (Bayne 2010, 334)

I don’t share Bayne’s skepticism because I don’t share the theory of belief upon which his skepticism is based. But, TAF2 supports another skepticism about bimagination proposal and other in-between proposals. For the sake of simplicity, let us talk about TAF here, instead of TAF2 (the same point can be made in terms of TAF2 actually). According to TAF, beliefs are the states that have the function of playing belief-like causal roles. Similarly, imaginations are the states that have the function of playing imagination-like causal roles. If there are such things as bimaginations, according to TAF, they are the states that have the function of playing bimagination-like causal roles. The problem about bimagination proposal is that it is not clear that we, human being, have the states that have the function of playing bimagination-like causal roles. We do have some states that actually play bimagination-like causal roles and, probably, problematic delusions are the examples. But, it doesn’t mean that they have the function of playing bimagination-like causal roles any

more than the fact that there are some kidneys that fail to filter metabolic wastes from blood means that these kidneys have the function of failing to do it. To say that we have some states with the function of playing bimagination-like causal roles is to say that we have some states that were selected for playing bimagination-like causal roles (or we have some mechanisms that were selected for producing and consuming some states in bimagination-like fashion). But, presumably, we do not have such states (or mechanisms), because having the states with bimagination-like causal roles do not seem to be evolutionary beneficial (or having the mechanisms that produce and consume states in bimagination-like fashion doesn't seem to be evolutionary beneficial).

This creates an important difference between belief or imagination on one hand and bimagination on the other. Having mental states with belief-like causal roles would be evolutionary beneficial. For instance, having the mental states that are sensitive to evidence and having impact on non-verbal behavior is enormously evolutionary beneficial. These state serve as "the maps by which we steer" (Ramsey 1931). Having states with imagination-like causal roles would also be evolutionary beneficial. They can be used for the simulation of non-actual scenario or perhaps the mind of other person in virtue of its insensitivity to evidence and its behavioral inertness. On the other hand, what is the evolutionary point of having the states with bimagination-like causal roles? What is the evolutionary point of having the states that are not sensitive to evidence as beliefs but not as

insensitive to evidence as imaginations? What kinds of evolutionary benefit is given by having the states that do not have as much impact on behavior, but not as behaviorally inert as imagination?

5.2 Pathological Nature of Delusions

5.2.1 Why Is Delusion Pathological?

Delusion is a pathological mental state, assuming that the anti-psychiatric idea that there is no mental disorder (e.g. Szasz (1960)) is false. It is a symptom of various disorders. Here is a question. What makes delusions pathological? In virtue of what are delusions pathological?

There are some potential answers to the question.

(1) One might think that delusions are pathological because their contents are too strange. The problem with this view is, among others, that it is not the case that all delusions have remarkably strange contents. Certainly, Capgras delusion (e.g. "my wife is replaced by an imposter."), Cotard delusion (e.g. "I am dead."), anosognosia (e.g. "this hand doesn't belong to me"), thought insertion (e.g. "the thoughts in my head are not mine but someone else's".), etc. are strange. But, other delusions such as erotomania (e.g. "a famous actor wants to marry me"), persecutory delusion (e.g. "some colleagues in my office are very jealous at me, and they are trying to prevent me from bein promoted") or grandiose delusion

(e.g. “I have a special power to save the world from evil”) tend not to be very strange. In addition, having strange content is not sufficient for a mental state to be pathological. For instance, we can imagine pretty strange things without losing mental health (e.g. “a man realized, when he woke up, that he had turned into a giant bug during the night”). Furthermore, philosophers seriously believe quite strange things without losing mental health (e.g. “for any objects, however arbitrary they are chosen, there is a further object that is composed by them” (unrestricted composition), “there are facts about the boundaries of a vague predicate which we can never discover” (epistemicism about vague predicates)) (for the comparison between delusional beliefs and philosophical beliefs, see Reimer 2010).

(2) Another answer would be that delusions are pathological because they are too irrational. The problem with this answer is, among others, that given empiricist theories of delusion formation according to which delusions are the responses to abnormal experiences, it is not clear that delusions are *too* irrational. As I already mentioned, Coltheart et al. (2010) argue that delusion formation process is actually quite rational according to Bayesian standard. Certainly, McKay (2012) argues, in response to Coltheart et al., that delusion is the product of Bayesian-irrational process with the bias of discounting prior probability ratio. But, even if McKay is correct, still, it is not clear that delusion formation process is significantly more irrational in comparison to the normal belief formation process, given the fact that similar biases in probabilistic reasoning are widely seen in normal populations. For

example, the famous study by Kahneman and Tversky on base-rate neglect (1973) showed that normal individuals also have the bias of neglecting prior probability ratio. Hemsley and Garety (1997) even suggested that the reasoning style of delusional subjects is actually more rational than that of healthy controls.

(3) Bortolotti (2010) suggests that delusions are pathological because they have significant negative impact on well-being. Delusions certainly negatively affect the well-being of delusional subjects in all sorts of ways. They cause psychological distress, prevent subjects from being fully involved in social life and personal relationship, prevent them from developing their abilities and skills, increase the risk of suicidal acts, and so on. Presumably, there is an important link between negative impact on well-being and pathology. For instance, it is plausible to think that negative impact on well-being is necessary for pathology. As Wakefield pointed out, "disorder is in certain respects a practical concept that is supposed to pick out only conditions that are undesirable and grounds for social concern" (Wakefield 1992b, 237). But, the problem about this answer is that negative impact on well being would not be sufficient for pathology, even if it is necessary. For instance, pain is not a pathological mental state, even though pain has significant negative impact on well-being. Of course, pain can be caused by various kinds of pathological conditions, but pain itself is not pathological. Rather, the lack of (appropriate) pain is pathological. Congential analgesia, the condition in which patients are unable to feel pains

caused by tissue damages, is clearly a pathological condition.

(4) Murphy (2012) proposes the idea that delusions are pathological because they defy all explanations by folk understanding of how mind works. Delusions are pathological because we can't give folk explanation of delusion formation and maintenance. Two problems. First, it is not clear that all kinds of delusions defy folk explanations. For instance, it is not clear that the delusions with clear motivational background defy folk explanations. For instance, Butler (2000) reported the case of B.X. who had sustained severe head injuries in a car accident. One year after his injury, he developed a delusional system that revolved around the continuing fidelity of his partner, N., who had in fact severed all contact with him soon after his accident. A pretty straightforward folk explanation of this seems to be that B.X. believes in the fidelity of N. because he desperately wants it to be the case.

> In the early April 1997 his delusional system began to break up and he accepted that he and N. were not married, although he continued to insist they would marry soon after his discharge from inpatient rehabilitation. Following this shift he became determined to personally contact N., and through April and May he telephoned her repeatedly, leaving messages on her answering machine that he "loved [her]" and "would be home soon." Facility staff also made repeated attempts to contact N., to discuss the extent of B.X.'s cognitive impairments, and to request her help in facilitating the dissolution of his erotomania. N. did not respond to any of these contacts until she wrote B.X. a long letter at the end of June. In it she starkly reiterated the permanence of their separation and her unwillingness ot see or speak to him. The letter concluded: "Thinking back, I feel nothing but anger and disgust. I would appreciate it if you would not contact me again." On receipt of this letter B.X. broke into tears, desperately insisting 3 hours later that the letter "had been a mistake" and that his relationship with N. was "back to normal." (Butler 2010, 88)

Second, lack of folk explanation wouldn't be sufficient for something to be pathological. As Churchland (1981) pointed out, there are so many psychological phenomena of which there is no interesting folk explanations.

> As examples of central and important mental phenomena that remain largely or wholly mysterious within the framework of FP, consider the nature and dynamics of mental illness, the faculty of creative imagination, or the ground of intelligence differences between individual. Consider our utter ignorance of the nature and psychological functions of sleep, that curious state in which a third of one's life is spent. Reflect on the common ability to catch an outfield fly ball on the run, or hit a moving car with a snowball. Consider the internal construction of a 3-D visual image from subtle differences in the 2-D array of stimulations in our respective retinas. Consider the rich variety of perceptual illusions, visual capacity for relevant retrieval. On these and many other mental phenomena, FP sheds negligible light. (Churchland 1981, 73)

Churchland agrees with Murphy in that mental disorder is not explained by folk psychology. But, he wouldn't agree with Murphy's claim that defying folk psychological explanation makes things pathological. After all, folk psychology doesn't explain very many things in the first place. And, many of the phenomena unexplained by folk psychology are clearly non-pathological ones. (For the sake of fairness, I need to mention the fact that Murphy's "folk understanding" is broader than "folk psychology" in narrow sense. It also includes, according to Murphy, the beliefs and expectations about the role of hot cognition and

personal interest in fixing beliefs, and the role of culture in shaping people's assumption about what counts as legitimate evidence. Still, this isn't very helpful. It is still the case that many of Chuchland's examples defy the folk understanding or folk explanation in this broader sense.)

Here is my diagnosis of the all of the problems above. These answers to the question about the pathology of delusion are detached from the considerations about what, in general, makes something pathological. In order to explain why *X* is pathological, first, we need to have an account of what makes things pathological in general and, then, show that *X* has the feature as well. But, all of these answers skip the first step and simply point out some remarkable negative features of delusions. This invites all sorts of counterexamples as well as theoretical difficulties.

So, a satisfactory explanation of the pathological nature of delusion needs to be given against the background of the view on what, in general, makes things pathological. Wakefield (1992a, 1992b, 1999a, 1999b, 2011) presented a general view of disorder, which is quite influential and, in my view, is the most promising among others. It is called "Harmful Dysfunction Analysis" or "HDA" for short. According to HDA, disorders are "harmful malfunctions" or "harmful dysfunctions." "Harmful" part is a value component of a disorder, and "malfunction" is a factual component of it. Because of this combination of value and factual component, HDA is described as a "hybrid" analysis of disorder. "Malfunction" is

etiologically defined. In other words, something malfunctions when it fails to perform the etiological function it has. (It turns out that this is strictly speaking imprecise. I will come back to this shortly.) For instance, a heart malfunctions when it fails to pump blood, a kidney malfunctions when it fails to filter metabolic wastes from blood, a corpus callosum malfunctions when it fails to facilitate interhemispheric communications, and so on. Being "harmful" means having negative impact on well-being. The harmful condition is required to deal with some cases where malfunctioning doesn't constitute disorder because of the lack of significant negative impact on well-being. For instance, harmless albinism and fused toe involve some kinds of etiological malfunctions, but they tend not to be regarded as disorders.

Now, given HDA, we can explain the pathological nature of delusions in the following way. Delusions are pathological because (1) they involve some kinds of malfunctionings, (2) they are harmful and, (3) harmful malfunctioning implies disorder or pathology.

In the next section, I will discuss (1) with details. I take (2) to be uncontroversial. Here are some remarks on (3).

HDA is quite influential, but it also attracts varieties of objections. Still, I don't think that these objections threaten my explanation of the pathological nature of delusion. First, most objections are irrelevant. They are irrelevant because these objections are aiming at refuting the necessity of harmful malfunction for disorders, rather than the sufficiency of it. On the other hand, my explanation says that delusions are pathological because they involve

harmful malfunctioning. Obviously, this explanation only needs the sufficiency of harmful malfunction for disorder, not the necessity of it. Second, Wakefield responded to many of these objections, and his responses are quite persuasive. Here is a quick summary of some notable objections and his replies.

- *Objection*: Viral infection is a clear case of disorder. But, the symptoms of viral infection (e.g. fever, cough, sneezing) are very often biological defenses, not malfunctions. But, then, malfunctioning is not necessary for disorder (Tengland 2001).
- *Answer*: Again, malfunctioning can occur in lower levels, such as tissue level or cell level. In cell level, viral infection involves malfunctioning. For instance, a virus enters a cell and reproduces itself within the cell and spreads to other cells by causing malfunctioning of the cellular machinery (Wakefield 2011).
- *Objection*: Vestigial organs such as appendix do not have any functions and, thus, they can't malfunction. But, they can be disordered. For instance, Appendicitis is clearly a disorder. Thus, harmful malfunctioning is not necessary for disorder (Murphy & Woolfolk 2000).
- *Answer*: Since etiological functions are ascribed to organisms in many different levels, malfunctioning happens in many different levels. In particular,

malfunctioning can occur in lower levels (than organ level), such as tissue level or cell level. This is the reason why Appendicitis is a disorder. Appendicitis causes malfunctions in lower levels and, hence, it counts as a disorder (Wakefield 2000).

- *Objection*: Evolutionary byproducts (Gould & Lewontin 1979) do not have any functions and, thus, they can't malfunction. But, it might be that some evolutionary byproducts are disordered. Then, harmful malfunctioning might not be necessary for disorder (Murphy & Woolfolk 2000).
- *Answer*: In some cases, the abnormality in evolutionary byproducts is the sign of some malfunctioning. In other cases, it isn't. The former case constitutes a disorder. The latter case doesn't. An example of the latter case is atheism. Religious beliefs are, presumably, the evolutionary by-products of some mechanisms that evolved for other purposes. Atheism (= lacking religious belief) is not a disorder, because atheism is not the sign of malfunctioning mechanisms. An example of the latter is dyslexia. Reading is also an evolutionary by-product of some mechanisms that evolved for other purposes. Unlike atheism, however, dyslexia is a disorder, because it is a sign of some malfunctioning mechanisms (Wakefield 2000).
- *Objection*: It might be that some disorders are caused not by malfunctioning, but by the mismatch between current environment and the environment where something was designed. In that case, malfunctioning is not necessary for disorder (Murphy &

Woolfolk 2010).

- *Answer*: The mismatch between current environment and designed environment doesn't constitute disorders. For instance, "we are designed to have a taste for fat and sugar, so even though these taste preferences are harmful to us in our current calorie-rich environment, they are not considered disorders. We believe that men are designed to be more aggressive than women, so even though the assumedly designed levels of male aggressiveness are arguably disastrous in the modern world, we do not consider them to be disorders. We are likely to be designed to desire multiple sexual relationships, so even though infidelity is highly disvalued in our society, infidelitous desires are not in themselves considered a disorder" (Wakefield 2010, 258).

- *Objection*: It can happen that some species are dying exactly because of the proper functioning of some of their features. For instance, it can happen that a type of bear is dying because of its warm fur together with global warming. HDA implies that this is not the case of disorder, because the fur is properly functioning. But, it is counterintuitive to say that the bears are in healthy condition (Nordenfelt 2007).

- *Answer*: The bears are not disordered. It is not counterintuitive to say that. They are not disordered but, rather, in an unlucky condition. "Imagine a moth species whose natural coloration is white because it is biologically adapted to be camouflaged

against the white bark on the trees in its environment to evade predators. Now, imagine that for either natural or environmental reasons (e.g., a new factory opens nearby and emits soot) the bark on the local trees rapidly turns brown, and the whiteness of the moth becomes a fatal signal to predators. The moth's coloration is killing them, and perhaps killing the whole species, due to the change in the environment. Yet one is not inclined to construe this misfortune as a vast epidemic of disorder among the moths in which there is a dysfunction of their coloration. Rather, the moths are simply unlucky" (Wakefield 2011, 168).

5.2.2 Delusions Involve Malfunctioning

Now, let us turn the idea that delusions involve some kinds of malfunctioning.

Strictly speaking, delusions themselves can't malfunction for the reason that I already mentioned. Mental states do not have any functions because they are not the products of natural selection. Functions are rather attributed to mechanisms. Thus, when I say that delusions involve malfunctioning, I am not saying that delusions themselves are malfunctioning, but that that there are some mechanisms, directly or indirectly related to delusions, that are malfunctioning.

There are some candidate delusion-related mechanisms that are malfunctioning. First, empiricist theories of delusion formation assume that delusions are formed in response to

some abnormal experiences. In this view, those mechanisms that are responsible for the production of abnormal experiences are good candidates for malfunctioning mechanisms. In Capgras delusion, for instance, it is hypothesized, with good empirical supports, that the abnormal perceptual-affective experience in response to which Capgras delusion is formed is the result of the disrupted connection between face recognition system and autonomic nervous system. Second, some kinds of malfunctioning mechanism might be involved in the process of delusion acceptance and maintenance. For instance, SYM view would posit some malfunctioning mechanisms that are responsible for the bias toward observational adequacy. Third, there might be some malfunctioning consumers of delusions, especially in the case of double-bookkeeping. According to malfunctioning consumer hypothesis of double-bookkeeping, the consumer of double-bookkeeping delusions is malfunctioning. On the other hand, according to missing motivation hypothesis, the consumer is not malfunctioning. Still, this hypothesis would posit some kinds of malfunctioning mechanisms that are responsible for the absence of appropriate motivational factors.

Here are some potential objections.

(1) *Psychological Defense Objection*: One might think that delusion-related mechanisms are successfully performing a function, namely, psychological defense function. In the case of B.X. that I mentioned earlier, for instance, it is tempting to say that his delusion about the fidelity of N. plays psychological defense functions. Indeed, this somewhat Freudian idea

that some delusions are psychological defense is becoming popular recently (Bental & Kaney 1996; Kinderman & Bentall 1996; McKay & Cipolotti 2007; McKay et al. 2005; McKay et al. 2007). For instance, it has been suggested that persecutory delusions are the products of externalizing attribution bias; they attributes negative events to other agents rather than themselves. This externalizing bias can easily be construed defensively; the bias is playing the role of defending self-esteem, for instance.

One version of this would be the idea that the defensive delusions are perfectly functional. There is nothing wrong with belief-related mechanisms in those cases. Another version, which is quite interesting, is the idea that defensive delusions are what McKay and Dennett (2009) calls "doxastic shear pins". According to this proposal, defensive delusions are malfunctioning in a sense, but they are *designed to be malfunctioning*. A shear pins is a metal pin installed in complex mechanistic systems, and it is designed to break in certain circumstances so as to protect other, more expensive parts of the system. Dennett and McKay discuss the possibility that a defensive function is a doxastic analogue of shear pin.

> 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 (analogues to the mechanical overload that breads 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 & Dennett 2009, 502)

In any case, however, the psychological defense objection is not very persuasive. I don't rule out the idea that some delusions are playing psychological defense function. The problem of this objection is, rather, that the fact that some delusions are playing a psychological defense function doesn't imply that they are successfully playing a *biological, etiological function*. Crucial point is that, as Dennett and McKay correctly pointed out, a psychological defense function might not be an etiological, biological function, because there is no guarantee that the psychological comfort achieved by psychological defense leads to reproductive success. Stich famously argued that, "natural selection does not care about truth; it cares only about reproductive success" (Stich 1990, 62). Similarly, he could also have said that natural selection only cares about reproductive success, not individual psychological comfort.

Nesse (1998) makes this point very powerfully. According to Nesse, negative emotions, such as fear or anxiety, have etiological, biological functions, such as the function of avoiding or preparing for danger or threats. Certainly, negative emotions are psychologically disturbing. But, it doesn't matter from a biological point of view, because "natural selection shaped the regulation mechanisms for maximal reproductive success, not for peace and

happiness" (Nesse 1998, 401). Presumably, we can even say that a negative emotion serve its biological function because of the psychological disturbances it causes, in the same way that a pain serves its biological function (of protecting bodily parts) because of the painful feeling it causes (or, it is). Thus, negative emotions are psychologically disturbing, but biologically functional. Nesse goes on to point out that some psychologically positive states are, in fact, biologically malfunctional. He argues,

> […] there are two kinds of anxiety disorder, those in which anxiety is expressed excessively and those in which in is expressed insufficiently. Can it be abnormal to have too little anxiety? Yes, in exactly the same way that there can be insufficient pain or an insufficient immune response, and which equally devastating consequences. Patients with insufficient anxiety do not, however, come to therapists to have their anxiety increased, even when their incaution actions result in accidents, losses of jobs and relationships and drug addiction. And pharmaceutical companies have yet to maket an agent designed to increase deficient anxiety to normal levels (although an agent that increases caution without increasing subjective anxiety would have huge market, with probable ability to prevent relapse into drug addiction, among other benefits). (ibid., 402)

Since anxiety has biological function, insufficient anxiety, not just excessive anxiety, is malfunctional, according to Nesse. People with insufficient anxiety would be psychologically more positive than other normal people. Nonetheless, they are biologically malfunctional and, presumably, they should be medically treated.

In short, even if it is the case that some delusions are playing psychological defensive

roles, it doesn't imply that they are correctly performing some etiological functions.

(2) *Adaptation Objection*: A better objection is that at least, some delusions are playing some genuine etiological functions. A good candidate of this is delusional jealousy or morbid jealousy. It is pretty likely that jealousy (or jealousy-related mechanisms), in general, has the function of protecting valuable relationship from partial or total loss (Buss et al. 1992, Daly et al. 1982). In morbid jealousy, the subjects irrationally believe the infidelity of their romantic partners without good evidence, and they often make unwarranted or unverified accusation about the infidelity. Morbid jealousy has some important characteristics in common with normal jealousy. For instance, men diagnosed with morbid jealousy are especially upset about a partner's sexual infidelity, whereas women diagnosed with morbid jealousy are especially upset about a partner's emotional infidelity, which is consistent with the pattern that is seen in normal jealousy (Easton et al, 2007). This seems to suggest that morbid jealousy is an exaggerated version of normal jealousy, where jealousy is activated in the cases where it is not really appropriate.

One might think that since jealousy is activated when it is not really appropriate in morbid jealousy, there must be some kinds of malfunctioning in jealousy-related mechanisms. But, this might not be the case. It may well be that jealousy is, by design, activated more often than it really needs to be. Smoke detectors are designed to give many false positives. In other words, they are designed to be activated more often than they really

need to be. This is because the cost of false positives are pretty low (i.e. some unnecessary evacuations), while the cost of false negatives is devastating (i.e. the building will be burnt down). This "smoke detector principle" (Nesse 2001) may well be true about jealousy. The cost of false positives seems to be not very high (i.e. the partner will be annoyed), while the cost of false negatives seem to be quite high (i.e. the partner leaves the subject). Thus, it may well be that jealousy is, by design, activated more often than it is really appropriate. Indeed, Nesse argues, based on the same principle, that anxiety is, by design, activated more often that it is really appropriate, and Haselton and Buss (2000) argue that men's perception of the sexual interest of woman is, by design, activated more often than it is really appropriate.

It is quite important to note that this objection is different from psychological defense objection. After all, jealousy never serves the role of defending psychological comfort. It rather disturbs it. Its role is, rather, to keep important relationship for reproductive success, which is clearly biological.

This objection is a good one. Thus, I will not claim, without sufficient empirical grounds, that all delusions involve some malfunctioning. It might turn out that, as suggested in this objection, that some delusions do not involve any etiological malfunctioning for the same reason that somewhat overactive smoke detectors are not malfunctioning.

Here is, however, a worry. What if, then, it turns out that morbid jealousy is not malfunctional, but a part of biological design? Does that mean that morbid jealousy is not

pathological? Doesn't that cause a problem? For, morbid jealousy seems to be pathological. For instance, some medical interventions are surely appropriate for morbid jealousy.

My answer to this is as follows. First, as I said earlier, I only need to be committed to the sufficiency of harmful malfunction for pathology, not necessity. Thus, it is still an open option to say that malfunction is not necessary for pathology. In taking this option, I am not forced to the allegedly problematic conclusion that morbid jealousy is not pathological. But, since I am also sympathetic to necessity claim (I think that Wakefied defended the necessity claim successfully), I am not very attracted to this reply. Second, I don't think it obvious that morbid jealousy is pathological when it turns out that it is a part of biological design. Certainly, some medical intervention would be appropriate for morbid jealousy. But, this doesn't imply that morbid jealousy is pathological. For instance, anesthesia is appropriate before a surgery, even though the pain that the surgery causes is not pathological in itself. The pain is rather normally performing its function of defending tissues. In some cases, medical treatment is appropriate for fever, even though the fever is not pathological in itself. The fever is rather a designed response to, say, viral infection. In some cases, morning sickness requires medical treatment even though it is not pathological in itself. Morning sickness is, presumably, a designed process for defending mother and fetus from potentially harmful substances. Medical intervention is appropriate in those cases not because these cases involve pathological conditions, but because they involve unnecessary suffering. The

same thing could be true about morbid jealousy. Medical intervention is appropriate for morbid jealousy not because morbid jealousy is a pathological condition, but it causes unnecessary (psychological and social) problems. Indeed, smoke detector principle predicts that most activations of jealousy are not necessary for the same reasons that most activations of smoke detectors are, strictly speaking, not necessary. It is very likely that the morbid jealousy is an instance of unnecessary activation of jealousy.

5.2.3 A Puzzle about Teleosemantics

Here is a puzzle. I argued that delusions are pathological because they involve harmful malfunctioning, and harmful malfunctioning is sufficient for disorder. Now, it is sometimes said that the fundamental idea of teleosemantics is that misrepresentations (e.g. non-veridical perceptions, false beliefs) involve some kinds of malfunctioning.

> The basic idea behind teleological theories of content is that this normative notion – and its distinction between proper functioning and malfunctioning – might somehow underwrite the normative notion of content – and its distinction between representation and misrepresentation. (Neander 1995, 112)

> Much of the original appeal of teleosemantics was its ability to employ teleo-functional notions of purpose in order to deal with apparently normative aspects of semantic phenomena. In particular, the biological notion of failure to perform a proper function was used to attack the problem of misrepresentation, which had caused a lot of trouble for information-based theories. (Godfrey-Smith 2006, 62)

But, then, given HDA, it follows that misrepresentations are pathological when they are harmful. But, this is clearly false. Some individuals might falsely believe that they are not as talented as their colleagues in the fields they are working, and the false beliefs might have some negative impacts on their well-being (e.g. psychological stress, anxiety, loss of confidence). But, nobody would say that these individuals are mentally disordered. Certainly, those false beliefs might sometimes lead to pathological level of depression. Still, it is clear that the false beliefs themselves are not pathological. Then, it looks as though I need to either reject teleosemantics or admit that there is something wrong with my explanation of the pathological nature of delusion.

This, I think, is too quick. In order to find what is actually going on here, we need to look more carefully at the idea that misrepresentations involve malfunctionning.

Think about the famous example of anaerobic bacteria.

> Some marine bacteria have internal magnets, magnetosomes, that function like compass needles, aligning themselves (and, as a result, the bacterium) parallel to the Earth's magnetic field. Since the magnetic lines incline downward (toward geomagnetic north) in the northern hemisphere, bacteria in the northern hemisphere, oriented by their internal megnetosomes, propel themselves toward geomagnetic north. Since these organisms are capable of living only in the absence of oxygen, and since movement toward geomagnetic north will take northern bacteria away from the oxygen-rich and therefore toxic surface water and toward the comparatively oxygen-free sediment at the bottom, it is not unreasonable to speculate, as Blakemore and Frankel do, that the function of this primitive sensory system is to indicate the

whereabouts of benign (i.e. anaerobic) environments. (Dretske 1986, 63)

Now, suppose that I use a bar magnet to lead a bacteria upward and, consequently, the bacteria dies because of the exposure to oxygen-rich surface water. Arguably, the magnetosome of this poor bacteria is not malfunctioning in this case. It is just unlucky. The crucial question here is whether or not the magnetosome is misrepresenting according to teleosemantics. If it is misrepresenting, then it turns out that teleosemantics is not committed to the view that misrepresentation necessarily involves malfunction.

Is the magnetosome misrepresenting, then? This depends on what the magnetosme represents. If it represents magnetic north, then it is not misrepresenting. After all, it successfully indicates magnetic north. If, on the other hand, it represents oxygen-free sediment, then it is misrepresenting. After all, it fails to indicate oxygen-free sediment.

Different teleosemanticists give different answers. Here, I discuss two teleosemanticists who give opposite answers; Millikan and Neander. Millikan argues that the magnetosome is misrepresenting because it represent oxygen-free sediment, not magnetic north. In Millikan's view, what a state of magnetosome represents is determined by what the state needs to correspond to if the consumer of the state is to perform its function in its normal way. And for the successful performance of the consumer (i.e. motor mechanism) in normal way, what the state needs to correspond to is oxygen-free sediment, not geomagnetic north.

After all, the purpose of the motor mechanism is to lead the bacteria to oxygen-free sediment. (The mechanism was selected for leading ancestral bacteria to oxygen-free sediment.) Thus, it looks as though it is not Millikan's view that misrepresentation always involves malfunction. The state of magnetosome misrepresents, but it is not malfunctioning.

But, the things are a little more complicated. As Millikan pointed out, there is a certain sense in which we can say that the magnetosome fails to perform its function.

> […] Dretske is right that the magnetosome that directs that bacterium in the wrong direction because someone holds a bar magnet overhead is not broken or malfunctioning. In that sense, it is functioning perfectly properly. But it doesn't mean that it is succeeding in performing all of its functions, any more than a perfectly functional coffeemaker is performing its function when no one has put any coffee in it. Very often things fail to perform their functions, not because they are damaged, but because the conditions they are in are not their normal operating conditions. (Millikan 2004, 83)

Millikan's view seems to be that that the magnetosome fails to perform a function in the same way that coffee maker fails to perform its function (of making coffee) when no coffee bean is put in. It is important to note that Millikan carefully distinguishes "malfunctioning" due to damage from "the failure of performing a function" due to unlucky condition. This distinction is much clearer in the following quote.

> When a devise does not perform some proper function only because the necessary background condition are absent, we do not consider that to be a malfunction – though

> it might, of course, be due to malfunction of some other devise whose job was to put the necessary conditions in place. Malfunction results only from abnormalities in the constitution of the devise itself. (Rylder et al. 2012, reply to Neander)

For the sake of avoiding confusion, I will use the term "misfunction" as referring to the failure of performing function due to unlucky situation or condition, distinguishing it from malfunction due to intrinsic damage or breakdown. In this terminology, the coffee machine is misfunctioning because it is in an unlucky situation where nobody puts coffee bean in it. The magnetosome is misfunctioning because it is in an unlucky situation where it is fooled by my bar magnet.

Think about the following case by Neander.

> Suppose a woman's Fallopian tubes are blocked, and as a result she is unable to conceive; sperm cannot reach the ova to fertilise them. Now here are two ways of describing her condition: (1) we could say that her reproductive system is malfunctioning, and further that all of its parts are malfunctioning, because none of them can achieve their higher 'purpose' of bringing a child to term and replicating her genes, or (2), we could say that her reproductive system is malfunctioning, but not all of its parts are malfunctioning, only her Fallopian tubes. In other words, on this second option, her ovaries don't malfunction (just) because her reproductive system malfunction because her Fallopian tubes are blocked. (Neander 1995, 119)

Her Fallopian tubes are malfunctioning. There is no question about that. The question Neander poses is whether other parts of the reproductive system, such as ovaries, are malfunctioning or not. She says that we could say that, for instance, ovaries are

malfunctioning because they fail to achieve their biological purpose, but we could also say that they are not malfunctioning because the failure is not their fault. In my terminology, the ovaries are not malfunctioning, but they are misfunctioning. It is misfunctioning because they fail to perform one of its functions (i.e. reproduction) not due to its own breakdown but due to unlucky situation where Fallopian tubes are blocked.

Summing up, it might be Millikan's view that misrepresentation always involves some kinds of misfunction ("failures of the cognitive systems to produce true beliefs generally result from absence of the right external-world background-condition" (Ryder et al. 2012, reply to Neander)), but it is not her view that misrepresentation always involve malfunction. Magnetosome misrepresents when it is fooled by a magnet bar, but it is not malfunctioning. It is, rather, just misfunctoining. Thus, Millikanian teleosemanticists, when they accept HDA, do not have to accept the conclusion that harmful misrepresentations are pathological.

This problem is, on the other hand, more serious in Neander's view. Neander's view is that the magnetosome doesn't misrepresent because it represents magnetic north, not oxygen-free sediment. According to Neander (1995, 2012), a representational state represent something in response to which the state is suppose to be caused, where "supposed to" is teleologically construed. The state of magnetosome represents magnetic north because the state is supposed to be caused by magnetic north, according to Neander.

Then, is Neander committed to the view that misrepresentation always involves

malfunction? She does seem to be committed to this. Neander discusses the example, which is also quite famous, of a frog (Rana Pipiens) that catches and eats flies. The frog, however, responds not just to fries, but to other small, dark, moving things with are not flies, such as bee bees. The point of this example is basically the same as the bacteria example. Let us say that R is the representational state of a frog caused by a fly or other small, dark, moving things. If we say that R represents flies, then R is misrepresenting when the R-token is caused by a flying bee bee. On the other hand, if we say R represents small, dark, moving things, then it is not misrepresenting when the R-token is caused by a flying bee bee. Neander's view is the latter. An R-token caused by a flying bee bee is not misrepresenting. Then, when does R-tokens misrepresent? Here is what Neander says;

> On the account that I favour, if the from R-tokens at anything which reflects onto its retina a pattern that falls outside of the specified parameters, then it misrepresents. The images cast by snails, for example, will fall outside these parameters, even when the snails are at their smallest and most sprightly, so R-tokening in response to a snail is a misrepresentation. […] frogs can make the mistakes mentioned above: some of them anyway. It's true that they can't/don't when they are functioning properly, but they don't always function properly. A sick frog might R-token at a snail if it was dysfunctional in the right way. Damaging the frog's neurology, interfering in its embryological development, tinkering with its genes, giving it a virus, all of these could introduce malfunction and error. (Neander 1995, 109)

According to Neander's view, R represents small, dark, moving things. Thus, an R-token misrepresents when it is caused by something other than small, dark, moving things. For

instance, it misrepresents when it is caused by a snail. But, this kind of mistake happens only when there is something wrong with Frog's perceptual mechanism ("Damaging the frog's neurology, interfering in its embryological development, tinkering with its genes, giving it virus"). Thus, it looks as though, after all, Neander is committed to the view that misrepresenting Rs always involve some kinds of malfunctions (not mere misfunctions).

But, the view that misrepresentation always involves malfunction is pretty implausible when it is applied to representations in general. It is pretty unlikely, for instance, that human misrepresentations such as healthy visual illusion or false beliefs involve malfunction in the same way that Frog's perceptual mechanism is malfunctioning due to neurological damage. Neander recognizes this problem. Here is her answer.

> Consider the case where we see a skinny cow in the dim distance and mistakengly represent it as a horse (Fodor's example). Here, we may suppose, we misrepresent without malfunctioning, and clearly the content of our perceptual representation goes beyond the physical parameters of the environmental features measured. But this sophisticated representation occurs after much visual processing has already taken place, at least, this is so on computational theories of vision. In such theories, early visual processing does not represent the cot as a horse (or as a cow) but as something which looks *a certain way* – as having a certain outline, texture, color, and so on. That is, according to conventional computational theories of perception, initially *there is a representation of the physical parameters of the environment as measured by the visual system*. It is much more plausible that there is no misrepresentation without malfunction at this level. (ibid. 132)

Neander's answer is that the claim that misrepresentation always involves malfunction is

applicable not to representations in general, but to a certain type of representations, including the representations in the early stages of visual processing and primitive representations in non-human animals. It doesn't applicable to sophisticated representations such as the visual representations in later stages, beliefs, and so on. Thus, after all, it is not Neander's view that healthy visual illusion or false beliefs involve malfunctions. From a broader point of view, thus, it is misleading to attribute the claim that misrepresentation always involves malfunction to Neander. This claim, according to her, only applies to a certain specific type of representations.

In sum, both kinds of teleosemantic theories, Millianian or Neanderian, do not hold that, in general, misrepresentations involve malfunctioning. Millikan would accept that misrepresentations always involve misfunction, but she would not accept that they always involve malfunction. Neander would accept that misrepresentation always involve malfunction in primitive type of representational states, but she would not accept that it is always the case. Thus, there is no real tension between teleosemantics and Wakefield's HDA.

Conclusion

The dissertation is about the two seemingly competing claims about delusions, namely, DD and CDT. Most, if not all, philosophers in the literature are incompatibilist. They think that DD and CDT are incompatible with each other and, thus, at least one of them needs to be rejected. The main aim of this dissertation is to develop an option that hasn't explored seriously so far, namely, a compatibilist option. Compatibilism is the view that that DD and CDT are compatible with each other. Compatibilists argue that they are compatible with each other according to some theories of belief and those theories are, at least, as plausible for independent reasons as functionalist ones according to which they are incompatible.

Here is a summary of each chapter.

In Chapter 1, I introduced DD and CDT with details and present some reasons for these claims. DD is supported by the fact that (1) delusional subjects sincerely assent to the content of their delusions, (2) the truth of DD is the part of the reason why delusion is a pathological condition, (3) DD is useful in distinguishing delusion from other pathological conditions, (4) most psychiatrists, the experts on the phenomenon, regard delusions as beliefs and (5) delusional subjects, who have authoritative first personal access to their own mental states, regard their delusions as beliefs. CDT, on the other hand, is supported by

numerous empirical and clinical observations suggesting that delusions are (A) extremely insensitive to evidence, (B) can be very incoherent with other beliefs, (C) can fail to evoke appropriate affective responses, and (D) can fail to cause appropriate action.

In Chapter 2, I motivated the compatibilist option by showing (a) that there is no good argument against the possibility of mental states without their distinctive causal roles, (b) that there are some examples, other than delusions, of mental states without their distinctive causal roles, and (c) that incompatibilist options (anti-DD and pro-DD) face some difficulties. On (a), I argued for instance that multiple-realizablity argument, which seems to support functionalist theories of belief, doesn't rule out the possibility of mental states without their distinctive causal roles. On (b), I suggested that pain asymbolia involves the pains without pain-like causal roles, addictive desires are the desires without desire-like causal roles, and so on. On (c), I showed that anti-DD incompabibilism has at least *prima facie* difficulty in accounting for the considerations that seem to favor DD and that Bortolotti's proposal (2010), which is the most well-developed pro-DD incompatibilist proposal, is problematic for a number of reasons.

In Chapter 3, I explored the best compatibilist theory of belief. Some candidates are examined and rejected. Type-identity theory is rejected because of its well-known problem about multiple-realizability of mental states. Humean phenomenalism is rejected because of its limited scope as a theory of belief. Lewis's statistical functionalism (1980) is rejected

because of its absurd consequences involving large-scale statistical changes. Bayne's normativism (2010) is rejected because of the indeterminacy as to whether a given mental state is a belief or not. The best candidate for the compatibilist theory of belief is what I call "teleo-attitude functionalism". The core idea of this theory is that beliefs are characterized in the same way that biological organs such as hearts or kidneys are characterized. For the same reason that there can be hearts that fail to pump blood, there can be beliefs that fail to play belief-like causal roles, according to teleo-attitude functionalism. This means that DD and CDT are compatible with each other according to this theory. Hearts that fail to pump blood are "malfunctioning hearts". Analogously, the beliefs that fail to play belief-like causal roles are "malfunctioning beliefs". Teleo-attitude functionalism has some important predecessors in philosophy and psychology including; Descartes's discussion on passion (Descartes 1649/1985), contemporary psychological accounts of emotion (Nesse 1990; Oatley & Johnson-Laird 1987), teleological functionalism (Lycan 1982, 1995; Sober 1985), and teleosemantics (Dretske 1986, 1991; Millikan 1984, 1989b).

Chapter 4 is the discussion of some theoretical issues about teleo-attitude functionalism. First, I discuss the issues on so-called "etiological analysis function" on which teleo-attitude functionalism rely. In particular, I argue that etiological analysis is the best option for teleo-attitude functionalists because it is a crucial part of the theory that beliefs can malfunction and etiological analysis is the most plausible analysis of function that allows for

malfunction (i.e. having the function of doing F without actually doing F). Many alternative analyses of function, such as systemic analysis (Cummins 1975), have difficulty in allowing for the possibility of malfunctioning. Counterfactual analysis (Nanay 2010) does allow for malfunction but it turns out to be less plausible than etiological analysis. Second, I respond to some expected objections to teleo-attitude functionalism, including philosophical objections with imaginary cases and empirical objections involving empirical considerations. Following Papineau (2001), I rejected philosophical objections, including Swampman objection and mad-belief objection, on the ground that teleo-attitude functionalisms is presented as the account of the way the world actually is. I critically examine the empirical, anti-adaptationist objections by Fodor (2001) and Gould (1991) and show that their attacks of adaptationist evolutionary psychology are based upon some theoretical confusions.

Chapter 5 explores the implications. First, I argued that DD is likely to be true, not merely compatible with CDT, according to teleo-attitude functionalism. According to the theory, a mental state is a belief just in case it has the right kind of etiological function. Since the etiological function of mental states are somehow derived out of etiological function of the mechanisms that produce and use the states (Millikan 1984), this amounts to the view that a mental state is a belief just in case it is produced by the right kind of mechanisms and used by the right kind of mechanisms. And I described the ways in which current and future empirical studies of delusion formation can help us to determine if

delusions are produced by the right kind of mechanisms and used by the right kind of mechanisms. Second, I showed that my proposal provides a nice account of the fact that delusion is, unlike self-deception or wishful thinking, regarded as a pathological condition. Delusions are, in my view, malfunctioning beliefs. I argue, on the basis of this, that delusion is a pathological condition because it is harmful (i.e. negatively affect well-being and social functioning) and it involves etiological malfunctions (of some underlying mechanisms). This account is just another application of Wakefield's influential account of disorder according to which a condition is a disorder just in case it involves harmful etiological malfunction.

Appendix: Two-Factor Theory and Prediction-Error Theory

6.1 Introduction

The two-factor theory (Davies et al. 2001; Coltheart 2007; Coltheart, Menzies & Sutton 2010) is an influential account of delusion formation. According to the theory, there are two distinct factors that are causally responsible for delusion formation. The first factor is supposed to explain the content of the delusion, while the second factor is supposed to explain why the delusion is adopted and maintained. Recently, another remarkable account of delusion formation has been proposed, in which the notion of "prediction error" plays the central role (Fletcher & Frith 2009; Corlett et al. 2009; Corlett et al. 2010). According to this account, the prediction-error theory, delusions are formed in response to aberrant prediction-error signals, those signals that indicate a mismatch between expectation and actual experience[1].

---

[1] There are many different versions of two-factor and prediction-error theory. In this chapter, I will put our focus on the most influential version of the two-factor theory, by Coltheart and others, and the most influential version of the prediction-error theory, by

Is the prediction-error theory a rival to the two-factor theory? Prediction-error theorists tend to be critical about the two-factor theory and present their views as an alternative to it. Fletcher and Frith wrote, on the two-factor theory: "symptoms reflecting false perception and false beliefs are so intertwined in schizophrenia that a theory relying on coincidental damage seems very unlikely." (Fletcher & Frith 2009, 51) Again, Corlett and colleagues argue that positing two factors is redundant because "a single deficit in Bayesian inference is able to explain more of what we know about the interactions between perception and belief-based expectation, the neurobiology of the delusions that occur in schizophrenia and the maintenance of delusions in the face of contradictory evidence." (Corlett et al. 2010, 357) In this chapter, I examine the relationship between the two-factor theory and the prediction-error theory in some detail. Our view is that the prediction-error theory does not have to be understood as a rival to the two-factor theory. I do not deny that there are some important differences between them. However, those differences are not as significant as they have been presented in the literature. Moreover, the core ideas of the prediction-error theory may be incorporated into the two-factor framework. For instance, the aberrant prediction-error signal that is posited by prediction-error theorists can be (or underlie) the first factor contributing to the formation of some delusions, and help explain the content of

---

Corlett and others.

those delusions. Alternatively, the aberrant prediction-error signal can be (or underlie) the second factor, and help explain why the delusion is adopted and maintained.

Sections 2 and 3 will offer an overview of the two-factor theory and the prediction-error theory respectively. Section 4 will host a discussion of the relationship between the two theories. First, I examine the major differences between them. Then, I explore the ways in which some central ideas of the prediction-error theory can be incorporated into the two-factor framework.

6.2 The Two-Factor Theory

6.2.1 Why Two Factors

The two-factor theory is primarily a theory of monothematic delusions (delusions concerning single themes). The most extensively discussed example of monothematic delusion in the literature is that of the Capgras delusion.

Ellis and Young (1990) argue that the Capgras delusion is formed in response to the abnormal experience of seeing familiar faces. The experience is abnormal in that it lacks the affective component that is usually a part of the experience. The Capgras delusion is formed as an attempt to explain this abnormal perceptual-affective experience.

> [Capgras patients] receive a veridical image of a person they are looking at, which stimulates all the appropriate overt semantic data held about that person, but they lack another, possibly confirming, set of information which […] may carry some sort of affective tone. When patients find themselves in such a conflict, […] they may adopt some sort of rationalisation strategy in which the individual before them is deemed to be an imposter, a dummy, a robot, or whatever extant technology may suggest. (Ellis & Young 1990, 244)

This hypothesis is supported by the finding that people with the Capgras delusion do not show the asymmetrical autonomic responses between familiar faces and unfamiliar faces (Ellis et al. 1997).

Is the abnormal perceptual-affective experience of familiar faces sufficient for the development of the Capgras delusion? If it were so, then anyone with this abnormal experience would, other things being equal, develop the delusion. This view that the abnormal experience is sufficient for the formation of the delusion is usually attributed to Maher (1974) who thinks that delusions are a perfectly normal response to abnormal experience. According to him, "a delusion is a hypothesis designed to explain unusual perceptual phenomena and developed through the operation of normal cognitive processes." (Maher 1974, 103)

Two-factor theorists, however, do not think that the abnormal experience is sufficient for the development of Capgras delusion. In addition to the abnormal experience[2], there has to

[2] Some two-factor theorists avoid the term "abnormal experience" because they think that

be another causal factor that is responsible for the development of the Capgras delusion. The main argument for the additional factor (Davies et al. 2001; Coltheart 2007; Coltheart, Langdon & McKay 2011) goes as follows. Just like people with Capgras delusion, people with damages to the ventromedial prefrontal cortex (VMPFC) fail to show the asymmetrical autonomic responses between familiar faces and unfamiliar faces (Tranal, Damasio & Damasio 1995). This suggests that people with Capgras delusion and people with VMPFC damage are in similar experiential conditions. However, people with VMPFC damage do not develop the Capgras delusion. This shows that the abnormal experience can be dissociated from Capgras delusion. One can have the abnormal experience without forming the delusion. There are two possible ways to account for the dissociation between experience and delusion. Either the abnormal experience may not be causally responsible for the development of Capgras delusion after all. Or, in addition to the abnormal experience, there must be an additional factor contributing to the development of the Capgras delusion. The first option is unlikely because the content of the Capgras delusion has an obvious connection to the abnormal experience. Therefore, the only available option is the second one. Let us call this "dissociability argument for the second factor".

---

the first factor is an unconscious event involving abnormal autonomic activity (Coltheart, Menzies & Sutton 2010). For the sake of simplicity, however, I will stick to this terminology in the rest of the paper.

The dissociability argument is applicable not only to the Capgras delusion, but also to many other monothematic delusions, because for each monothematic delusion, there are some people who might have experiences quite similar to those of the relevant delusional subjects, but nonetheless do not develop the relevant delusions. For example, people with schizophrenia who experience global affective flattening might be experientially quite similar to people with the Cotard delusion (who believe that they are dead or disembodied), but do not develop the Cotard delusion. Some people with depersonalization disorder might be experientially quite similar to the people with delusion of alien control (who believe that the parts of their bodies are controlled by someone else), but do not develop the delusion of alien control. People with severe prosopagnosia (face blindness) might be experientially quite similar to the people with mirrored-self misidentification (who believe that the reflected images on the mirror are someone else), but do not develop the delusion of mirrored-self misidentification. And the list could go on.

Suppose that the dissociabillity argument about is successful. This establishes, for example, that the abnormal experience is not sufficient for the development of the Capgras delusion. Now, there are broadly two possible interpretations of the insufficiency. First, the abnormal input is sufficient for the adoption of the Capgras hypothesis (e.g., the hypothesis that this woman is not my wife) but not sufficient for its maintenance. In that case, the role of the second factor is to explain the maintenance of the Capgras delusion. Coltheart,

Menzies and Sutton (2010) hold such a view. According to their version of the two-factor theory, the second factor is the failure of incorporating the evidence against delusional beliefs. The second factor explains why the Capgras delusion is maintained despite the overwhelming couterevidence (see below for more details). Second, the abnormal input is not even sufficient for the adoption the Capgras hypothesis. In that case, the role of the second factor is to explain the adoption (and the maintenance) of the Capgras hypothesis. McKay (2012) holds such a view. According his version the two-factor theory, the second factor is the bias of discounting prior probability ratio. The second factor explains why the Capgras delusion is formed (and maintained) despite the low prior probability of the Capgras hypothesis (see blow for more details).

One fundamental commitment of the two-factor theory is the idea that *there are two distinct factors that are causally responsible for the development (including adoption and maintenance) of delusions*. Here, to say that two factors are distinct is to say that they can be dissociated from each other. For instance, the second factor of Capgras delusion, whatever it is, is dissociable from the first factor, the abnormal experience. They are actually dissociated, for instance, in people with VMPFC damage.

Another important commitment of the two-factor theory is the idea that two factors play different explanatory roles. In particular, *the first factor explains the content of the delusion, whilst the second factor explains the fact that the delusional belief is developed (i.e., adopted*

and maintained) (Coltheart 2007). For instance, the abnormal perceptual-affective experience of familiar faces, which is often regarded as the first factor in the Capgras delusion, is certainly explanatory with regard to the content of the Capgras delusion. It explains why, for instance, a person with Capgras believes that his wife has been replaced by an imposter. It does not equally explain other beliefs the person might have, for instance, that he is followed by CIA agents or that the world is coming to an end. On the other hand, the abnormality in the belief evaluation system located in right hemisphere, which is the second factor according to Coltheart (2007), explains why delusional hypotheses are adopted and maintained. For instance, it explains why, unlike people with VMPFC damage[3], people with Capgras adopt and maintain the imposter hypothesis.

There are some popular assumptions about the first and the second factors. First, two-factor theorists tend to think that the first factors are different in different kinds of delusions. Different kinds of delusions have very different content. Thus, assuming that the first factors explain content, different kinds of delusions need to have different kinds of the first factors. On the other hand, the second factor is often said to be the same in all kinds of delusions (Coltheart 2007). Second, it is often assumed that the second factor has something

---

[3] Some two-factor theorists maintain that people with VMPFC damage do not *adopt* the imposter hypothesis (e.g. McKay 2012), whilst the other argue that they adopt it but they do not *maintain* it (e.g. Cotlheart, Menzies & Sutton 2010).

to do with some sort of right hemisphere abnormalities. This assumption follows from the observation that people with delusions due to neurophysiological impairments tend to have deficits in right hemisphere, in particular, right prefrontal cortex (Coltheart 2007).

6.2.2 Problems with the Two-Factor Theory

There are two main difficulties for the two-factor theory. One is related to the scope of the theory, while the other is related to the nature of the second factor.

Problem of Scope: The first problem is that the scope of the theory seems to be rather limited. The two-factor theory is, as I already noted, primarily about the formation of monothematic delusions. It is not clear how the theory could be applied to other kinds of delusions. In particular, its applicability to polythematic delusions (i.e., delusions involving many different themes) is unclear. Coltheart admits that the two-factor theory "has focused principally on monothematic delusion, largely neglecting polythematic delusions, including some that are especially frequently seen in clinical patients." (Coltheart 2013, 105)

The applicability of the theory to motivationally formed delusions is also unclear. McKay, Langdon and Coltheart (2005) refer to a case of a man with reverse Othello Syndrome who, after a serious car accident that caused severe head injury and quadriplegia, believed that he and his former partner had recently married, despite the fact that the former partner severed all contact with him after the accident (Butler 2000). A plausible story about this

case is that the man's delusional belief about his recent marriage was formed (partly) because he did not want to accept the stark reality. Does the two-factor theory have anything to say about this case? McKay, Langdon and Coltheart write, "[t]he problem for the two-factor account as it currently stands is that there are myriad cases of delusions that similarly resist ready identification of potential first deficits, but for which plausible motivational stories can be told." (McKay, Langdon & Coltheart 2005, 313)

Two-factor theorists have some ideas to solve these problems. Coltheart, Langdon and McKay (2011) suggest three potential scenarios in which two-factor accounts are applicable to polythematic delusions. First, polythematic delusions might be caused by multiple first factors. Second, polythematic delusions might be developed when the second factor is very serious. Third, polythematic delusions might be developed when the first factor is relatively ambiguous and open to different explanations. According to McKay, Langdon and Coltheart (2005), the two-factor account can be applied to motivationally formed delusions if the first factor includes motivational states such as a desire or a suggestion from a psychological defense mechanism. For instance, the case of Reverse Othello syndrome above might be understood as a case where the first factor is the desire not to accept the loss of one's former romantic partner.

Problem of the Second Factor: The most serious problem of the two-factor theory is that the nature of the second factor is underspecified. All two-factor theorists agree on the

existence of the second factor. But, there is no agreement about its nature. Coltheart's proposal, that the second factor is an abnormality in the belief evaluation system located in the right hemisphere, does not specify exactly which kind of abnormality it is. Could it be, for instance, a reasoning bias?

There have been several suggestions as to the nature of the abnormality. The second factor might be the presence of a jumping-to-conclusion bias (the bias of coming to a conclusion with less evidence in comparison to non-clinical samples, see Huq, Garety & Hemsley 1988), an externalizing attribution bias (the bias of attributing negative events to external objects or people instead of oneself, see Young 2000), a bias toward observational adequacy (the bias of putting too much weight on observational data, discounting existing beliefs, see Stone & Young 1997) or a failure to inhibit pre-potent doxastic responses to perceptual experience (the failure to adopt a critical stance towards perceptual content and accept it as veridical, see Davies et al. 2001). Although some of the above suggestions hold some promise, it is not clear whether one bias would be sufficient in isolation from the other ones to underlie the second factor.

Recently, the nature of the second factor has been investigated by appealing to quantitative models of the inferential step in the delusion formation process (Coltheart, Menzies & Sutton 2010; McKay 2012; Davies & Egan 2013). Coltheart, Menzies and Sutton, for instance, argue that the inferential step in the Capgras delusion formation process is

perfectly Bayesian-rational. Let Hs be the hypothesis that the woman in front of me is a stranger, Hw be the hypothesis that the woman is my wife, and O be the abnormal perceptual-affective experience. So it follows from Bayes' theorem that:

$$\frac{\mathrm{P(Hs|O)}}{\mathrm{P(Hw|O)}} = \frac{\mathrm{P(Hs)}\times\mathrm{P(O|Hs)}}{\mathrm{P(Hw)}\times\mathrm{P(O|Hw)}}$$

Coltheart, Menzies and Sutton argue that:

$$\mathrm{P(Hs)} < P(\mathrm{Hw})$$
$$\mathrm{P(O|Hs)} \gg \mathrm{P(O|Hw)}$$
$$\mathrm{P(Hs|O)} > P(Hw|O)$$

In other words, Hw has a higher prior probability than Hs (i.e., Hw is more plausible than Hs before O becomes available). But, it is outweighed by the likelihood ratio that overwhelmingly favors Hs (i.e., Hs explains O overwhelmingly better than Hw). As a result, Hs gets a higher posterior probability than Hw (i.e., Hs is more plausible than Hw given O). And, assuming that this is the way people with Capgras actually reason, Coltheart, Menzies and Sutton argue that the Capgras delusion is adopted in a Bayesian-rational probabilistic inference. Since this inferential step is Bayesian-rational, they argue, there is no second factor here. The second factor rather operates at the post-adoption stage where

counterevidence to Hs becomes available. For instance, after a person with Capgras adopts Hs, he will realize that trusted friends or seemingly reliable psychiatrists are saying that the woman in his house is his wife, not an imposter. He will also find that the "imposter" does not just look exactly like his real wife, but also behaves exactly like his real wife, knows everything his real wife knows, and so on. All of this evidence supports Hw instead of Hs. The person with Capgras fails to incorporate the new body of evidence into his belief system, because he neglects new evidence due to a bias. And this bias of neglecting new evidence is the second factor.

This is certainly an interesting proposal. But, does it really solve the problem of the second factor once and for all? Presumably it does not. If people with Capgras have the bias of neglecting new evidence, then why they do not neglect the abnormal experience, which is new evidence acquired at the beginning of the process? People with Capgras adopt Hs because Hs is more plausible given the abnormal experience, according to the proposal. Here, they do not neglect the abnormal experience at all. Why is it that they do not neglect the abnormal experience despite having a bias that leads them to neglect new evidence? To be consistent, presumably Coltheart, Menzies and Sutton need to say that the bias of neglecting new evidence is acquired after Hs is adopted (so that the abnormal experience is not neglected) and before counterevidence to Hs becomes available (so that the counterevidence is neglected). But, as McKay (2012) points out, this chronological story is

quite unrealistic and implausible[4].

Summing up, the two-factor theory consists of two main claims: (1) there are two distinct factors that are causally responsible for the development of delusion; (2) the first factor explains the content of delusion, while the second factor explains why the delusional hypotheses are adopted and maintained. The two-factor theory is supported by the dissociability argument in support of the second factor; the existence of the second factor is defended on the basis of the dissociability between the first factor (e.g., an abnormal perceptual-affective experience) and the presence of delusional beliefs (e.g., the Capgras delusion). A problem for the two-factor theory is that the scope of the theory seems to be limited. It is not clear how the theory is applicable to polythematic delusions and motivationally formed delusions. Another problem is that there is no agreement about the nature of the second factor. Even the recent quantitative two-factor accounts might not be fully satisfactory.

[4] For more discussions on the proposal, see McKay (2012) and Davies and Egan (2013).

6.3 Prediction-error theory

6.3.1 Salient Experience and Prediction Errors

Kapur (2003) argues that the abnormality in dopamine transmission in schizophrenia leads to an inappropriate attribution of "salience". The attribution of salience is "the process whereby events and thoughts come to grab attention, drive action, and influence goal-directed behavior because of their association with reward or punishment." (Kapur 2003, 14) When salience is attributed inappropriately to events that are not very interesting as a matter of fact, they grab special attention. Such events, all of a sudden, seem to be important and are invested with special meaning. This hypothesis explains the fact that many people with schizophrenic delusions report that some events have "special meaning" for them, that they have an "altered experience" of the world, that their awareness is "sharpened", and so on. The delusion, according to Kapur, is produced as the explanation of the inappropriately salient events.

> "Delusions are 'top-down' cognitive explanation that the individual imposes on these experiences of aberrant salience in an effort to make sense of them. [...] Once the patient arrives at such an explanation, it provides an 'insight relief' or a 'psychotic insight' and serves as a guiding cognitive scheme for further thoughts and actions." (Kapur 2003, 15)

The prediction-error theory of delusion provides a more detailed story about how salience is

inappropriately attributed in people with schizophrenia and delusions (Corlett et al. 2010; Fletcher & Frith 2009). According to the theory, *inappropriate attribution of salience is the product of aberrant prediction-error signals*.

In general, a prediction error tells us that what we experience does not match our predictions. It indicates that the internal model of the world from which the prediction is derived is incorrect and needs to be updated. By updating the model in such a way as to minimize prediction errors, we improve our understanding of the world. In this framework, prediction-error signals play a fundamental role in allowing us to update our model of the world, and to update our beliefs. Prediction-error theorists posit abnormal prediction-error signaling in people with delusions. In particular, it is hypothesized that, in people with delusions, *prediction-error signals are excessive in the sense that they are produced when there is no real mismatch between prediction and actual inputs*. The excessive prediction-error signals falsely indicate that our internal model of the world needs to be updated even though, as a matter of fact, it does not have to be.

Prediction-error signals also play a crucial role in allocating attention in such a way that we allocate attention to the events that defy our expectations. This makes sense intuitively since the predictable events do not bring any new information and, thus, do not deserve attention. Due to excessive prediction-error signals, people with delusions allocate attention to events that do not actually deserve it. This is the process where uninteresting events

become inappropriately salient, and delusions are produced as explanatory responses to the inappropriately salient events due to aberrant prediction-error signals. Corlett and colleagues (2009) nicely summarize the main claim of the prediction-error theory.

> "Prediction error theories of delusion formation suggest that under the influence of inappropriate prediction error signal, 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)

The remarkable study by Cortlett and colleagues (2007) strongly supports this aberrant prediction-error signal hypothesis. In the study, two groups of participants, people with delusions and people without, are tested with a task involving learning the association between certain foods and allergic reactions. The activity of right prefrontal cortex (which was identified as a reliable marker of prediction-error processing in previous studies) was monitored with fMRI. They found that, in the delusional group but not in the control group, the magnitude of the activity of the right prefrontal cortex is not significantly different between the cases where the expectations about allergic reaction were confirmed and the cases where they were violated. They also found that the severity of this abnormal prediction-error signaling was correlated with the severity of their delusion.

6.3.2 Problems with the Prediction-Error Theory

There are two main challenges for the prediction-error theory. First, aberrant prediction-error signals might be dissociable from delusions. Second, the theory (or, at least, its main claim) does not say anything about how delusions are maintained after they are adopted.

Dissociability Problem: Prediction-error theorists typically assume that aberrant prediction-error signals are sufficient for the development of delusion. This assumption, however, is problematic. Here is the account of the Capgras delusion offered by Corlett and colleagues (2010). Corlett and colleagues accept Ellis and Young's hypothesis that people with Capgras delusion have abnormal perceptual-affective experience of familiar faces. They interpret this as a kind of prediction error where the expected affective experience does not match the actual experience. The Capgras delusion is formed as an explanatory response to this affective prediction error. Unlike two-factor theorists, Corlett and colleagues are explicitly committed to the idea that this affective prediction error is sufficient for the development of the Capgras delusion. "[W]e argue that phenomenology of the percepts are such that bizarre beliefs are inevitable; surprising experiences demand surprising explanations." (Corlett et al. 2010, 360)

The problem with this commitment is that the affective prediction error is dissociable from the Capgras delusion. For instance, it might be that the affective prediction error is

occurring not only in people with Capgras delusion, but also in people with VMPFC damage. But, the latter do not develop the Capgras delusion. As we previously discussed, this shows that either the affective prediction error is causally irrelevant to the formation of the Capgras delusion or it is not sufficient for it. The first option cannot be accepted by prediction-error theorists. Thus, it looks as though they need to accept that the affective prediction error is not sufficient.

There are several possible responses from prediction-error theorists. For example, they could argue that the affective prediction error in the people with VMPFC damage is not as serious as the one in people with the Capgras delusion and "the salience of the data in question has simply not passed a certain threshold" (McKay 2012, 348) in the former. Alternatively, they might argue that people with VMPFC damage actually develop something like the Capgras delusion. For example, Corlett and colleagues (2010) argue that some people with damages to the ventromedial and the lateral prefrontal cortex develop delusion-like spontaneous confabulation.

Problem of Maintenance: The second problem for the prediction-error theory is that it explains the development of delusion only up to the point where delusional beliefs are adopted as an account of inappropriately salient events. But, another thing to be explained is the fact that, after the initial adoption, delusions are maintained despite overwhelming counterevidence. Remember that the definition of delusion in DSM-5 says that delusion "is

firmly sustained despite what almost everyone else believes and despite what constitutes incontrovertible and obvious proof or evidence to the contrary." (American Psychiatric Association 2013, 819) It looks as though the account of this aspect of the delusion is missing in the main claim of prediction-error theory. Corlett and colleagues admit that "this model accounts for why delusions emerge but not for why they persist." (Corlett et al. 2013, 2)

This explanatory gap is more evident when we compare the prediction-error theory with the two-factor theory. The maintenance of delusion is the central issue for two-factor theorists. A part of the aim of positing the second factor is to explain the maintenance of delusion. For example, as we already saw, Coltheart, Menzies and Sutton (2010) propose that the second factor is the bias of neglecting new evidence. If people with delusions really have this bias, it certainly explains why delusions are firmly maintained despite overwhelming counterevidence. The counterevidence is simply ignored because of the bias.

Presumably, it is not just that the account of maintenance is missing in the prediction-error theory, but also that the theory has a peculiar difficulty explaining it. It is hypothesized, in the theory, that prediction-error signals are excessive in people with delusions. This presumably does not suddenly come to an end when a delusional hypothesis is adopted. In other words, prediction-error signals will remain excessive even after the adoption of delusional hypotheses. But, this seems to predict that delusions, after adoption, will be unstable instead of firmly maintained. Suppose that a subject adopts a delusional

hypothesis as an account of inappropriately salient events due to excessive prediction-error signals. Now, the delusional hypothesis becomes an important part of his internal model of the world. After adopting it, he will be again exposed to recurrent excessive prediction-error signals that indicate, this time, that the delusional internal model of the world is inaccurate and needs to be updated. It is expected, then, that he will update the delusional internal model of the world so that prediction error will be minimized. But, as a matter of fact, people with delusions do not update their delusional hypotheses in this way. They just stick to their commitments even in the face of overwhelming counterevidence. This is quite puzzling.

Corlett and colleagues (2009; 2013) recognize this problem and try to solve it by offering an account of the maintenance of delusion on the basis of the idea of memory reconsolidation[5]. Belief (including delusional belief) is, according to Corlett and colleagues, a kind of memory. Thus, belief is maintained in the same way that memory is maintained. In particular, memory reconsolidation plays a crucial role in the maintenance of delusional beliefs. After the initial adoption of delusion, due to the abnormal activity in the midbrain reminder system, "aberrant prediction errors might re-evoke the representation of the delusion without definitively disconfirming it. This would drive preferential reconsolidation over and above any new extinction learning. The net effect would be a strengthening of the

[5] See also Murray (2011).

delusion through reconsolidation rather than a weakening by extinction." (Corlett et al. 2009, 3)

To summarize, according to the prediction-error theory, a delusion is formed as an explanation of events that are inappropriately salient due to aberrant prediction signals. The hypothesis is consistent with the subjective report of delusional subjects, and is supported by empirical studies suggesting abnormal prediction-error signaling in delusional subjects. A problem of the theory is that aberrant prediction-error signals might not be sufficient for the development of the delusion, contrary to the assumption made by many prediction-error theorists. Another problem is that the account of the maintenance of the delusion in the face of overwhelming counterevidence is missing in the main claim of prediction-error theory. Thus, it needs to be supplemented by an additional account of maintenance, such as the one appealing to the idea of memory reconsolidation.

6.4 Two-factor theory vs. prediction-error theory

In this section, I examine the relationship between two-factor theory and prediction-error theory in detail. First, I discuss the main differences between the two theories, and then explore the ways in which the ideas of prediction-error theory can be incorporated into the two-factor framework.

6.4.1 Differences

There are many differences between two-factor and prediction-error theory. But, the main differences between them seem to be about (1) target phenomena, (2) the number of factors, and (3) the perceiving-believing relation.

(1) *Target Phenomena*: Two-factor theory is primarily about monothematic delusions. In particular, the theory is quite persuasive about monothematic delusions of neuropsychological or organic origins (such as a stroke, traumatic brain injury, etc.). People with monothematic delusions of neuropsychological origins often have abnormalities in the right hemisphere, which supports the existence of the second factor independently from the dissociability argument. On the other hand, prediction-error theory is primarily about delusions in schizophrenia. Often, the theory is motivated by general neurological hypotheses about schizophrenia, such as the dopamine hypothesis, and presented as a part of the general theory of the positive symptoms of schizophrenia, including delusion and hallucination. Fletcher and Frith note that their proposal is that "a common mechanism, involving minimization of prediction error, may underlie perception and inference, and that a disruption in this mechanism may cause both abnormal perceptions (hallucinations) and abnormal beliefs (delusions)." (Fletcher & Frith 2009, 48) In addition, the empirical support for prediction-error theory, such as the study with allergy detection task by Corlett and

colleagues (2007), comes from studies on people with schizophrenic delusions.

(2) *Number of Factors*: It is the essential commitment of the two-factor theory that there are two factors that are causally responsible for the development of delusion. Instead, the prediction-error theory is typically presented as a one-factor account, where the factor is the aberrant prediction-error signal.

> "Prediction error driven Bayesian models of delusions subsume both factors into a single deficit in Bayesian inference; noise in predictive learning mechanism engender inappropriate percepts which update future priors, leading to the formation and maintenance of delusions." (Corlett et al. 2010, 357)

> "[…] the positive symptoms of schizophrenia seem to reflect two underlying abnormalities, suggesting that a two-factor explanation is required. However, recent computational models of perception and learning suggest that the same fundamental mechanism (Bayesian inference), by which a model of the world is updated by prediction errors, applies to both perception and belief-formation. […] We suggest that positive symptoms of schizophrenia are caused by an abnormality in the brain's inferencing mechanisms, such that new evidence (including sensations) is not properly integrated, leading to false prediction errors." (Fletcher & Frith 2009, 56)

(3) *Perceiving and Believing*: Prediction-error theorists typically assume that there is no categorical distinction between perceiving and believing. Accordingly, they typically do not draw a sharp distinction between abnormal perception (hallucination) and abnormal belief (delusion). The reason why there is no categorical distinction between perceiving and believing is that, according to their view, both are operating on the same principle, namely,

minimizing prediction error.

> "The boundaries between perception and belief at the physiological level are not so distinct. An important principle that has emerged is that both perception of the world and learning about the world (and therefore beliefs) are dependent on predictions and the extent to which they are fulfilled. This suggests that a single deficit could explain abnormal perceptions and beliefs." (Fletcher & Frith 2009, 51)

> "Within this theoretical framework belief itself must, like perception, be a matter of prediction error minimisation. The difference between belief and perception lies in the time scale of the represented processes and their degree of invariance or perceptive independence. There is no further special difference between them and the issue of rationality applies equally to perception and belief." (Hohwy & Rajan 2012, 10)

Two-factor theorists, on the other hand, are typically committed to the categorical distinction between perceptual/experiential processes and believing processes. This is to be expected, because it is the core claim of the two-factor theory that the experiential factor (the first factor) is not sufficient for the development of delusional beliefs.

> "As a two-factor theorist, my view is that the distinction between perception and belief is not easily dispensed with – particularly at the personal level of description. As Martin Davies (personal communication) notes, 'One indication of this is that obliterating the distinction between perception and belief seems to have the consequence that illusory perceptual states that are informationally encapsulated from most of a person's belief are liable to be counted as delusions." (McKay 2012, footnote 26)

These are the main differences between two-factor and prediction-error theory that anyone

interested in the relationship between them should be aware of. However, these differences are, in our view, not as significant as one might expect.

The difference in terms of target phenomena is not very significant, given that monothematic delusions and schizophrenic delusions often overlap with each other. For example, the Capgras delusion is a typical example of monothematic delusion, but can be observed in people with schizophrenia. In addition, as Coltheart (2010) points out, if the distinction between monothematic and polythematic is mapped onto the distinction between neuropsychological and non-neuropsychological delusions, then it is not clear-cut. We cannot be sure that there are any non-neuropsychological delusions. Furthermore, it might turn out that the two delusion-formation theories, even though they start from monothematic delusions and schizophrenic delusions respectively, are applicable to other kinds of delusions in the end. For example, as I already noted, there are some possible scenarios in which the two-factor theory is applicable to polythematic and motivationally formed delusions (without obvious neuropsychological origins). Again, if the prediction-error theory is a persuasive account of, say, the Capgras delusion in schizophrenia, it is perfectly possible that the same account is applicable to the Capgras delusion occurring in other contexts. As Coltheart, Langdon and McKay noted, the Capgras delusion is likely to be "cognitively homogeneous" even though it is "etiologically heterogeneous." (Coltheart, Langdon & McKay 2007, 645) If the account of the Capgras delusion given by

prediction-error theorists is plausible in the context of schizophrenia, then it is also a possible candidate for a general account of the Capgras delusion that is applicable outside the context of schizophrenia.

The difference with regard to the number of factors might not be very significant either. It is not obvious that the prediction-error theory is necessarily a one-factor theory. It is certainly true that the prediction-error theory is, as a matter of fact, often presented as a one-factor account. But, are there any reasons to think that the prediction-error theory needs to be a one-factor theory? According to Corlett and colleagues (2010), the affective prediction error is sufficient for the formation of the Capgras delusion, and theirs is certainly a one-factor account. But, not all prediction-error theorists have to follow Corlett and colleagues on this issue. Prediction-error theorists could think that the Capgras delusion is an abnormal response to an affective prediction error and that there has to be an additional factor that explains that the abnormality of the response. The commitment to an additional factor might give a better account of the dissociability between the Capgras delusion and the affective prediction error. According to the version of the prediction-error theory we are considering, people with VMPFC damage would not develop the Capgras delusion even though they were experiencing affective prediction error because the affective prediction error would not be sufficient for the development of Capgras delusion. There would have to be an additional factor.

As I noted, the prediction-error theory is often presented as a one-factor account. But, it is not obvious that, strictly speaking, it qualifies as a one-factor account. For example, an account of the maintenance of delusion is not explicit in the main body of the prediction-error theory. This means that prediction-error theorists need to tell some further story about maintenance. It might turn out that the further story posits an additional abnormal factor. In this case, prediction-error theorists would posit two factors after all: an aberrant prediction error signal to account for the adoption of the delusion, and an additional factor to account for its maintenance. For instance, think about the account by Corlett and colleagues of the maintenance of the delusion, according to which delusions are maintained through the process of memory reconsolidation. If it turns out that the memory reconsolidation process is abnormal in people with delusions, and this abnormality is dissociable from aberrant prediction-error signals, then the proposal does not qualify as a one-factor account. It posits two distinct factors: aberrant prediction-error signals and an abnormality in the memory reconsolidation process. Such a proposal would also be consistent with the claim by two-factor theorists that the two factors play different explanatory roles. The affective prediction error would explain why a hypothesis with delusional content is adopted, and the abnormality in the memory reconsolidation process would explain why the delusional hypothesis is maintained. In this case, the proposal would be just another version of the two-factor theory.

Sometimes, prediction-error theorists show a subtle attitude towards the number of factors. Hohwy and Rajan, for instance, argue that the prediction-error theory is different from the Maher-style one-factor account because it "goes beyond one- vs two-factor view." (Hohwy & Rajan 2012, 9) Unlike Maher, they posit more than one factor, but, unlike two-factor theorists, they do not believe that there is a categorical distinction between the two factors. The two factors might be *numerically* distinct (in the sense that they are dissociable from each other), but they are not *categorically* distinct (in the sense that they do not belong to different neurological or psychological categories).

The third difference between the two theories concerns the perceiving-believing relation. Is perceiving categorically distinct from believing? This seems an important issue, but we should not overestimate this difference between the two theories. First, the notion of belief in the vocabulary of prediction-error theorists is sometimes peculiar and, accordingly, we need to be careful in interpreting their claim that there is no categorical distinction between perceiving and what they call "believing". Corlett and colleagues characterize belief as "probability distributions that are represented by the brain." (Corlett et al. 2010, 347) This notion of belief might not overlap perfectly with the ordinary, folk psychological notion of belief. After all, believing is what we, not our brain, do, in the ordinary usage of the term. We, not our brains, believe something. It seems to be possible that my brain assigns a high probability to a certain hypothesis, but the hypothesis is not introspectively accessible to me

or, even though it is accessible, it does not guide my actions or thoughts in the way normal beliefs do. Presumably, I do not believe the hypothesis in this case, at least not in the ordinary usage of the term, even though my brain assigns a high probability to it.

Second, it is not obvious that the two-factor theory is necessarily committed to the claim that perceptual processes and believing processes are categorically different. As I already noted, the central commitments of the two-factor theory are: (1) there are two distinct factors that are causally responsible for the development of delusion, and (2) those two factors play different explanatory roles. The question that arises is whether these two claims imply a categorical difference between perceptual processes and believing processes.

One might reason in the following way. Those claims imply the categorical difference between the first and the second factor and, since the first-second factor distinction amounts to the perceiving-believing distinction, the commitment to the first distinction implies the commitment to the second distinction. The problem in this line of reasoning is that neither (1) nor (2) entails the categorical difference between the first and the second factor. (1) is committed to the distinctness between the first and the second factors. But, as we already noted, this notion of "distinctness" only means the dissociability between the first and the second factors, not a categorical difference between them. It is perfectly possible that two factors are dissociable, and yet they belong to the same neurological or psychological category. Again, (2) is committed to the explanatory difference between the

first and the second factor. But, it is not obvious that this entails a categorical difference between them. It is possible that the two factors play different explanatory roles, and they belong to the same neurological or psychological category.

When emphasizing the continuity between perceiving and believing, prediction-error theorists often appeal to Helmholtz's (1878/1968) famous idea that perceiving involves unconscious inferential processes based upon prior knowledge and sensory inputs. Belief formation is often regarded as an inferential process but, according to the Helmholzian picture, perception is inferentially formed too. Thus, there is no categorical distinction between believing and perceiving in this regard. They both involve inferences. This Helmholzian picture would not be incompatible with the two-factor theory. Indeed, Coltheart, Menzies and Sutton (2010) are very sympathetic to the Helmholzian picture. Their proposal is that the Capgras delusion is formed through an unconscious Bayesian inferential process based upon prior beliefs as well as abnormal data involving reduced autonomic response to familiar faces. They admit that this proposal is similar to the Helmholzian picture.

Again, Coltheart (2010) makes it explicit that the core claims about perception and learning by Corlett and other prediction-error theorists are perfectly compatible with the two-factor theory.

> "The first is that much of what we perceive is based in our expectation. That conception is also central to the two-factor theory of delusion because the first factor in most delusions is a violation of expectation (e.g., in Capgras delusion the violation of the expectation of a strong autonomic response upon seeing the face of one's spouse.) The second is that perception expectations are learned; that is also consistent with the two-factor theory, because we learn how a mirror works, we learn that objects and people of value to us will generate strong autonomic responses, etc. The third is that mismatches between what we expect and what we experience drive the updating of our expectancies: that too is what the two-factor theory proposes, merely replacing the term 'updating of our expectancies' with the term 'revision of our belief system'." (Coltheart 2010, 25)

6.4.2 Incorporating Elements of the Prediction-Error Theory into the Two-Factor Theory

I just reviewed the major differences between the two-factor and the prediction-error theories. The differences, I argued, are not as significant as one might expect. In this subsection, I explore ways in which some ideas from the prediction-error theory can be incorporated in the framework of the two-factor theory. I begin by motivating such a hybrid strategy.

First, both theories have remarkable theoretical virtues and empirical adequacy. If we choose one and deny the other, then we loose the theoretical virtues and empirical adequacy of the theory that we do not choose. When, for example, we choose the two-factor theory and deny the prediction-error theory, our view loose the empirical support for the prediction-error theories, such as the support from the study by Corlett and colleagues

(2007). If the prediction-error theory is wrong, then how do we make sense of the study? When, on the other hand, we choose the prediction-error theory and deny the two-factor theory, we have difficulty in making sense of the asymmetry between people with the Capgras delusion and with the VMPFC damages. If there is no such thing as the second factor, then how do we make sense of the asymmetry? A hybrid strategy is free from these problems, since the hybrid view incorporates the core ideas of both theories. The hybrid theory can enjoy the theoretical virtues and empirical adequacy of the two-factor theory and the prediction-error theory at the same time. The theory, for instance, might be consistent with the Corlett study and make sense of the asymmetry between people with the Capgras delusion and with the VMPFC damages.

Second, both theories have some difficulties. As I already noted, a problem of the two-factor theory is about its scope. It is not clear how the theory is applied to polythematic delusions or motivated delusions. A problem of the prediction-error theory is about maintenance. Prediction-error theorists tend to focus on the stage of adoption. It is far from obvious that the prediction-error theory can say something informative about the stage of maintenance. A hybrid theorist might deal with these difficulties by combining the ideas from both groups of theories in a certain way. For instance, a possible idea is that a hybrid theory explains the maintenance stage by positing something other than the aberrant prediction-error signals as the second factor.

There are, broadly speaking, two possibilities. First, *the aberrant prediction-error signal is, or underlies, the first factor*. This means that the content of the relevant delusions is explained in terms of aberrant prediction-error signals. Second, *the aberrant prediction-error signal is, or underlies, the second factor*. This means that the adoption and maintenance of delusional hypotheses are explained in terms of aberrant prediction-error signals. (Please note that those possibilities are not exclusive to each other. Both can happen at the same time. In that case, *aberrant prediction-error signal is, or underlies, the first and the second factor*.)

The first idea, the idea that aberrant prediction-error signal is the first factor, seems to be an attractive option. According to the prediction-error theory, delusions are explanations of events that are inappropriately salient due to aberrant prediction-error signals. And, most two-factor theorists (explanationist two-factor theorists) think that delusions are explanations of the first factor[6]. For instance, Coltheart, Menzies and Sutton regard the probabilistic inference in the delusion adoption stage as an inference to the best *explanation* (of the first factor). Then, it is tempting to identify "events that are inappropriately salient due to aberrant prediction-error signals" with "the first factor".

---

[6] However, endorsement-type two-factor theory (e.g. Davies et al. 2001) maintains that delusions are not the explanation, but the endorsement of the content of abnormal experience.

This proposal has some potential benefits. First, it would solve the problem of dissociation (for the prediction-error theory). Since, in this proposal, aberrant prediction-error signal is merely the first factor, it is not expected to be sufficient for the development of delusion. People with VMPFC damage have the first factor, but they presumably do not have the second factor. Second, the problem of maintenance (for the prediction-error theory) might be solved if the second factor helps explain the maintenance of delusion. It is no longer an issue if an aberrant prediction-error signal does not satisfactorily explain maintenance. The second factor explains it. Third, the problem of scope (for the two-factor theory) might be solved if aberrant prediction-error signals explain the first factor in the formation of delusions whose first factors are unknown. For example, it is not clear how the two-factor theory is applied to the delusion of reference because there the first factor is unknown. But, according to this proposal, the first factor in the formation of delusions of reference might just be the inappropriate salience attached to some uninteresting events.

To say that the aberrant prediction-error signal is the first factor in the formation of some delusions is to say that it helps explain the content of those delusions. But how exactly can aberrant prediction-error signals be explanatory with regard to the content of delusions? The key would be the role of prediction error in attention allocation. Aberrant prediction-error signals render some events inappropriately attention-grabbing. Delusions

are formed as the explanations of those events. Here, aberrant prediction-error signals influence the allocation of attention, which is certainly explanatory with regard to the content of delusions that are formed as the explanations of the attention-grabbing events. Corlett and colleagues (2009) discuss a subject who formed the delusional belief that "The Organization" painted the doors of the houses on a street as a message for him. This was based on his observation that many doors on the right hand side were red, while many doors of the left side were green (Chadwick 2001). In this case, aberrant prediction-error signals render this uninteresting pattern of red doors and green doors inappropriately attention-grabbing, and this attention-grabbingness is explanatory with regard to the content of his delusion. It explains, for instance, why he thinks that the pattern of the color of the doors, as opposed to cars or trees on the street, conveys the message for him.

There are two possibilities to explore. One possibility is that the aberrant prediction-error signal underlies the first factor in the formation of some, but not all, delusions. For instance, one might think that inappropriate prediction-error signaling is the first factor in the formation of delusions of reference or persecutory delusions, but it has nothing to do with the first factor in the formation of monothematic delusions.

The other possibility is that the first factor in the formation of all delusions, including monothematic ones, has something to do with prediction error. Two-factor theorists often assume that the second factor is shared by all (monothematic) delusions, but the first factor

is different in different delusions. But, in our proposal, the first factor would also be shared by all delusions in the sense that it would have something to do with prediction error. This idea is attractive because, as Coltheart points out in the earlier quote, "the first factor in most delusions is a violation of expectation." As Corlett and colleagues (2010) suggest, the abnormal perceptual-affective experience, which is a good candidate for the first factor in the formation of the Capgras delusion, constitutes an *affective* prediction error. The same thing would be true about the Cotard delusion, whose first factor is often regarded as an abnormal affective experience about things in general. Corlett and colleagues argue that this abnormal experience also constitutes an affective prediction error. Again, delusions of control, thought control and thought insertion are very likely to be linked to prediction error. These delusions are likely to be generated by the failure to identify self-generated physical behavior or mental events as self-generated (Blakemore, Oakley & Frith 2003; Frith 2005). This failure can be easily linked to a prediction error because the most obvious feature that differentiates self-generated physical behavior and mental events from externally generated ones is their predictability. Due to the erroneous prediction error attributed to self-generated physical behavior and mental events, such events are mistakenly regarded as defying predictions and, hence, as externally generated.

Let us move on to the second idea that the aberrant prediction-error signal is, or underlies, the second factor. The potential advantage of this idea is that it might give us

some further insight into the nature of the second factor, which is still underspecified. The conciliatory strategy might help us to solve the main problem of the two-factor theory, namely, the problem of the second factor. Coltheart (2010) argues that the allergy detection task study by Corlett and colleagues is consistent with and supportive of the claim by him and colleagues that the second factor concerns the hypothesis evaluation system and is located in the right hemisphere. The study, according to Coltheart, is informative about the exact location of the second factor (right lateral prefrontal cortex) as well as the nature of it.

> "In the patients, surprising and unsurprising events *both* evoked responses of RLPFC, and to an equal degree. Hence *unsurprising* events evoked such activity in the patients when no such activity was evoked in the controls. The implication for two-factor theory, if one assumes that RLPFC is a signature of hypothesis *evaluation*, is that what is abnormal about the patients is that they are evaluating hypotheses even on those trials where the hypothesis was confirmed and so did not need evaluation." (Coltheart 2010, 24)

The general idea seems to be as follows: assuming that belief updating is operating on the principle of prediction-error minimization, excessive prediction-error signals in people with delusions would lead to revising beliefs that do not have to be revised. In short, people with delusions would be too revisionist when they update their beliefs. This bias might constitute the second factor. This idea is not entirely novel. Stone and Young (1997) already proposed that the second factor in delusion formation is a bias toward observational adequacy, which is due to putting too much weight on observational data and discounting existing beliefs.

Obviously, this bias makes people too revisionist in belief updating.

McKay (2012) develops Stone and Young's idea in detail and connects the proposal with the prediction-error theory. Unlike Coltheart, Menzies and Sutton (2010), McKay does not think that the inferential step in the delusion adoption stage is Bayesian-rational. The step deviates from Bayesian rationality and the deviation is characterized by the bias of discounting the prior probability ratio. Here is McKay's story about how people with the Capgras delusion reason.

$$\mathrm{P(Hs)} \ll P(\mathrm{Hw})$$
$$\mathrm{P(O|Hs)} \gg \mathrm{P(O|Hw)}$$
$$\mathrm{P(Hs|O)} > P(Hw|O)$$

Hw has a much higher prior probability than Hs. But, the likelihood ratio overwhelmingly favors Hs. Due to a bias, the prior probability ratio in favor of Hw is discounted, and the likelihood ratio in favour of Hs is overemphasized. As a result, Hs gets a higher posterior probability than Hw. This bias of discounting the prior probability ratio is the second factor. The difference between people with Capgras and people with VMPFC damage is explained by the assumption that this bias is peculiar to the former and absent in the latter. Unlike the proposal by Coltheart, Menzies and Sutton, there is no need to posit the bias of neglecting new evidence as the second factor, which might be problematic as I already noted.

McKay suggests that excessive prediction-error signals might underlie the bias of discounting the prior probability ratio.

> "Prediction error signals are triggered by discrepancies between the data expected and the data encountered. Such signals render salient the unexpected data and initiate a revision to accommodate these data. If there is an excess of prediction error signal, an appropriately heightened salience is attached to the data, and belief revision is excessively accomodatory – biased towards explanatory adequacy.[7]" (McKay 2012, 348)

This is certainly a very interesting proposal. However, one might want to add a twist to it. The distinction between prediction-error themselves and the precision of prediction-errors is crucial in the prediction-error framework. Prediction-errors are the indicators of the mismatch between predictions and actual inputs. The precision of prediction-errors are, on the other hand, the indicator of the reliability or the trustworthiness of the prediction-errors. Prediction errors with high precision are regarded as reliable and trustworthy and, hence, have huge impact in prediction-error processing. In particular, they tend to cause the revision of prior beliefs. Prediction-errors with low precision, on the other hand, are regarded as unreliable and untrustworthy and, hence, do not have much impact in prediction-error processing. They tend to be neglected.

---

[7] For McKay, "bias toward explanatory adequacy" is another name for the bias of neglecting prior probability ratio.

With the distinction between prediction-errors and the precision of prediction-errors at hand, we might want to argue that what is responsible for the bias of discounting prior probability ratio is, strictly speaking, not the excess of prediction-error themselves but rather the excessive precision of prediction-errors. There are some reasons to accept this proposal.

First, the people with the Capgras delusion and with VMPFC damage might be equivalent with respect to relevant prediction-errors. This is, at least, Coltheart's view (2010); there are the mismatches between the expected and the actual autonomic activities in both groups. If this is the case, then, assuming McKay's view that prediction-errors are driving the bias of discounting prior probability ratio, we need to conclude that both the people with the Capgras delusion and with VMPFC damages exhibit the bias of discounting prior probability ratio, which contradicts McKay's own view. Second, precision is exactly the kind of factor that can drive the bias of discounting prior probability ratio. The bias of discounting prior probability ratio is the bias concerning the balance between new observations (or likelihood) and existing beliefs (or prior probabilities). And, precision is exactly the kind of factor that determines the balance. In particular, when some empirical prediction-errors are regarded as very precise (or when relevant prior beliefs are regarded as very imprecise) the empirical prediction-errors will be heavily weighed in the process of updating beliefs. This is, in effect, the bias of discounting prior probability ratio.

The proposal here, then, is that the second factor is the bias of discounting prior probability ratio driven by the high (relative to prior beliefs) precision of the relevant prediction-errors. The proposal has some interesting implications about the nature of the bias.

The first implication is about the rationality of the bias. McKay's view is that the bias is irrational, say, from a Bayesian point of view. Bayes' rule requires that we treat likelihoods and prior probabailities equally. The bias seems to be the violation from the requirement. According to the current proposal, the bias is not strictly speaking irrational as long as we think that what we calculate are not prediction-errors per se but precision-weighed prediction-errors. There is nothing irrational in emphasizing the prediction-errors that are highly precise. In a sense, then, this proposal shares an important idea (namely, the idea that delusions are formed as the rational response to the first factors) with Maher and Coltheart who are, in in fact, the opponents of McKay's view.

The second implication is about the generality of the bias. In McKay's view, the second factor is a general reasoning bias that has pervasive influence on the way people reason quite generally probably even outside the context of delusions. Let us call it the "global bias of discounting probability ratio." The current proposal, however, does not have to be committed to this idea. The reason is that it is at least theoretically possible that some particular prediction-errors in some particular context are very precise, which means that it

is at least theoretically possible that one exhibits the bias of discounting prior probability ratio with respects to some particular prediction-errors in some particular contexts. The hybrid view, thus, allows for the possibility of the "local bias of discounting probability ratio."

A possible benefit of the local conception of the bias is that it is free from some worries about the global bias hypothesis.

A worry about the global bias hypothesis is related to the specificity of monothematic delusions (Coltheart et al. 2011). The global bias has the pervasive influence on the way people reason generally, even outside the context of delusions. Then, the global bias hypothesis seems to predict that people with delusions can have strange or abnormal beliefs about wide range of topics. People with monothematic delusions, however, do not have strange or abnormal beliefs outside the topics of their delusions. As Davies and Egan put it, monothematic delusions are "islands of delusion in a sea of apparent normality" (Davies & Egan 2013, 690).

Relatedly, the global bias hypothesis seems to predict that people with delusions have stronger tendency to be fooled by perceptual illusions (e.g., Muller-Lyer illusion) and to form incorrect beliefs about them than normal individuals. For instance, Muller-Lyer illusion provides the observation that two lines are not equal in length. With the global bias of discounting prior probability ratio, individuals with delusions overemphasize the

observation, which means that they tend to believe that two lines are not equal in length. But, this prediction might not be true. Davies et al. call it "an unwanted prediction " and say "we would prefer an account of the second factor that avoided this prediction" (Davies et al. 2001, 153).

The local bias hypothesis is free from the worries about the specificity and perceptual illusions because the hypothesis allows for the possibility that the influence of the bias is only seen with respect to some particular inputs in some particular contexts. Unlike the global bias hypothesis, the local bias hypothesis does not predict that people with delusions can have strange beliefs about wide range of topics. It might turn out that people with delusions show the bias of discounting prior probability ratio with respect to some particular strange inputs (such as the abnormal data about familiar faces), but do not show the bias in other cases. For the same reason, the hypothesis does not predict that they have strong tendency to be fooled by visual illusions.

6.5 Conclusion

In this chapter, I reviewed the basic commitments of the two-factor as well as the prediction-error theory and discussed the relationship between the two. I acknowledged that there are some important differences between the two. They have been developed as

accounts for different kinds of delusions. The prediction-error theory is often presented as a one-factor account, whereas the two-factor theory is committed to two distinct factors. Prediction-error theorists tend to blur the distinction between perceiving and believing, whereas two-factor theorists tend to draw a sharp line there. However, when we examine these differences further, we realize that they are not as significant as one might expect. Moreover, it is possible to incorporate the basic ideas of the prediction-error theory into the two-factor framework.

This hybrid strategy might offer some interesting solutions to the existing challenges that the two theories face when considered in isolation. The aberrant prediction-error signal might explain the content of the delusion via its impact on attention allocation, in which case it could underlie the first factor in delusion formation. Alternatively, aberrant prediction-error signals might explain the adoption and maintenance of delusional hypotheses by making belief updating excessively revisionist, in which case it could underlie the second factor in delusion formation. If both theories of delusion formation, the two-factor theory and the prediction-error theory, continue to be largely supported by the empirical evidence and provide plausible enough explanations for the phenomenon of delusion formation, then the conciliatory strategy strikes us as appealing. A good story about delusion formation will probably need to incorporate ideas from both.

Reference

- American Psychiatric Association. (2013) *Diagnostic and Statistical Manual of Mental Disorders.* 5th edition, American Psychiatric Publishing.
- Anselmetti, S., Cavallaro, R., Bechi, M., Angelone, S. M., Ermoli, E., Cocchi, F., et al. (2007). Psychopathological and neuropsychological correlates of source monitoring impairment in schizophrenia. *Psychiatry Research*, *150*(1), 51-59.
- Armstrong, D. M. (2002). *A Materialist Theory of the Mind*. Routledge.
- Bayne, T. (2010). Delusions as doxastic states: Contexts, compartments, and commitments. *Philosophy, Psychiatry, & Psychology*, *17*(4), 329-336.
- Bayne, T., & Fernández, J. (2009). Delusion and self-deception: Mapping the terrain. in T. Bayne & J. Fernández (eds.) *Delusion and Self-deception: Affective and Motivational Influences on Belief Formation*, Psychology Press. 1-21.
- Bayne, T., & Hattiangadi, A. (2013). Belief and its bedfellows. in N. Nottelmann (ed.) *New Essays on Belief: Constitution, Content and Structure*, 124-144.
- Bayne, T., & Pacherie, E. (2005). In defence of the doxastic conception of delusions. *Mind & Language*, *20*(2), 163-188.
- Bentall, R. P., & Kaney, S. (1996). Abnormalities of self-representation and persecutory delusions: A test of a cognitive model of paranoia. *Psychological Medicine*, *26*(6), 1231-1238.
- Berrios, G. E. (1991). Delusions as "wrong beliefs": A conceptual history. *The British Journal of Psychiatry*, *159*(suppl. 14), 6-13.
- Berthier, M., Starkstein, S., & Leiguarda, R. (1988). Asymbolia for pain: A sensory-limbic disconnection syndrome. *Annals of Neurology*, *24*(1), 41-49.
- Bigelow, J., & Pargetter, R. (1987). Functions. *Journal of Philosophy*, *84*, 181-196.
- Bisiach, E., & Geminiani, G. (1991). Anosognosia related to hemiplegia and hemianopia. in G. P. Prigatano & D. L. Shacter (eds.) *Awareness of Deficit After Brain Injury*, Oxford University Press, 17-39.
- Blakemore, S., Oakley, D. A., & Frith, C. (2003). Delusions of alien control in the normal brain. *Neuropsychologia*, *41*(8), 1058-1067.

- Bleuler, E. (1924). *Textbook of Psychiatry*. J. Zinkin (transl.) International Universities Press.
- Bleuler, E. (1950). *Dementia praecox or the Group of Schizophrenias*. A. A. Brill (transl.) Macmillan.
- Block, N. (1998). Is experiencing just representing?. *Philosophical and Phenomenological Research*, *58*(3), 663-670.
- Bloom, P. (2012). Religion, morality, evolution. *Annual Review of Psychology*, *63*, 179-199.
- Boorse, C. (1976). Wright on functions. *The Philosophical Review*, *85*, 70-86.
- Boorse, C. (2002). A rebuttal on functions. in A. Arew (ed.) *Functions: New essays in the Philosophy of Psychology and Biology*, Oxford Uniersity Press, 63-112.
- Bortolotti, L. (2010). *Delusions and Other Irrational Beliefs*. Oxford University Press.
- Bortolotti, L. (2011). Double bookkeeping in delusions: Explaining the gap between saying and doing. in J. Aguilar, A. Buckareff, & K. Frankish (eds.) *New Waves in the Philosophy of Action*, 237-256.
- Bortolotti, L. (2012). In defence of modest doxasticism about delusions. *Neuroethics*, *5*(1), 39-53.
- Bortolotti, L., & Broome, M. R. (2012). Affective dimensions of the phenomenon of double bookkeeping in delusions. *Emotion Review*, *4*(2), 187-191.
- Bovet, P., & Parnas, J. (1993). Schizophrenic delusions: a phenomenological approach. *Schizophrenia Bulletin*, *19*(3), 579.
- Boyer, P. (2001). *Religion Explained: The Evolutionary Origins of Religious Thought*. Basic Books.
- Boyer, P. (2003). Religious thought and behaviour as by-products of brain function. *Trends in Cognitive Sciences*, *7*(3), 119-124.
- Brakoulias, V., Langdon, R., Sloss, G., Coltheart, M., Meares, R., & Harris, A. (2008). Delusions and reasoning: A study involving cognitive behavioural therapy. *Cognitive Neuropsychiatry*, *13*(2), 148-165.
- Brébion, G., Amador, X., David, A., Malaspina, D., Sharif, Z., & Gorman, J. M. (2000). Positive symptomatology and source-monitoring failure in schizophrenia—an analysis of symptom-specific effects. *Psychiatry Research, 95*(2), 119-131.
- Buller, D. J. (1998). Etiological theories of function: A geographical survey. *Biology*

and Philosophy, *13*(4), 505-527.

- Buller, D. J. (2005). *Adapting Minds: Evolutionary Psychology and the Persistent Quest for Human Nature*. The MIT press.
- Buss, D. M., Larsen, R. J., Westen, D., & Semmelroth, J. (1992). Sex differences in jealousy: Evolution, physiology, and psychology. *Psychological Science*, *3*(4), 251-255.
- Butler, P. V. (2000). Reverse Othello syndrome subsequent to traumatic brain injury. *Psychiatry*, *63*(1), 85-92.
- Campbell, J. (2001). Rationality, meaning, and the analysis of delusion. *Philosophy, Psychiatry, & Psychology*, *8*(2), 89-100.
- Chadwick, P. K. (2001). Psychotic consciousness. *International Journal of Social Psychiatry*, *47*(1), 52-62.
- Christodoulou, G. (1977). The syndrome of Capgras. *The British Journal of Psychiatry*, *130*(6), 556-564.
- Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy*, *78*, 67-90.
- Clark, T., Day, E., & Fergusson, E. C. (2005) *Core Clinical Cases in Psychiatry*. Hodder Arnold.
- Coltheart, M. (2007). Cognitive neuropsychiatry and delusional belief: The 33rd Sir Frederick Bartlett lecture. *The Quarterly Journal of Experimental Psychology*, *60*(8), 1041-1062.
- Coltheart, M. (2010). The neuropsychology of delusions. *Annals of the New York Academy of Sciences*, *1191*(1), 16-26.
- Coltheart, M. (2013). On the distinction between monothematic and polythematic delusions. *Mind & Language*, *28*(1), 103-112.
- Coltheart, M., Langdon, R., & McKay, R. (2007). Schizophrenia and monothematic delusions. *Schizophrenia Bulletin*, *33*(3), 642-647.
- Coltheart, M., Langdon, R., & McKay, R. (2011). Delusional belief. *Annual review of psychology*, *62*, 271-298.
- Coltheart, M., Menzies, P., & Sutton, J. (2010). Abductive inference and delusional belief. *Cognitive Neuropsychiatry*, *15*(1-3), 261-287.
- Corlett, P. R. (2012). Latent inhibition: Cognition, neuroscience and applications to schizophrenia: The art of explaining psychosis. *Cognitive Neuropsychiatry*, *17*(3),

287-289.

- Corlett, P. R., Cambridge, V., Gardner, J. M., Piggot, J. S., Turner, D. C., Everitt, J. C., et al. (2013). Ketamine effects on memory reconsolidation favor a learning model of delusions. *PloS one*, *8*(6), e65088.
- Corlett, P. R., Krystal, J. H., Taylor, J. R., & Fletcher, P. C. (2009). Why do delusions persist?. *Frontiers in Human Neuroscience*, *3*, 12.
- Corlett, P., Murray, G., Honey, G., Aitken, M., Shanks, D., Robbins, T., et al. (2007). Disrupted prediction-error signal in psychosis: evidence for an associative account of delusions. *Brain*, *130*(9), 2387-2400.
- Corlett, P., Taylor, J., Wang, X., Fletcher, P., & Krystal, J. (2010). Toward a neurobiology of delusions. *Progress in Neurobiology*, *92*(3), 345-369.
- Cummins, R. (1975). Functional analysis. *Journal of Philosophy*, 72, 741-764.
- Cummins, R. C. (1983). *The Nature of Psychological Explanation*. The MIT Press
- Cummins, R., & Roth, M. (2010). Traits have not evolved to function the way they do because of a past advantage. in F. J. Ayala & R. Art (eds.) *Contemporary debates in Philosophy of Biology*, 72-88.
- Currie, G. (2000). Imagination, delusion and hallucinations. *Mind & language*, *15*(1), 168-183.
- Currie, G., & Jureidini, J. (2001). Delusion, Rationality, Empathy: Commentary on Martin Davies et al. *Philosophy, Psychiatry, & Psychology*, *8*(2), 159-162.
- Currie, G., & Jureidini, J. (2003). Art and delusion. *The Monist*, *86*(4), 556-578.
- Currie, G., & Ravenscroft, I. (2002). *Recreative Minds: Imagination in Philosophy and Psychology*. Oxford University Press.
- Daly, M., Wilson, M., & Weghorst, S. J. (1982). Male sexual jealousy. *Ethology and Sociobiology*, *3*(1), 11-27.
- Davidson, D. (1984). *Essays on Truth and Interpretation*. Oxford University Press.
- Davidson, D. (1987). Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association*, *61*, 441-458..
- Davies, M., Coltheart, M., Langdon, R., & Breen, N. (2001). Monothematic delusions: Towards a two-factor account. *Philosophy, Psychiatry, & Psychology*, *8*(2), 133-158.
- Davies, M., & Egan, A. (2013). Delusion: Cognitive approaches Bayesian inference and compartmentalisation. in K. W. M. Fulford, M. Davies, R. G. T. Gipps, & G. Graham (eds.) *Oxford Handbook of Philosophy and Psychiatry*. Oxford University

Press, 689-727.

- Davies, P. S. (2000). Malfunctions. *Biology and philosophy*, *15*(1), 19-38.
- Davies, P. S. (2003). *Norms of Nature: Naturalism and the Nature of Functions*. The MIT Press.
- Dennett, D. C. (1989). *The Intentional Stance*. The MIT press.
- Dennett, D. C. (1991). *Consciousness Explained*. Penguin Press.
- Descartes, R. (1649/1985). *The Passions of the Soul*. in J. Cottingham, R. Stoothoff, & D. Murdoch (eds.), *The Philosophical Writings of Descartes*, Vol. 1. Cambridge University Press, 325-404.
- Dretske, F. (1986). Misrepresentation. in R. Bogdan (ed) *Belief: Form, Content and Function*, Oxford University Press, 17–36.
- Dretske, F. I. (1991). *Explaining Behavior: Reasons in A World of Causes*. The MIT press.
- Dretske, F. I. (1997). *Naturalizing the Mind*. The MIT Press.
- Drury, V., Birchwood, M., & Cochrane, R. (2000). Cognitive therapy and recovery from acute psychosis: a controlled trial 3. Five-year follow-up. *The British Journal of Psychiatry*, *177*(1), 8-14.
- Easton, J. A., Schipper, L. D., & Shackelford, T. K. (2007). Morbid jealousy from an evolutionary psychological perspective. *Evolution and Human Behavior*, *28*(6), 399-402.
- Egan, A. (2008). Imagination, delusion, and self-deception. in T. Bayne and J. Fernández (eds.) *Delusions and Self-deception; Motivational and Affective Influences on Belief Formation*, 263-280.
- Ellis, H. D., & Young, A. W. (1990). Accounting for delusional misidentifications. *The British Journal of Psychiatry*, *157*(2), 239-248.
- Ellis, H. D., Young, A. W., Quayle, A. H., & De Pauw, K. W. (1997). Reduced autonomic responses to faces in Capgras delusion. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *264*(1384), 1085-1092.
- Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, *10*(1), 48-58.
- Fodor, J. A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. The MIT Press.

- Fodor, J. A. (2001). *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. The MIT press.
- Foussias, G., & Remington, G. (2010). Negative symptoms in schizophrenia: avolition and Occam's razor. *Schizophrenia Bulletin*, *36*(2), 359-369.
- Fowler, D., Garety, P., & Kuipers, L. (1995). *Cognitive Behaviour Therapy for Psychosis: Theory and Practice*. Wiley.
- Frankish, K. (2009). Delusions: A two-level framework. in M. Broome & L. Bortolotti (eds.) *Psychiatry as Cognitive Neuroscience: Philosophical Perspectives*, Oxford University Press. 269-284.
- Frankish, K. (2012). Delusions, Levels of belief, and non-doxastic acceptances. *Neuroethics*, *5*(1), 23-27.
- Frith, C. D. (1992). *The Cognitive Neuropsychology of Schizophrenia*. Psychology Press.
- Frith, C. (2005). The neural basis of hallucinations and delusions. *Comptes Rendus Biologies*, *328*(2), 169-175.
- Gendler, T. S. (2000). The puzzle of imaginative resistance. *Journal of Philosophy*, 55-81.
- Ginet, C. (2001). Deciding to believe. in M. Steup (ed.) *Knowledge, Truth, and Duty*, Oxford University Press, 63-76.
- Godfrey-Smith, P. (1989). Misinformation. *Canadian Journal of Philosophy*, *19*(4), 533-550.
- Godfrey-Smith, P. (1993). Functions: consensus without unity. *Pacific Philosophical Quarterly*, *74*(3), 196-208.
- Godfrey-Smith, P. (1994). A modern history theory of functions. *Noûs*, 28(3), 344-362.
- Godfrey-Smith, P. (1998). *Complexity and the Function of Mind in Nature*. Cambridge University Press.
- Godfrey-Smith, P. (2006). Mental representation, naturalism, and teleosemantics. in G. MacDonald & D. Papineau (eds.) *Teleosemantics. Oxford University Press*, 42-68.
- Goldman, A. I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press.
- Gould, S. J. (1991). Exaptation: A crucial tool for an evolutionary psychology. *Journal of Social Issues*, *47*(3), 43-65.

- Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *205*(1161), 581-598.
- Gould, S. J., & Vrba, E. S. (1982). Exaptation – a missing term in the science of form. *Paleobiology*, 4-15.
- Guthrie, S. (1995). *Faces in the Clouds*. Oxford University Press.
- Haselton, M. G., & Buss, D. M. (2000). Error management theory: a new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology*, *78*(1), 81.
- Hemsley, D. R., & Garety, F. (1997). *Delusions: Investigations into the Psychology of Delusional Reasoning*. Psychology Press.
- Helmholtz, H. (1878/1968). The facts of perception. in R. P. Warren (ed.) *Helmhotz on Perception: Its Physiology and Development*, 205-231.
- Hirstein, W., & Ramachandran, V. S. (1997). Capgras syndrome: a novel probe for understanding the neural representation of the identity and familiarity of persons. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *264*(1380), 437-444.
- Holton, R., & Berridge, K. (2013). Addiction between compulsion and choice. in N. Levy (ed.) *Addiction and Self-Control: Perspectives from Philosophy, Psychology, and Neuroscience*, Oxford University Press, 239-268.
- Hohwy, J., & Rajan, V. (2012). Delusions as forensically disturbing perceptual inferences. *Neuroethics*, *5*(1), 5-11.
- Hume, D. (1748/2007). An Enquiry Concerning Human Understanding. P. Millican (ed.) Oxford University Press.
- Huq, S., Garety, P., & Hemsley, D. (1988). Probabilistic judgements in deluded and non-deluded subjects. *The Quarterly Journal of Experimental Psychology*, *40*(4), 801-812.
- Jenkinson, P. M., Edelstyn, N. M., Drakeford, J. L., & Ellis, S. J. (2009). Reality monitoring in anosognosia for hemiplegia. *Consciousness and Cognition*, *18*(2), 458-470.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237.
- Kapur, S. (2003). Psychosis as a state of aberrant salience: a framework linking

biology, phenomenology, and pharmacology in schizophrenia. *American Journal of Psychiatry*, *160*(1), 13-23.

- Kinderman, P., & Bentall, R. P. (1996). Self-discrepancies and persecutory delusions: evidence for a model of paranoid ideation. *Journal of Abnormal Psychology*, *105*(1), 106.
- Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*, 190-194.
- Kripke, S. A. (2011). *Philosophical Troubles: Collected Papers*, Vol. 1. Oxford University Press.
- Lewis, D. (1980). Mad pain and Martian pain. in N. Block (ed.) *Readings in the Philosophy of Psychology*, *1*, Harvard University Press, 216-222.
- Lucchelli, F., & Spinnler, H. (2007). The case of lost Wilma: a clinical report of Capgras delusion. *Neurological Sciences*, *28*(4), 188-195.
- Lycan, W. G. (1982). Toward a homuncular theory of believing. *Cognition and Brain Theory*, *4*, 139-159.
- Lycan, W. G. (1995). *Consciousness*. The MIT Press.
- Maher, B. A. (1974). Delusional thinking and perceptual disorder. *Journal of Individual Psychology*, *30*, 98-113.
- Marušić, J. S. (2010). Does Hume hold a dispositional account of belief?. *Canadian Journal of Philosophy*, *40*(2), 155-183.
- McKay, R. (2012). Delusional inference. *Mind & Language*, *27*(3), 330-355.
- McKay, R., & Cipolotti, L. (2007). Attributional style in a case of Cotard delusion. *Consciousness and Cognition*, *16*(2), 349-359.
- McKay, R. T., & Dennett, D. C. (2009). The evolution of misbelief. *Behavioral and Brain Sciences*, *32*(6), 493-510.
- McKay, R., Langdon, R., & Coltheart, M. (2005). "Sleights of mind": Delusions, defences, and self-deception. *Cognitive Neuropsychiatry*, *10*(4), 305-326.
- McKay, R., Langdon, R., & Coltheart, M. (2007). Models of misbelief: Integrating motivational and deficit theories of delusions. *Consciousness and Cognition*, *16*(4), 932-941.
- Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories: New Foundations for Realism*. The MIT press.
- Millikan, R. G. (1989a). In defense of proper functions. *Philosophy of Science*, *56*(2),

288.

- Millikan, R. G. (1989b). Biosemantics. *Journal of Philosophy*, 86, 281-297.
- Millikan, R. G. (1995). Pushmi-pullyu representations. *Philosophical Perspectives*, 9, 185-200.
- Millikan, R. G. (1996). On swampkinds. *Mind & Language*, *11*(1), 103-117.
- Millikan, R. G. (2002). Biofunctions: Two paradigms. in A. Arew (ed.) *Functions: New essays in the Philosophy of Psychology and Biology*, Oxford Uniersity Press, 113-143.
- Millikan, R. G. (2004). *Varieties of Meaning: the 2002 Jean Nicod Lectures*. The MIT press.
- Murphy, D. (2012). The folk epistemology of delusions. *Neuroethics*, *5*(1), 19-22.
- Murphy, D., & Woolfolk, R. L. (2000). The harmful dysfunction analysis of mental disorder. *Philosophy, Psychiatry, & Psychology*, *7*(4), 241-252.
- Murphy, D., & Woolfolk, R. L. (2000). Conceptual analysis versus scientific understanding: An assessment of Wakefield's folk psychiatry. *Philosophy, Psychiatry, & Psychology*, *7*(4), 271-293.
- Murray, G. (2011). The emerging biology of delusions. *Psychological Medicine*, *41*(1), 7.
- Nanay, B. (2010). A modal theory of function. *Journal of Philosophy*, *107*(8), 412.
- Neander, K. (1991a). Functions as selected effects: The conceptual analyst's defense. *Philosophy of Science*, *58*(2), 168.
- Neander, K. (1991b). The teleological notion of 'function'. *Australasian Journal of Philosophy*, *69*(4), 454-468.
- Neander, K. (1995). Misrepresenting & malfunctioning. *Philosophical Studies*, *79*(2), 109-141.
- Neander, K. (1996). Swampman meets swampcow. *Mind & Language*, *11*(1), 118-129.
- Neander, K. (2013). Toward an informational teleosemantics. in D. Ryder, J. Kingsbury, & K. Williford. (2012). *Millikan and Her Critics*. John Wiley & Sons, 21-36.
- Nesse, R. M. (1998). Emotional disorders in evolutionary perspective. *British Journal of Medical Psychology*, *71*(4), 397-415.
- Nesse, R. M. (1990). Evolutionary explanations of emotions. *Human Nature*, *1*(3),

261-289.

- Nesse, R. M. (2001). The smoke detector principle. *Annals of the New York Academy of Sciences*, *935*(1), 75-85.
- Nichols, S. (2004). Imagining and believing: The promise of a single code. *The Journal of Aesthetics and Art Criticism*, *62*(2), 129-139.
- Nichols, S., & Stich, S. (2000). A cognitive theory of pretense. *Cognition*, *74*(2), 115-147.
- Nordenfelt, L. (2007). The concepts of health and illness revisited. *Medicine, Health Care and Philosophy*, *10*(1), 5-10.
- Oatley, K., & Johnson-Laird, P. N. (1987). Towards a cognitive theory of emotions. *Cognition and Emotion*, *1*(1), 29-50.
- Okasha, S. (2003). Fodor on cognition, modularity, and adaptationism. *Philosophy of Science*, *70*(1), 68-88.
- Papineau, D. (1984). Representation and explanation. *Philosophy of Science*, 51, 550-572.
- Papineau, D. (2001). The status of teleosemantics, or how to stop worrying about swampman. *Australasian Journal of Philosophy*, *79*(2), 279-289.
- Pinker, S. (1997). *How the Mind Works*. Norton.
- Price, H. H. (1970). *Belief*. Allen & Unwin..
- Prior, E. W. (1985). What is wrong with etiological accounts of biological function?. *Pacific Philosophical Quarterly*, *66*(3-4), 310-328.
- Pyysiäinen, I., & Hauser, M. (2010). The origins of religion: evolved adaptation or by-product?. *Trends in Cognitive Sciences*, *14*(3), 104-109.
- Ramsey, F. P. (1931). *The Foundations of Mathematics and Other Logical Essays*. Routledge.
- Reimer, M. (2010). Only a philosopher or a madman: Impractical delusions in philosophy and psychiatry. *Philosophy, Psychiatry, & Psychology*, *17*(4), 315-328.
- Roe, K., & Murphy, D. (2011). Function, dysfunction, and adaptation?. in P.R. Adriaens & A. De Block (eds.) *Maladapting Minds: Philosophy, Psychiatry, and Evolutionary Theory*, 216-237.
- Ryan, S. (2003). Doxastic compatibilism and the ethics of belief. *Philosophical Studies*, *114*(1), 47-79.
- Ryder, D., Kingsbury, J., & Williford, K. (2012). *Millikan and Her Critics*. Wiley.

- Ryle, G. (1949). *The Concept of Mind*. Hutchinson.
- Sass, L. A. (1994). *The Paradoxes of Delusion: Wittgenstein, Schreber, and the Schizophrenic Mind*. Cornell University Press.
- Schwitzgebel, E. (2010). Belief. *Stanford Encyclopaedia of Philosophy*. http://plato.stanford.edu/entries/belief/#2.5
- Schwitzgebel, E. (2001). In-between believing. *The Philosophical Quarterly*, *51*(202), 76-82.
- Schwitzgebel, E. (2002). A phenomenal, dispositional account of belief. *Noûs*, *36*(2), 249-275.
- Sober, E. (1985). Panglossian functionalism and the philosophy of mind. *Synthese*, *64*(2), 165-193.
- Sterelny, K. (1990). *The Representational Theory of Mind*. Blackwell.
- Stich, S. P. (1990). *The Fragmentation of Reason: Preface to a Pragmatic Theory of Cognitive Evaluation*. The MIT Press.
- Stone, T., & Young, A. W. (1997). Delusions and brain injury: The philosophy and psychology of belief. *Mind & Language*, *12*(3-4), 327-364.
- Szasz, T. S. (1960). The myth of mental illness. *American Psychologist*, *15*(2), 113.
- Cosmides, L., & Tooby, J. (2000). Evolutionary psychology and the emotions. in M. Lewis & J. M. Haviland-Jones (eds.) *Handbook of Emotions*, Vol. *2*, Guilford Press, 91-115.
- Tengland, P. (2001). *Mental health: A Philosophical Analysis*. Springer.
- Tumulty, M. (2011). Delusions and dispositionalism about belief. *Mind & Language*, *26*(5), 596-628.
- Tumulty, M. (2012). Delusions and not-quite-beliefs. *Neuroethics*, *5*(1), 29-37.
- Tranel, D., Damasio, H., & Damasio, A. R. (1995). Double dissociation between overt and covert face recognition. *Journal of Cognitive Neuroscience*, *7*(4), 425-432.
- Tye, M. (1998). Inverted earth, swampman, and representationism. *Noûs*, *32*(S12), 459-477.
- Wakefield, J. C. (1992a). The concept of mental disorder: on the boundary between biological facts and social values. *American Psychologist*, *47*(3), 373-388.
- Wakefield, J. C. (1992b). Disorder as harmful dysfunction: A conceptual critique of DSM-III-R's definition of mental disorder. *Psychological review*, *99*(2), 232-247.
- Wakefield, J. C. (1999a). Evolutionary versus prototype analyses of the concept of

disorder. *Journal of Abnormal Psychology*, *108*(3), 374-399.

- Wakefield, J. C. (1999b). Mental disorder as a black box essentialist concept. *108*(3), 465-472.
- Wakefield, J. C. (2000). Spandrels, Vestigial Organs, and Such: Reply to Murphy and Woolfolk's" The Harmful Dysfunction Analysis of Mental Disorder." *Philosophy, Psychiatry, & Psychology*, *7*(4), 253-269.
- Wakefield, J. C. (2011). Darwin, functional explanation, and the philosophy of psychiatry. in P.R. Adriaens & A. De Block (eds.) *Maladapting Minds: Philosophy, Psychiatry, and Evolutionary Theory*, 43-172.
- Wallis, G. (1986). Nature of the misidentified in the Capgras syndrome. *Bibliotheca Psychiatrica*, *164*, 40.
- Young, A. W. (2000). Wondrous strange: The neuropsychology of abnormal beliefs. *Mind & language*, *15*(1), 47-73.