Constructing and Applying Explanatory Models in Clinical Psychology

Von der Philosophischen Fakultät der Gottfried Wilhelm
Leibniz Universität Hannover

zur Erlangung des Grades Doktorin der Philosophie (Dr. phil.)
genehmigte Dissertation

von Julia Salome Pfeiff, B.A., M.Sc.

Erscheinungs- bzw. Druckjahr 2021

Referentin	Univ.-Prof. Dr. Uljana Feest
Korreferentin	Univ.-Prof. Dr. Marie Kaiser
Tag der Promotion	23.07.2020

Affidavit

Ich versichere hiermit an Eides statt, dass:

- ich die vorliegende Dissertation selbstständig angefertigt habe;
- ich keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe;
- ich diese Dissertation nicht schon früher als Prüfungsarbeit verwendet habe.

Unterschrift:

Datum:

Abstract

This thesis deals with the following question: How are clinical psychology's explanatory models of mental disorders constructed and how are they utilized in psychotherapy? *Clinical psychology* is understood as an applied scientific discipline that is directed at developing treatments (e.g., psychotherapy) for mental disorders and at understanding them. *Explanatory models* are important objects of research that are also used in therapeutic practice. I investigate two exemplary models, one model of *Major Depressive Disorder*, and one model of *Obsessive-Compulsive Disorder.* These models, as I argue in chapter 1, are intended to account for the disorder's *etiology* and *maintenance*, utilizing a mixture of functional, causal and folk-psychological vocabulary. In chapter 2, a comparative analysis of earlier and later versions of these two models shows that (1) clinical observations play a major role in how these models are formulated and (2) considerations of applicability in psychotherapy constitute an important influence in how they changed over time. In chapter 3, I argue that explanatory practices in psychotherapy are intended to make the patient not feel at fault for developing their disorder but nonetheless take over responsibility for their symptoms in the future. This overarching goal is connected to three sub-goals, namely, (1) attributing limited responsibility for falling ill to the patient, (2) attributing agency to her, and (3) presenting possibilities for intervention. In chapter 4, I argue that achieving these aims is connected to using two different notions of rationality, namely, (1) *theoretical rationality* and (2) *pragmatic rationality.* While, according to (1), someone is rational just in case she adopts beliefs that cohere with her relevant background knowledge, according to (2), she is rational just in case she adopts beliefs that square well with her considered goals. These notions allow for an enlightening reconstruction of disputation techniques found in therapeutic practice. In chapter 5, I similarly engineer a notion of *dysfunctionality* that is compatible with an understanding of a patient being rational in holding a particular dysfunctional belief. According to this notion, something is dysfunctional for an agent just in case it causally counteracts her needs and produces significant harm as a result. In chapter 6, I draw out the most important results of my work.

Keywords: philosophy of psychology, models of mental disorders, dysfunctionality

Abstract

Diese Dissertation beschäftigt sich mit der folgenden Frage: Wie werden die Modelle psychischer Störungen in der klinischen Psychologie konstruiert, und wie werden sie in der Psychotherapie angewendet? *Klinische Psychologie* wird als eine angewandte wissenschaftliche Disziplin verstanden, die darauf ausgerichtet ist, Behandlungsmöglichkeiten (etwa Psychotherapie) für psychische Störungen zu entwickeln und diese Störungen zu verstehen. *Erklärungsmodelle* sind wichige Objekte der wissenschaftlichen Forschung, die auch in der therapeutischen Praxis angewandt werden. Ich untersuche zwei exemplarische Modelle, ein Modell der *Depression* und eines der *Zwangsstörung*. Diese Modelle, wie ich in Kapitel 1 ausführe, sind dazu intendiert, für die *Ätiologie* und die *Aufrechterhaltung* der Störung aufzukommen. Dazu wird eine Mischung aus funktionalem, kausalem und alltagspsychologischen Vokabular verwendet. In Kapitel 2 zeigt eine komparative Analyse von früheren und späteren Versionen der beiden Modelle, dass (1) klinische Beobachtungen eine große Rolle darin spielen, wie diese Modelle formuliert sind und (2) Überlegungen der Anewendbarkeit in der Psychotherapie beeinflussen, wie diese Modelle sich über die Zeit verändert haben. In Kapitel 3 argumentiere ich, dass Erklärungspraktiken in der Psychotherapie dazu intendiert sind, der Patientin den Eindruck zu vermitteln, dass sie keine Schuld daran trägt, ihre Störung entwickelt zu haben, sie aber trotzdem davon zu überzeugen, für ihre Symptome in der Zukunft Verantwortung zu übernehmen. Dieses übergreifende Ziel ist mit drei Teilzielen verknüpft, (1) der Patientin nur eingeschränkte Verantwortlichkeit dafür, krank geworden zu sein, zuzuschreiben, (2) ihr Handlungsfähigkeit zuzuschreiben, und (3) Interventionsmöglichkeiten aufzuzeigen. In Kapitel 4 zeige ich, wie das Erreichen dieser Ziele mit zwei verschiedenen Begriffen von Rationalität verknüpft ist, (1) theoretische Rationalität und (2) pragmatische Rationalität. Während jemand laut (1) dann rational ist, wenn sie Überzeugungen annimmt, die mit ihren Hintergrundüberzeugungen kohärent sind, ist sie laut (2) rational genau dann, wenn sie Überzeugungen annimmt, die gut zu ihren expliziten Zielen passen. Diese Begriffe erlauben eine einleuchtende Analyse von Disputationstechniken, die in der therapeutischen Praxis verwendet werden. In Kapitel 5 konstruiere ich auf ähnliche Weise einen Begriff von *Dysfunktionalität*, der kompatibel ist damit, eine Patientin als rational darin zu verstehen, eine dysfunktionale Überzeugung zu haben. Laut dieser Begrifflichkeit ist etwas dysfunktional für eine Agentin genau dann, wenn es spezifischen Bedürfnissen von ihr kausal entgegenarbeitet und als Folge signifikante Mengen an Leid produziert. In Kapitel 6 expliziere ich die wichtigsten Resultate meiner Arbeit.

Schlagwörter: Philosophie der Psychologie, Modelle psychischer Störungen, Dysfunktionalität

Acknowledgements

This project has profited from the help of and support from so many people that I can hardly name all of them. Nonetheless, I will give it a try.

I would like to start by thanking Uljana Feest, my first supervisor, for her extremely helpful, detailed comments on all those manuscripts that I dared to send her, and for many supporting conversations. Further thanks go to my second supervisor, Marie Kaiser, for her very helpful remarks on my manuscripts. Furthermore, I would like to thank Torsten Wilholt, Anke Büter, Dietmar Hübner, Saana Jukola and Mathias Frisch for commenting extensively on my work. I am grateful also to the other members of our graduate school for their helpful feedback. Special thanks go to my office mates, in particular to Philipp, Stefano, David, Enno and Leon, as well as Rafael. Thanks for countless coffee breaks, for inspiring discussions of philosophical work, for lots of nonsense, and especially, for being kind when I needed it. Further thanks goes to six anonymous psychotherapists who shared their extensive knowledge about clinical practice with me, and to one anonymous professor of clinical psychology.

My deepest gratitude goes to my parents for always supporting me (even though philosophy confused you quite a bit). I am heavily indebted also to my friends, who amazingly stayed with me even during the last weeks of this project when I was a zombie running on caffeine and carbohydrates. Thank you for reminding me that I'm more than a brain on a stick. I couldn't have done it without you. Thank you, Kristin, Lukas, Yann, Daria, Elisa, and Anne-Kathrin, as well as Anja, Rusty, Cornelia, Leo, and Felix. For countless hugs, thousands of messages, probably tons of food, sending packages full of chocolate, for crafting shiny objects for me, dragging me into the sun, putting weird drawings into my manuscripts, and for teaching me how to make things my own. Most importantly, thank you for all the glamour.

Contents

List of Abbreviations

APA American Psychiatric Association.

CBT Cognitive-Behavioral Therapy.

CT Cognitive Therapy.

DSM-5 Diagnostic and Statistical Manual of Mental Disorders, fifth edition (published in 2013).

DSM-IV Diagnostic and Statistical Manual of Mental Disorders, fourth edition (published in 1994).

EEA Environment of Evolutionary Adaptedness.

ICD International Classification of Diseases.

MCT Metacognitive Therapy.

MDD Major Depressive Disorder.

NPD Narcissistic Personality Disorder.

OCD Obsessive-Compulsive Disorder.

PTSD Posttraumatic Stress Disorder.

REBT Rational-Emotive Behavior Therapy.

RET Rational-Emotive Therapy.

SAD Social Anxiety Disorder.

Chapter 1

Explanatory Models of Mental Disorders

1.1 Introduction and Motivation

In this chapter, I will set the stage for what is to come by presenting two explanatory models that are used in the applied discipline of clinical psychology for the purpose of understanding and treating specific mental disorders, namely, *Major Depressive Disorder* (*MDD*) and *Obsessive-Compulsive Disorder* (*OCD*).[1]

In this very first section, I will motivate my investigation of explanatory models from clinical psychology and provide some background knowledge that is helpful for understanding both my interest in them and the status of these models in clinical psychology. In the second section, I will characterize the kind of phenomenon[2] they aim to account for, that is, mental disorders. In the third and fourth section, I will present these two models. For an overview over the structure of this dissertation, see the fifth section of this chapter.

I understand clinical psychology as an *applied scientific discipline* in its own right[3] that differs in important respects – that is, over and above its methodol-

[1] Over the course of this investigation, I will both use the terms "MDD" and "depression" for the phenomenon in question. They are intended to denote the exact same thing, even though "depression" is also often used to denote a certain kind of *mood* (e.g., in Lindert et al. 2014). This is consistent with how Beck uses the term, for example, when he states that "Depression may appear as a primary disorder or it may accompany a wide variety of other psychiatric or medical disorders." (Beck 1967, p. xiii) and later in his book presents symptoms of this disorder that are very similar to the ones that we know from the DSM-5. Similarly, in the recent article from 2016, the authors state that "it [the model] should provide a framework to explain the natural history of depression: predisposition, precipitation, and recovery from the *disorder*" (Beck & Bredemeier 2016, p. 596, my italics).

[2] For the purposes of this dissertation, I intend to use the notion of "phenomenon" in a relatively neutral manner, thinking of phenomena mainly as objects of investigation in the sciences that have some repeatable features (think of, for example, Bogen & Woodward 1988, p. 317)

[3] In contrast to research areas like *strollology* (Burckhardt 2015), which mainly consists in a

ogy – from psychiatry, for example. While psychiatry is, in my understanding, a subdiscipline of medicine, clinical psychology is a subdiscipline of *psychology*. Clinical psychology, as I see it, is mainly directed at developing tools for effective psychotherapy and at understanding mental disorders (my characterization thus focuses more on the research side than several other characterizations, see, for example, Plante 2005). In this investigation, I will focus on the specific disciplinary practices of clinical psychology and how they can be characterized and understood philosophically.

Which features set clinical psychological models apart from other models of the same phenomena, especially those from neuroscience and psychiatry? I ask this question for several reasons. Firstly, there are movements within both psychology and philosophy according to which psychiatry – understood as a field that is concerned with mental health and has clinical psychology as a part (Kendler 2008, p. 1) – should develop to become *clinical cognitive neuroscience* (e.g. Murphy 2006, p. 108). Secondly, some authors have proposed that psychological explanations really are only elliptical *mechanistic* explanations that should aim to represent the spatial structure of a system as well (Piccinini & Craver 2011). My thesis is that these proposals are neither plausible, nor are they good reconstructions of how psychologists go about creating and improving models. Thirdly, many authors take psychiatry as having clinical psychology as a part, or at least use the term "psychiatry" to cover all the fields that are concerned with mental health and pathology (Kendler 2008, p. 1). I oppose this view: While models from psychiatry normally focus on the physiological level to explain diseases (or disorders), clinical psychology usually focuses on mental states and processes like beliefs or cognitive biases in explaining mental disorders. Mentioning intentional entities is essential for those models to be used for their actual purposes in psychotherapy, as I will show in chapter three. It is these two features, namely, referring to intentional states and having a dual purpose in research and practice, that sets these models apart from models in medical fields. Surprisingly, perhaps, the process of construction and especially the process of applying these models in psychotherapeutic practice have not received much philosophical analysis thus far.

particular methodology for studying specific phenomena that belong to the domains of sociology or cultural studies.

Allow me to make some remarks on what the reader should not expect from this investigation. Firstly, I will largely ignore the question of whether either these models or these explanatory practices really provide genuine *explanations* on any philosophical account. Instead, I will focus more on how they are actually used in research and psychotherapy. Secondly, for the same reason, I do not put too much emphasis on the question whether the objects at the center of my investigation really are *models* or whether they should rather be understood as *theories*. I take them to be objects that represent particular mental disorders in an idealized manner, and as serving important functions in clinical psychological research and psychotherapy, which makes them suitable objects of investigation. Calling them "explanatory models" despite these reservations has more to do with how psychologists refer to them. With these caveats in mind, let me start by characterizing these models. Usually, they are intended to provide answers to at least the following two questions (see, e.g. Wittchen & Hoyer 2011, p. 791):

1. the etiological question: How did this disorder come about?
2. the question of maintenance: Why do the symptoms recur, instead of fading away?

To answer the first question, psychologists present factors that temporally precede and predispose someone for the disorder. The idea is that, in the presence of appropriate stimuli – that is, particular *activating events* – certain symptoms are brought about. To answer the second question, they aim to show how different factors work together in leading to either temporally stable or recurring symptoms. These factors partially comprise the disorder's symptoms, partially independent *maintaining factors*.[4]

Plausibly, to achieve an understanding of (1), psychologists need to put forward a *causal-contrastive* explanation, highlighting which temporally preceding factors in the patient's[5] history brought about the specific mental disorder in

[4] Often, they also provide a (partial) answer to the question of what the disorder actually *is*. We can see this particularly clearly when considering the model of depression that was provided by Beck & Bredemeier (2016). Depression thus emerges as the result of an evolutionarily adaptive cognitive-behavioral program. By contrast, the model of Salkovskis et al. (1998) provides less of an understanding of what the disorder at hand *is*. But importantly, this is not its main goal.

[5] For lack of a better term, I will mainly use the term "patient" when referring to individuals with mental disorders, although I am of course aware both of the fact that not all individuals who suffer from mental disorders *are*, in fact, patients.

question. These factors are contrastive, since they do not occur in the history of someone who is healthy (this sense of "contrastive" is inspired by Lewis 1986, p. 229). To understand (2), psychologists need to refer to certain causal factors that make clear why the symptoms recur – usually, this will happen by pointing to particular, relatively stable features of the patient.[6]

Explanatory models of mental disorders are used for the purposes of (1) explaining *individual case histories* (Cooper 2007, p. 67) of subjects suffering from a mental disorder and of (2) explaining the disorder *as such*.[7] Explanations of individual case histories are usually presented in psychotherapy, while explanations of mental disorders are usually put forward in research contexts – in particular, in academic training of future psychologists. Since the models mention only generic features of these disorders, they constitute explanations of mental disorders as such. In chapter three, I will analyze how these models are used by mental health professionals to generate explanations of individual case histories.

In describing the explananda, I will make use of the diagnostic criteria presented in the most recent edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-5*). I have decided to use the DSM-5 as opposed to the *International Classification of Diseases* (*ICD*) for several reasons: Firstly, the DSM is more widely used in clinical psychological research, and, secondly, the ICD and the DSM criteria of mental disorders are usually geared to each other. Thus, the differences in their understanding of those two disorders are small enough to be ignored for my purposes within this thesis.

The two models that I am concerned with for the rest of this book are (at least partially) *cognitive* models of these disorders that are conceptually tied to the framework of *Cognitive-Behavioral Therapy* (*CBT*).[8] Before delving into the details of these models, I will, firstly, lay out my reasons for choosing them.

I focus especially on models from the cognitive-behavioral framework, since it is

[6] But it might, if we think that mental disorders are actually networks of interacting symptoms (Borsboom 2017), also happen by showing how symptoms stabilize one another, that is, we need not necessarily postulate further factors.

[7] This distinction resembles one made by Rachel Cooper between "natural-history style explanations" and explanations of individual cases (Cooper 2007, p. 44), but it is not strictly identical with it, because, while she focuses on those objects that these explanations invoke in order to explain – in one case, that is, natural kinds –, I only want to, for the moment, distinguish two different kinds of explananda.

[8] The terms "Cognitive Behavioral Therapy" and "Cognitive Behavior Therapy" are used here to denote the very same form of therapy, and I will switch between the two occasionally.

arguably the most influential framework for conceptualizing mental disorders, it is most in line with current knowledge about psychological processes, and because CBT is often called the *gold standard* for treating mental disorders by clinical psychologists (David et al. 2018).

CBT is a widely-used form of psychotherapeutic treatment of mental disorders. Among these treatments, it has been researched most thoroughly, with several meta-analyses showing its effectiveness (see, for example, Beck 2005). Research to date has shown that, for many mental disorders, CBT is effective (Nathan & Gorman 2015, p. xv-xxvii), that it usually is at least as effective as other psychotherapeutic treatments, and that it appears to be superior to medication for many mental disorders (e.g. Barth et al. 2016, Butler et al. 2006, Hofmann et al. 2012, Tolin 2010).

Furthermore, models developed in the cognitive-behavioral tradition are among the few explanatory models in clinical psychology that are both relevant for research on these disorders[9] and actually used within psychotherapeutic practice with the explicit aim of explaining disorders to patients[10]. This is mirrored by the fact that handbooks of clinical psychology and psychotherapy often incorporate reference to so-called "psychoeducation" of patients, that is, teaching the patient important facts about her disorder (e.g. Wittchen & Hoyer 2011).

Many researchers in clinical psychology are trained as cognitive behavioral therapists, and many psychotherapists make use of the theoretical explanatory models in their therapeutic practice, leading to strong interactions of theory and practice in this discipline, a phenomenon that is of particular interest to me.[11]

[9] We can see this clearly when considering the number of papers from clinical psychology, neuropsychology and the like that have referred to the original paper of Salkovskis et al. (1998). At the time of this writing (that is, 01/21/2019), GoogleScholar listed 216 citations of the original paper that treat topics as diverse as the neuropsychology of the disorder, psychotherapy research, additional testing of the components of the cognitive model, the development of psychometric instruments and further cognitive factors – e.g., attentional control – that are relevant to the disorder. Since this paper is only available as a supplement, citations of the first paper treating this model must in fact be counted in as well, and these were, at the time of this writing, 646.

[10] Even though, due to the number of participants, inconclusive, I infer this from several qualitative interviews and private conversations with psychotherapists and psychotherapists in training.

[11] This becomes clear when considering the ideal of the "scientist-practitioner", which arguably had and still has a huge impact on how clinical psychologists and cognitive-behavioral psychologists view themselves. According to this model, clinical psychologists should not *only* carry out effective interventions, but they should at the same time also "contribute to the development of the knowledge base itself." (Lane & Corrie 2007, p. 14) Carrying out effective interventions, for the proponents of this position, also means to continually engage with the current research in clinical psychology.

The model of MDD that I will discuss has been proposed by the founder of CBT, Aaron Beck, and his colleague Keith Bredemeier in 2016, and the model of OCD by Salkovskis and colleagues in 1998.[12] I decided to engage with these specific explanatory models for several different reasons.

For one, MDD and OCD are two very common mental disorders, with incidences of roughly 3% and 2%, respectively (World Health Organization 2009). Both are considered major factors in the global burden of disease, with MDD being the leading cause of so-called "years lost to disability" worldwide (World Health Organization 2009, p. 8). OCD might, according to some studies, lead to a stronger reduction of quality of life than schizophrenia (Stengler-Wenzke et al. 2006). Thus, these disorders are both relatively common and have quite debilitating effects on patients. Furthermore, a lot of research has already been done on them and is still being done. We may thus assume that the psychological understanding of these disorders is quite advanced when compared to other mental disorders.[13]

Furthermore, the currently important models of these disorders are based on the assumption that they should be – at least partially – explained by reference to cognitive factors such as beliefs or desires (e.g. Beck & Bredemeier 2016, Teasdale & Barnard 1993, Salkovskis et al. 1998).[14] On that matter, MDD and OCD differ from several other mental disorders (e.g., schizophrenia), which supposedly have their most important contributing factors in specific genetic or brain abnormalities (e.g. Insel 2010). Thus, these specific explanatory models are well suited for my investigation into clinical psychological models and explanations of mental disorders.

As I will show in the remainder of this chapter, these explanatory models exhibit several noteworthy features. Some of these features set them apart from explanatory models of the same phenomena from neuroscience or psychiatry,

[12] I will sometimes use expressions like "Beck's 2016 model" or "Salkovskis' model from 1998" for the more recent versions of these models that were authored not only by these two authors alone. When I do so, I merely want to keep things brief.

[13] A further reason is that I intend to only deal with paradigmatic cases of *mental* or *psychological* disorders here. This means that I have tried to exclude neurological disorders or such mental disorders that seem to normally have physiological causes – think of schizophrenia, for example – and to reduce the discussion of personality disorders to a minimum.

[14] Interestingly, even psychodynamic models of mental disorders seem to (at least implicitly) refer to (non- or pre-conscious) cognitive factors: For example, two psychodynamic models of depression that are intended to integrate previous psychodynamic models of the disorder (Busch et al. 2016, p. 27-28) appear to make reference either to cognitions or to emotions that have a cognitive component. While one of them makes reference to *perceived rejection*, the second model makes reference to *disappointment*.

for example. These rather refer to non-mental objects like neurotransmitters or neural networks[15].

One such feature is the usage of *folk-psychological vocabulary*, for example, referring to thoughts or affective states. These models are, nonetheless, intended as causal models and they use functional vocabulary (this was also pointed out by Bolton & Hill 1996). Furthermore, these models mainly refer to entities that are *in principle* observable by the patient either through introspection (even though there is, of course, a lot of controversy over the reliability of introspection) or through simple observation.

The models provided in this tradition stand in an interesting contrast to the views of particular philosophers about how explanatory models *should* be and *are* in fact set up. Think of, for example, Churchland (1981), according to whom good explanatory models of mental phenomena should be free of folk-psychological vocabulary.

Furthermore, these models often describe behavioral strategies or beliefs as "dysfunctional", thereby identifying these factors as harm-inducing.[16] In doing so, clinical psychological models of mental disorders differ substantially from explanatory models that emerged from, e.g., neuroscience. Even when researchers working on either side of this divide use the same word – "dysfunction", for example – they mean very different things.

Thus, I am interested in the cognitive-behavioral tradition since it arguably provides both plausible psychological models of mental disorders that are relevant in clinical psychological research and that are, at the same time, used in one psychotherapeutic framework that has proven to be robustly (if moderately) successful (David et al. 2018). In my understanding, it thus qualifies as *successful applied clinical psychology.*

1.2 Background

Now that I have given some reasons to be interested in these specific models, I will start delving into a more sustained analysis of them. I will begin with a

[15] Here, I would like the term "non-mental" simply to mean objects or processes without mental content. Of course, such objects may nonetheless be correlates of processes with mental content.

[16] This is very rough, and does not cover the myriad of different meanings of "dysfunctional" and related terms. Nonetheless, my point here is only that these apparently value-laden terms are normally used in clinical psychological models of mental disorders. A more in-depth analysis of these concepts will be provided in chapter five.

characterization of *mental disorder* from the fifth edition of the DSM-5. The authors characterize this notion as follows:

> "A mental disorder is a *syndrome* characterized by *clinically significant disturbance* in an individual's cognition, emotion regulation, or behavior that *reflects* a dysfunction in the psychological, biological, or developmental processes underlying mental functioning. Mental disorders are usually associated with significant *distress* or *disability* in social, occupational or other important activities." (American Psychiatric Association 2013, p. 20; my italics)

In other words: According to the *American Psychiatric Association* (*APA*), mental disorders are sets of clinically relevant symptoms that are brought about[17] by an underlying dysfunction. Importantly, these symptoms result in distress or disability, that is, they are *harmful* to the individual.

The APA's formulation is ambiguous in interesting ways concerning the contrast between the so-called *latent variable model* and the *network model* of mental disorders (e.g. Bringmann & Eronen 2018). For one, suggesting that there is exactly *one* underlying dysfunction for the symptoms of each mental disorder is reminiscent of a common cause perspective, according to which mental disorders are dysfunctions that bring about specific symptoms, but that are themselves not identical to these symptoms. For the other, the APA refers to mental disorders as *syndromes*, that is, sets of symptoms. This view is closer to network approaches to psychopathology, at least if we take it literally, that is, as meaning that mental disorders are *identical* to the respective sets of symptoms.[18]

Although this characterization of mental disorders can only be a first pass at what these conditions are[19], this suffices for the purposes of the time being.

To provide some background on my discussion of these two models, let me just sketch some main assumptions of CBT. One tenet of CBT is that what is

[17] In a not necessarily causal sense of "bringing about".

[18] Now, although the question whether mental disorders are actually latent variables or rather networks of interacting and self-stabilizing symptoms is an intriguing one that has generated much debate (e.g. Borsboom 2008, Borsboom & Cramer 2013, Bringmann & Eronen 2018, Cramer et al. 2010), I will, for the purposes of this dissertation, set it aside – mainly because it does not have too much bearing on those practices that are of interest to me here.

[19] One reason is that it is far from clear what psychologists mean when they refer to the "clinical significance" of the symptoms in question. Furthermore, the term "dysfunction" requires further clarification as well.

relevant for the kinds of emotions an individual feels in response to an event is how she *appraises* that very event. More precisely, psychologists assume that human beings react to events with particular so-called *automatic thoughts* that influence their emotional reactions to the event. These automatic thoughts are sub- or pre-conscious, and they are supposedly the output of more generic cognitive structures, so-called cognitive *schemas* (Beck 1995). Schemas are understood as follows:

> "[...] a *structure* used for screening, coding, and evaluating impinging stimuli. In terms of the individual's adaptation to external reality, it is regarded as the mode by which the environment is broken down and organized into its many psychologically relevant facets [...] schemas are conceived as relatively stable cognitive structures which channel thought processes [...] The schema abstracts and molds the raw data into thoughts or cognitions. A cognition, in the present usage, refers to any mental activity which has a verbal content [...] the notion of schemas is utilized to account for the regularities and repetitive themes [...] in the reactions to [inner and] environmental events." (Beck 1964, p. 562-563, my italics).

Very roughly, cognitive schemas can be understood as made up of sets of very basic beliefs that determine how someone reacts both cognitively and emotionally to events.

This should give the reader a first understanding of the bare bones of CBT that is sufficient for understanding the two explanatory models of interest.

In the next two sections, I will present the two explanatory models. They are organized identically: In each section, I will first describe the phenomenon, that is, either OCD or MDD. Secondly, I will present the respective explanatory model. Finally, I will identify the central explanatory strategy of each model.

I will first analyze the model of OCD presented by Salkovskis et al. (1998) and subsequently discuss the model of MDD from Beck & Bredemeier (2016).

1.3 Obsessive-Compulsive Disorder

1.3.1 Diagnosis

What do clinical psychologists mean when they speak of "Obsessive-Compulsive Disorder"? Let me start by presenting an example of someone suffering from the condition:

> “Mrs. M., 44 years old, has been suffering from compulsory washing for 18 years, which severely affects her daily life. After being in contact with [what she perceives as] “dirt”, she performs extensive washing rituals, for example after taking the subway, touching money, but also if she believes that her car has had contact with dead animals (pigeons, frogs, ...). She usually disinfects her money, her bag, washes all of her clothing, stands under the shower for $1\frac{1}{2}$ hours, sometimes drives through the carwash several times per day. [...] Grave problems with her husband are caused by frequent washing and sometimes disposing of his clothes [...]” (Wittchen & Hoyer 2011, p. 1006; my translation)

It seems fairly clear that, for all we know, the woman presented here has a mental disorder in the sense specified in the APA’s characterization: She shows a set of symptoms that are recurrent and lead to suffering in various areas of her life. The DSM-5’s diagnostic criteria are the following:

A. Presence of obsessions, compulsions, or both:
   Obsessions are defined by (1) and (2):

   1. Recurrent and persistent thoughts, urges, or images that are experienced, at some time during the disturbance, as intrusive and unwanted, and that in most individuals cause marked anxiety and distress.
   2. The individual attempts to ignore or suppress such thoughts, urges, or images, or to neutralize them with some other thought or action.

   Compulsions are defined by (1) and (2):

   1. Repetitive behaviors [...] or mental acts [...] that the individual feels driven to perform in response to an obsession or according to rules that must be applied rigidly.
   2. The behaviors or mental acts are aimed at preventing or reducing anxiety or distress, or preventing some dreaded event or situation; however, these behaviors or mental acts are not connected in a realistic way with what they are designed to neutralize or prevent, or are clearly excessive. [...]

B. The obsessions or compulsions are time-consuming [...] or cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.

C. The obsessive-compulsive symptoms are not attributable to the physiological effects of a substance [...] or another medical condition.

D. The disturbance is not better explained by the symptoms of another mental disorder [...]. (American Psychiatric Association 2013, p. 237)

The APA also specifies exclusion criteria for conditions when the diagnosis should not be given. Roughly put, the symptoms should neither be more plausibly attributable to another mental disorder, to some medical condition, or to the effects of a particular substance – that is, alcohol, drugs or medication – that the individual has consumed (American Psychiatric Association 2013, p. 237).[20]

There are two questions to ask here: Firstly, how exactly do the diagnostic criteria for OCD satisfy the characterization of mental disorder that I have presented above?, and secondly, how does Mrs. M. fit into the diagnostic criteria of OCD?

As we have seen above, the APA's characterization of mental disorder has three components: (1) Mental disorders are syndromes, that is, sets of symptoms, which (2) are associated with clinically significant distress and (3) reflect a dysfunction that underlies the syndrome.

I consider it quite clear that (1) and (2) are fulfilled here: OCD is characterized by specific symptoms, that is, obsessions and compulsions, thus satisfying (1). Furthermore, (2) is either fulfilled by the fact that the obsessions are time-consuming[21] or by the definition of obsessions as entities that lead to anxiety or distress. Finally, there is nothing in the example which explicitly satisfies (3). It thus seems like an underlying dysfunction is merely assumed.[22]

How would the woman in my case-study satisfy the criteria of OCD? Firstly, she clearly suffers from compulsions as defined above: She performs washing

[20] The interested reader can find the complete list, including the APA's specifications for possible other factors accounting for these symptoms, in appendix A.5.

[21] At least if we assume that being time-consuming in the way specified by the DSM (more than one hour per day) leads to clinically significant suffering.

[22] Of course, the question about the underlying dysfunction *might* be an important one. However, I think that this is not important here. Since the DSM only specifies criteria for the mere *diagnosis* of a condition, it must not also be in a position to pinpoint the underlying dysfunction.

rituals that – in accord with criterion (1) – she applies rigidly. The first conjunct of criterion (2) is satisfied, since the behavior is directed to prevent being in contact with dirt. It also is clearly excessive, thereby satisfying the second conjunct. Thereby, she satisfies criterion A.

Secondly, the time-criterion (B) is satisfied, since she stands under the shower for $1\frac{1}{2}$ hours per day. Additionally, there is significant impairment of her social relationships: There are grave problems with her husband that are connected to her compulsions. Finally, since neither drug abuse nor additional mental disorders are mentioned here, we may assume that criterion C and D are satisfied as well.

Given that we now have a rough understanding of the disorder in question, I will turn to the cognitive psychological model of Salkovskis et al. (1998).

1.3.2 Explaining OCD

Salkovskis et al. (1998) presented what arguably now has the status of a *standard model* for explanations of OCD from a clinical point of view. It is a *cognitive* model, since it identifies particular cognitive entities as important factors that bring about the disorder and contribute to its maintenance.[23]

This model is a more developed version of Salkovskis' first and widely-received model of OCD (Salkovskis 1985), according to which the syndrome might be due to a particular interpretation of *normal* intrusive thought (see fig. 2.1). This model, in turn, is based on the so-called "cognitive model" (Beck 1995, p. 14), one of the major theoretical postulates underlying *Cognitive Therapy* (*CT*) that was developed when Beck tried to make sense of the phenomenon of depression.[24] According to it, a person's reaction to a given event crucially depends on how she *construes* this event (Beck 1995, p. 14).[25] In consonance with

[23] Although the model is first presented in the article by Salkovskis et al. (1998), I will mainly make reference to the later paper by Salkovskis (1999) here, since the earlier paper focuses on the understanding of obsessional thinking provided by this model, while the later paper deals more explicitly with how this model helps explain OCD *in general*.

[24] Interestingly, Aaron Beck himself has, at a later point in time, presented the so-called "Generic Cognitive Model" (Beck & Haigh 2014, p. 1), that is taken to cover the underlying, common features of several mental disorders, including MDD and OCD. But it does not entail the characterization of OCD that Salkovskis et al. (1998) presented – even though the two models share the same conceptual framework, namely, Beck's first cognitive model.

[25] As Aaron Beck mentions himself, this idea did of course not originate with him. In fact, it can be found already in the writings of several ancient philosophers that have heavily influenced Beck himself as well as Albert Ellis, the founder of of RET. The latter has also exerted enormous influence on what today is CBT, to the point of sometimes being referred to as "the oldest form of

the cognitive conceptualization, Salkovskis' model of OCD focuses on how individuals with the condition *interpret* a certain class of cognitions.[26] In a later paper on the same model, Salkovskis writes that he takes this model to cover both "the *origins* and the *maintenance* of obsessional problems" (Salkovskis 1999, p. S33, caption of figure 1. my italics), and thus, of OCD.[27]

In essence, this model identifies the misinterpretation of intrusive thoughts as the source of the symptoms and the disorder's maintenance. Intrusive cognitions are a specific class of cognitions that differ from other cognitions by "interrupting the person's current stream of consciousness" (Salkovskis 1999, p. S31) and having content which is deemed "upsetting, unacceptable, or otherwise unpleasant" (Salkovskis 1999, p. S31) by the individual. One crucial background assumption is that such thoughts are not, in isolation, pathological, but appear in healthy individuals as well (Rachman & de Silva 1978). Thus, not the occurrence of such cognitions *per se* is problematic, but their *appraisal*. Because the individual perceives them as implying the possibility of harm or danger (Salkovskis 1999, p. S31), these thoughts become problematic. This interpretation differs from the one of healthy individuals, who manage to simply ignore these intrusive thoughts or interpret them correctly, that is, as insignificant thoughts that they just happen to have. Since intrusive thoughts appear in healthy individuals as well, there is no need – at least not for the *clinical* psychologist – to explain their origin.

According to the model of Salkovskis et al. (1998), this systematic misinterpretation of intrusive thoughts depends upon pre-existing and relatively enduring *dysfunctional beliefs* of the person. These beliefs are similar to "ordinary" beliefs in many of their features – most importantly, they are usually acquired on the basis of and justified by the patient's experiences and other beliefs that she might hold.[28] While dysfunctional beliefs are assumed to be at the core of a

cognitive-behavior therapy" (Ellis et al. 2010, p. v).

[26] I think that it is fair to assume that having a specific kind of cognition qualifies as a special kind of event.

[27] Taking a mechanistic perspective on these models, one might argue that they are mixed models that combine mechanistic explanations of a phenomenon's production and explanations of something's maintenance (Craver & Darden 2013, p. 65-66). Although the question of whether these models can be understood in the mechanistic framework is a fascinating one, I will not be concerned with it here.

[28] Concerning this feature, they differ from certain other mental phenomena that occur in mental disorders. Think of, for example, delusions, which are not usually well integrated into the subject's usual experience and do not – at least from an outside perspective – seem to stand in the right kind of justificatory relation to the patient's experiences and her other beliefs.

number of different mental disorders, the particular *content* of these beliefs is taken to be specific for each disorder, and to be causally relevant for its specific symptoms (Beck 1967, p. 267).[29] Usually, dysfunctional beliefs in individuals with OCD are centered around themes such as (1) the need to be perfect, (2) the idea of being responsible for current or future harm and (3) the need to control one's thoughts (e.g. Obsessive Compulsive Cognitions Working Group 1997).

Such beliefs are taken to be acquired as a reaction to specific early life experiences. When they are adopted, they normally are useful for the individual (Salkovskis & Forrester 2002, p. 46), at least in helping the individual to make sense of her experience of the world. Thus, the individual is taken to have *good reasons* for acquiring these beliefs.[30]

Dysfunctional beliefs are assumed to be either active most of the time or active at least during periods of time when the patient is experiencing the symptoms of her mental disorder (Beck 1995, p. 15).[31] In an asymptomatic period of time, these beliefs are thought to be mostly inactive. Then, at a particular point in time, critical incidents – which the individual perceives as related to her dysfunctional beliefs – lead to the *activation* of those beliefs. As a consequence, they become thought- and action-guiding. Having such dysfunctional core beliefs presents a first hint of why some people are more vulnerable for OCD or MDD than others – in keeping with the *diathesis-stress model* (Wittchen & Hoyer 2011, p. 21).

How are these ideas related to the second part of the model that concerns OCD's *maintenance*? According to the model, an individual with these dysfunctional beliefs reacts in a particular way when she is confronted with an intrusive thought: She reacts with thinking that she is *responsible* for (preventing) past, current or future harm that – in the case of future harm – would otherwise afflict the self or others (Salkovskis 1999, p. S31). The ideas of both being responsible for potential harm and being in a position to prevent

[29] This hypothesis is also sometimes called "cognitive content specificity".

[30] The precise way in which he or she has good reasons will be spelled out in a later chapter, when dealing with the concepts of rationality and irrationality in psychological and, in particular, psychotherapeutic explanation of mental disorders such as OCD and MDD.

[31] The precise understanding of this varies. Some clinical psychologists have proposed that these beliefs are "dealt with" through other, counteracting beliefs that allow the individual to cope with the dysfunctionality of these beliefs (meaning that he or she does not develop a full-blown mental disorder) without actually needing to challenge the problematic beliefs in question (Beck 1995, p. 20-21).

it lead to *negative mood* such as distress, *attentional bias* in favor of particular kinds of information – specifically information that concerns danger –, *counterproductive safety strategies* and the desire to engage in *neutralizing behavior*. "Neutralizing behavior" denotes compulsive behavior which the individual engages in to feel less upset about these responsibility beliefs.[32] In the short term, neutralizing leads to a subjectively experienced decline in feelings of responsibility (Rachman et al. 1976, p. 450). In the long term, this *stabilizes* the dysfunctional beliefs by preventing them from being disproved by contrary experiences (Salkovskis 1999, p. S32f). *Counterproductive safety strategies* are aimed at reducing the intrusive thoughts themselves through, e.g., avoidance of specific situations associated with the occurrence of these thoughts. These four components increase the probability of further intrusions, elevate the amount of perceived threat and increase the perception of responsibility, thus "leading to a cycle of negative thinking and neutralizing." (Salkovskis 1999, p. S32). Figure 1.1 serves as a graphic representation of this model.

In short, I understand this second part of the model to claim that OCD is maintained by an interplay of the following factors: (i) intrusive thoughts that lead to (ii) misinterpretations of significance of – statistically normal – intrusions, that, in turn, bring about all of the following and are reinforced by (iii)-(vi): (iii) attention and reasoning biases, (iv) mood changes, (v) counterproductive 'safety' strategies and (vi) neutralizing actions, that, again, are causally relevant for further intrusive thoughts.

1.3.3 Making Sense of the Model

In this section, I will make some remarks on how the model is best understood and interpreted. I will draw the reader's attention to the most important features of the model. Furthermore, I would like to note some problems surrounding particular of its features and suggest a way to solve them.

Firstly, it is important to note that this explanatory model is intended to

[32] One might discuss why *beliefs* would stand in need of being neutralized. Most plausibly, the author's idea is that these beliefs are somehow charged with evaluative or emotional content that needs to be neutralized. Another interpretation would be that the beliefs in question must be *falsified*, but this seems confusing, since we would want to allow the possibility that two beliefs, e.g., "I could be a danger to others." and "I am a danger to others." are equally in need of neutralization – but only one of them can plausibly be falsified (even if we take the modal operator to refer to a suitably restricted kind of possibility). In short: Falsification of beliefs does not seem to be what this is about.

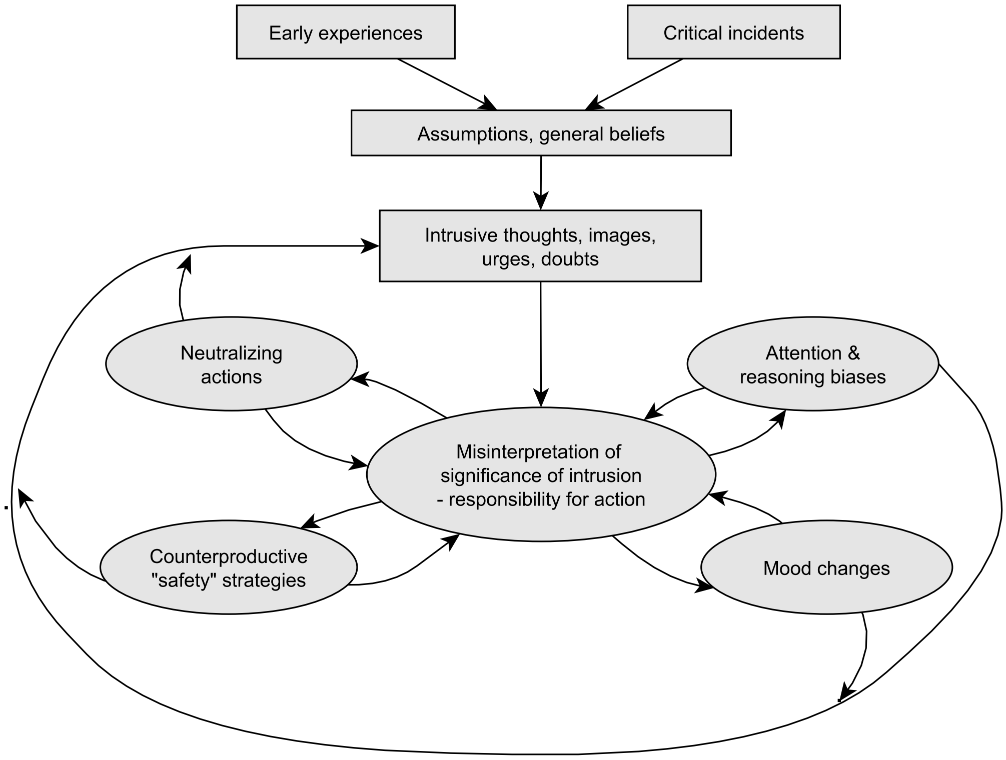
Early experiences
Critical incidents
Assumptions, general beliefs
Intrusive thoughts, images, urges, doubts
Neutralizing actions
Attention & reasoning biases
Misinterpretation of significance of intrusion - responsibility for action
Counterproductive "safety" strategies
Mood changes

Figure 1.1: Cognitive model of the origins and maintenance of Obsessive-Compulsive Disorder, put forward by Salkovskis et al. (1998), slightly adapted.

convey the *causal structure* underlying the disorder's development and maintenance. This becomes clear when considering the relations between model components depicted in fig. 1.1: The relations that they supposedly[33] represent are thought to satisfy interventionistic or manipulability accounts of causation (e.g. Woodward 2003).[34] Consider the relation between neutralizing actions and dysfunctional beliefs: According to the model, if someone was to *intervene* on the person's dysfunctional beliefs – assuming a particular set of variables and under the supposition that we are holding relevant background variables fixed –, the person's neutralizing actions would occur less frequently. Thus, the relation between these two features is counterfactually stable under interventions on the independent feature, and thus, emerges as causal on this account. Plausibly, the same holds for the other relationships represented in

[33] This is a bit tricky, as there might be differences between some of the relations that were thought to hold between components of the model in 1995 and the relations that *actually* hold between the represented objects. I only want to make a claim about the kind of relation that would hold, given that the model were (by and large) correct.

[34] At least if we idealize those relations somewhat.

the model as well.[35]

The causal relation between such beliefs and symptoms of mental disorders is at the heart of CT (Beck 1995, p. 14), which intends to treat mental disorders by changing the individual's pervasive dysfunctional beliefs.

That said, it is worth pointing out some problems that arise with the graphical representation in fig. 1.1, if we assume that the arrows depicted there are indeed intended to represent causal relations. Two problems arise here, one pertaining to the form of this model, the other to its content.

One very obvious difficulty is the fact that, prima facie, several causal relations would seem to emerge as *circular*. In other words: Certain events would be causally relevant for themselves. This can't be right.

This issue is easily circumvented by requiring that the relevant two dependency relations are not to be understood as pertaining to the very same (token) event, but to two different token events of the same *type*: Thus understood, these circular structures merely point to *feedback-loops*, in which, e.g., catastrophic misinterpretations of the significance of an intrusive thought give rise to neutralizing actions, and these neutralizing actions enhance the probability of further – but different – misinterpretations of significance to occur. One example might be the misinterpretation "Thinking that I might be infected with a serious disease means that I might be a danger to my friends". This thought is causally relevant for neutralizing actions such as washing one's hands, which again heightens the probability of further misinterpretations of significance to occur.[36] The mutually reinforcing nature of these symptoms is exactly what psychologists want to point out in this context.

Let us now turn to tensions between the graphic representation in figure 1.1 and the way in which the model is usually described.

Problematically, the proposition that intrusive thoughts are *caused* by general

[35] This is very rough. As has been pointed out, among others, by Kästner (2018), classical interventionist theories like the one by Woodward (2003) are actually too restrictive to account for those causal relations that are represented in psychiatric models – and the same holds, I think, for clinical psychological models of mental disorders. Nonetheless, just as she claims, I think that a somewhat "relaxed" variant of this, that is, *difference-making interventionism*, can make sense of the causality that is represented in these models. This account leaves the central idea of interventionism intact – that is, the view that causation is essentially tied to difference-making –, but changes the account somewhat, such that it fits better with how causal reasoning is actually employed in such psychiatric and clinical psychological models.

[36] One might think that what is at issue here is really that a misinterpretation of significance leads to compulsions, which strengthens the underlying *disposition* to misinterpret one's intrusions.

beliefs that are specific to individuals with OCD – which seems to be implied by fig. 1.1 – does not seem to be what the authors have in mind. After all, a main background assumption of the model is that the problematic intrusive thoughts in individuals with OCD originate in *normal* intrusive cognitions (Salkovskis 1999, p. S31). This is based upon the finding that intrusive cognitions occur in healthy individuals just as frequently and with roughly the same content as in individuals with OCD (Rachman & de Silva 1978, p. 233). Thus, not the individual's dysfunctional beliefs bring about misinterpretations of significance by leading to intrusive thoughts. Instead, the individual's completely normal intrusive thoughts *together* with particular dysfunctional beliefs – which set the individual apart from health individuals – lead to misinterpretations of significance of the thoughts in question. In other words: The contrastively relevant cause here is the activated dysfunctional belief. As I have pointed out before, in most characterizations of the model, the origins of intrusive thoughts are actually *not* accounted for – and as we have seen, they do not have to be, since clinical psychologists are only interested in modelling the symptoms and those causal factors that actually differ between healthy individuals and individuals who suffer from OCD. I thus assume that the graphic representation misrepresents the actual content of the model slightly.

I think that one way to account for this is the following: It seems plausible to say that those *dysfunctional beliefs* that are thought to *bring about* misinterpretations of significance are actually best understood not as causal factors for the tendency to misinterpret the significance of one's intrusions, but as partially constituting this tendency for misinterpretation. One source of evidence for this claim can be found in the following quotation, specifying one of the goals of cognitive therapy:

> "To identify and modify underlying dysfunctional assumptions and beliefs *which predispose* [the patient] to negative automatic thoughts." (Robertson 2010, p. 4)

This makes sense of the fact that getting rid of dysfunctional beliefs in the individual is one of the most important goals of cognitive behavioral therapy (Salkovskis 1999, p. S40f): once they are abandoned, the *tendency* to misinterpret one's intrusive thoughts breaks down. Thus, one could say that the dysfunctional belief accounts for the disposition: misinterpreting the significance of one's intrusions, that is, reacting to intrusive thoughts with appraisals like "I have to do something to prevent harm from happening." is based on dysfunctional beliefs like "Thinking about something negative means that it is likely to happen.", and thus, will stop occurring if this belief is absent.

Furthermore, it is important to note that what clinical psychologists describe here is a set of *beliefs*, *emotional reactions* and *behaviors* that mutually reinforce one another. Thus, to account for this model, I will need to say something about the role of folk-psychological vocabulary. In chapter three, I will pay particular attention to it.

Let me draw your attention to another pattern in what we have covered so far that makes further consideration necessary: The above-presented description of the model suggests that there are specific kinds of behavior – I am thinking of neutralizing behavior – exhibited by the individual that have a particular *function* for this individual. At the same time, the notion of "dysfunctional belief" is, as I have pointed out, central for the explanatory power of the model. Interestingly, these two senses of "function" and "dysfunctionality" do not coincide. This leaves an understanding of the functional terminology employed in this model as another task to be accomplished. I will attend to this issue in the fifth chapter of this dissertation.

Let me now take a step back and consider the general explanatory strategy that is used in this model. I take it that the predisposition for the syndrome and its maintenance are explained mainly by making two moves: Firstly, the symptoms are understood as being brought about by the misinterpretation of intrusive thoughts and specific cognitive biases that accompany it; and both are due to specific dysfunctional beliefs. Secondly, these dysfunctional beliefs are, in turn, characterized as *reasonable* reactions to particular experiences. Similarly, avoidance behavior emerges as a *counterproductive safety strategy*, while obsessions emerge as *neutralizing behaviors*. It seems clear that this model emphasizes the *functions* that the apparently irrational and erratic behavior of subjects with OCD has for them. Very generally, this model *rationalizes* this mental disorder in the sense of making the symptoms appear more reasonable than they appeared at the outset. Furthermore, the individual's disorder is, as Bolton pointed out, "normalized" (Bolton 2008, p. 52) in the sense of being reduced to the workings of (relatively) *normal* mental processes and causal relations.[37] Regarding psychopathological symptoms as *variants* of normal emotional reactions is actually an explicit part of Beck's theory of emotional disorders (Weishaar 1993). This is actually slightly different from what Bolton means when he says that "[...] psychological models [...] are just those that find meaning even where it seems to have run out" (Bolton 2008, p. 184).[38]

In summary, the model of OCD that Salkovskis et al. (1998) proposed has several noteworthy features: there is (1) the usage of folk-psychological vocabulary, (2) the attempt to describe a causal structure underlying the disorder, and (3) the presence of functional vocabulary. As I pointed out (again, in keeping with Bolton 2008), this model *normalizes* OCD. Stated intuitively, this means that it emphasizes that particular kinds of behavior and emotional reaction are *reasonable*, as soon as we take their function for the individual into account: The adoption of *dysfunctional* beliefs makes sense when the context of their adoption is taken into account.

So far, I have described several important features of this model as well as the central explanatory strategy. With these findings in mind, let us now turn to the more recent, but also more complex model of MDD, put forward by Beck & Bredemeier (2016).

1.4 Major Depressive Disorder

1.4.1 Diagnosis

In the DSM-5, MDD is distinguished from other depressive disorders as follows:

> "The common feature of all of these [depressive] disorders is the presence of sad, empty, or irritable mood, accompanied by somatic and cognitive changes that significantly affect the individual's capacity to function. [...] Major depressive disorder [...] is characterized by discrete episodes of at least 2 weeks' duration [...] involving clear-cut changes in affect, cognition and neurovegetative functions [...]." (American Psychiatric Association 2013, p. 155)

[37] Bolton (2008) then analyzes the notion of "normality" in this context to have several potential meanings. Statistical normality and adaptiveness are two of them.

[38] By "meaning", the author means "various concepts, such as appropriateness of affect and behaviour, rationality of belief and action, functionality of behaviour, strategy, and regulation by information processing." (Bolton 2008, p. 182)

The term “affect” appears to be used here as a theory-neutral umbrella-term for those mental states that involve “some degree of motivation, intensity, and force” (e.g. Barrett & Bliss-Moreau 2009, p. 167). Thus, it can be used to cover (at least) emotions, feelings, and moods (de Sousa 2017). The precise changes in affect, cognition and neurovegetative functions are encoded in the diagnostic criteria. To be diagnosed with MDD, an individual has to exhibit at least five out of nine relevant symptoms, namely:

Changes in affect that are operationalized as

1. feeling sad, empty or hopeless,
2. showing markedly diminished interest or pleasure in most activities most of the time and,
3. experiencing feelings of worthlessness or excessive guilt.

Furthermore, the following *changes in cognition* may occur:

4. diminished ability to think concentrate and
5. recurrent thoughts of death.

Finally, there are *changes in neurovegetative functions*. The individual might either show

6. weight gain or weight loss,
7. insomnia or hypersomnia,
8. psychomotor agitation or retardation – that is, unintentional and purposeless motions together with emotional distress and restlessness or slowed-down thought and a reduction of physical movements –, and finally,
9. fatigue or energy loss.

In addition to this, the American Psychiatric Association (2013, p. 160f) specifies several exclusion criteria that are roughly equivalent to those that I have already mentioned when describing OCD. For the full list of diagnostic criteria, see appendix A.3.

Even a superficial look at these criteria makes it easy to see why there has been so much debate about the *heterogeneity* of depression in clinical psychology (e.g. Goldberg 2011), and relatedly, about the question whether mental disorders could possibly be natural kinds (e.g. Kincaid & Sullivan 2014): There are several symptom constellations satisfying these diagnostic criteria that do not have a single symptom in common. Nonetheless, I will put such considerations aside, taking a look at the actual models that clinical psychologists employ when trying to make sense of this disorder and its many different possible instantiations.

1.4.2 Explaining MDD

Etiology of MDD

The model of MDD put forward by Beck & Bredemeier (2016) is intended as a *unification* of different theoretical – that is, "[c]linical, [c]ognitive, [b]iological, and [e]volutionary" (Beck & Bredemeier 2016, p. 596) – perspectives on MDD. Most importantly, it conceptualizes depressive symptoms as the result of a cognitive-behavioral program that has an evolved function.

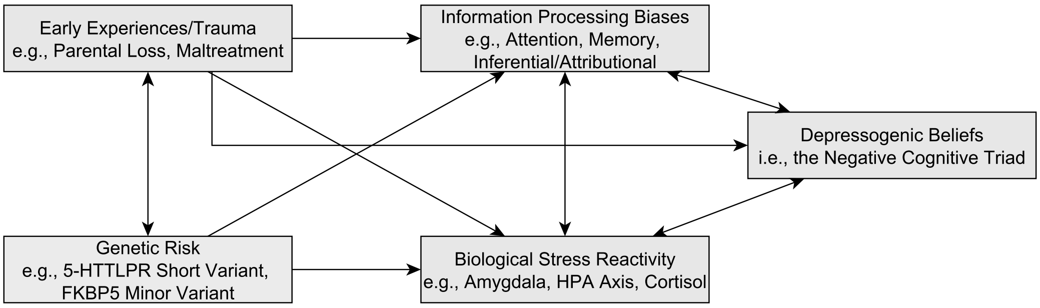
Early Experiences/Trauma
e.g., Parental Loss, Maltreatment
Information Processing Biases
e.g., Attention, Memory,
Inferential/Attributional
Depressogenic Beliefs
i.e., the Negative Cognitive Triad
Genetic Risk
e.g., 5-HTTLPR Short Variant,
FKBP5 Minor Variant
Biological Stress Reactivity
e.g., Amygdala, HPA Axis, Cortisol

Figure 1.2: Factors underlying the predisposition for MDD, put forward by Beck & Bredemeier (2016), slightly adapted.

The first part of the model supposedly explains the predisposition for MDD as follows: There are several cognitive processes implicated in the predisposition for MDD. The experience of *traumatic events* can lead to the development of *information-processing biases* that skew incoming information in a negative fashion, facilitate the retrieval of negatively charged memories and impact the attribution of success and defeat, for example. I take it that the term "bias" is used here to denote a tendency of an individual's processing system that differs systematically from how human beings usually process information, thus leading to problematic outcomes. These problematic outcomes may either be holding more false beliefs than the average person holds, or holding more negatively charged beliefs than the average person holds.[39]

Either together or in isolation, traumatic events and information-processing biases can give rise to *depressogenic beliefs.* According to Beck, depressogenic beliefs result from a particular activation of three *schemas* that form the so-called "cognitive triad".

These three schemas, according to him,

> "simultaneously operate to determine the meaning/value of life events [...] and generate appropriate responses. These include the self-image (lovable vs. unlovable), image of the world (friendly vs. unfriendly, accepting vs. rejecting), and expectations of the future (hopeful vs. hopeless)." (Beck & Bredemeier 2016, p. 597).

Due to the traumatic events and information-processing biases in question, these three schemas are consistently activated, thus generating depressogenic beliefs that represent the individual's self as unlovable, the world as unfriendly or rejecting, and the future as hopeless.

Furthermore, on the biological level, someone's *genetic risk* can interact with her early traumatic experiences in producing *information-processing biases.* All of the aforementioned factors may, collectively or in isolation, lead to *enhanced biological stress reactivity*, which strengthens existing depressogenic beliefs.

Furthermore, in this model, (quasi-)environmental[40], folk-psychological, cogni-

[39] Even though the term "bias" is, in the classical literature on *cognitive biases*, usually held to denote tendencies of individuals that lead to systematic errors in judgements under uncertainty (Tversky & Kahneman 1974) it seems that either (1) "information-processing biases" must denote something different here, since debates around so-called "depressive realism" suggest that there might be forms of depression where the individual's information-processing differs from "normal" information-processing by not being as *inaccurately positive* (Moore & Fresco 2012, p. 497), thus resulting in *more* correct beliefs, or (2) Beck is wrong in claiming that depressed individuals really exhibit information-processing biases (Haaga & Beck 1995). This latter interpretation seems puzzling – that is, *if* depressive realism actually exists, which is unclear (e.g. Allan et al. 2007, Moore & Fresco 2012) –, since the author still refers to *information-processing biases* in his latest model of depression (Beck & Bredemeier 2016).

[40] I am using the term "quasi-environmental" here to account for the fact that traumatic events seem to be hybrid entities, as they might be understood as being comprised of a particular kind of event plus a particular (perhaps intra-individually shared) evaluation of it which singles the event out as *traumatic.*

tive and biological entities are taken to work together in producing the predisposition for MDD. This feature of many clinical psychological and psychiatric models has been discussed in parts of the philosophical literature under the header of "multilevel models" (e.g. Kendler 2005, p. 435).

Maintenance of MDD

How is the *maintenance* of MDD's symptoms explained? As I see it, the explanation works as follows: Most importantly, Beck & Bredemeier (2016) suggest that the symptoms of depression arise as manifestations of an evolved cognitive-behavioral program. This program has a specific *function*, namely, conserving energy in an individual who has just lost her investment in a vital resource. How does this work, specifically?

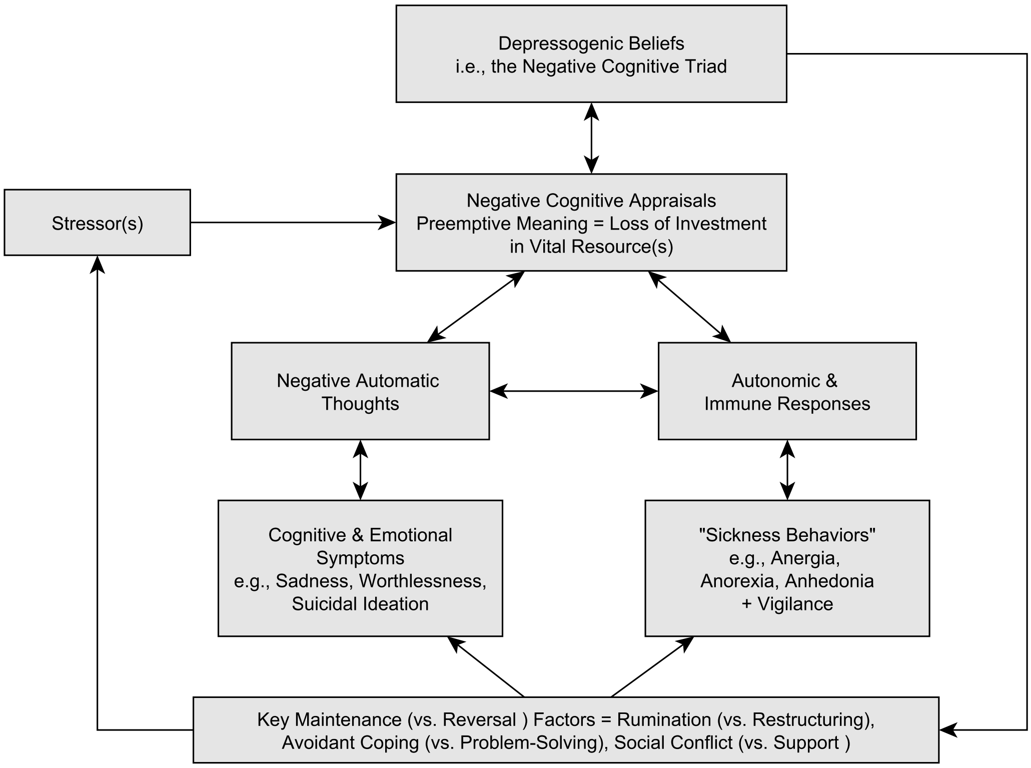
Depressogenic Beliefs
i.e., the Negative Cognitive Triad
Stressor(s)
Negative Cognitive Appraisals
Preemptive Meaning = Loss of Investment in Vital Resource(s)
Negative Automatic Thoughts
Autonomic & Immune Responses
Cognitive & Emotional Symptoms
e.g., Sadness, Worthlessness, Suicidal Ideation
"Sickness Behaviors"
e.g., Anergia, Anorexia, Anhedonia
+ Vigilance
Key Maintenance (vs. Reversal) Factors = Rumination (vs. Restructuring), Avoidant Coping (vs. Problem-Solving), Social Conflict (vs. Support)

Figure 1.3: Model of MDD as due to the execution of an evolved program and maintenance factors stabilizing these symptoms, put forward by Beck & Bredemeier (2016), slightly adapted.

This evolved program is triggered when an individual interprets environmental stressors as indicating that she has lost her investment in a vital resource. One such stressor could be a divorce: Plausibly, when a couple gets divorced, one or both of the (ex-)partners may appraise their current state as implying that they have just lost their investment in a vital resource, namely, their partner. One of the most important reactions to this evaluation are negative *automatic thoughts*. These are pre-conscious thoughts with – at least in the case of depression – negatively charged content (Beck 1995, p. 14-15). They bring about the *cognitive and emotional symptoms* of MDD, that is, as we have seen, sadness, worthlessness or suicidal ideation. Additionally, the thought of having lost a vital resource supposedly activates the individual's immune system, pushing it into overdrive and thereby producing *sickness behaviors*: that is, loss of energy, reduced food intake and diminished interest or pleasure in most activities.

In other words, the general idea is that the *symptoms* of MDD arise because, in the *Environment of Evolutionary Adaptedness* (*EEA*), losing a vital resource threatened an individual's survival and reproduction. The EEA "refers to the selective 'environment' that has shaped humans during the course of their evolution" (Foley 1995, p. 194). Losing a vital resource was threatening in the EEA, because the satisfaction of her vital needs was not ensured anymore, or her ability to reproduce was reduced. Thus, it became necessary for these individuals to conserve energy.

Now, responding to this loss with these emotional, behavioral and cognitive symptoms conserved energy in several ways – one factor, for example, being that the subject is not seen as a threat anymore by other individuals (Beck & Bredemeier 2016, p. 604) –, and thus, enhanced the probability of survival and reproduction. On the other hand, such individuals might, as they showed sickness behaviors, have been seen as easy prey. This possibility, according to the authors, accounts for the increased vigilance that can often be found in depressed individuals (Beck & Bredemeier 2016, p. 605-606).

Manifesting this evolved program is facilitated if individuals already have depressogenic beliefs at the outset, since these beliefs increase the probability of interpreting environmental stressors as indicating that one has just lost a vital resource.

Importantly, instantiating this program *once* is not sufficient for suffering from MDD. For this to be the case, the individual must repeatedly instantiate this program due to the operation of particular maintenance factors. One such factor are maladaptive behavioral strategies, for example, *rumination*, meaning that the individual repeatedly thinks about her past mistakes and wrongdoings, which becomes generalized into a negative self-concept. Another such strategy is *avoidant coping*, that is, avoiding to deal with the problem at hand and instead withdrawing from situations in which one might be confronted with it – among other things, from social situations. This has several consequences: For one, the problem is not solved, but remains virulent. For the other, by avoiding social situations, the individual is not exposed to information that might correct her view of herself. And finally, *social conflict* is a stressor that often strengthens depressogenic beliefs. Importantly, depressogenic beliefs increase the probability that the individual shows maladaptive coping strategies.

Let me make some general remarks about this explanatory model. Firstly, the symptoms of the disorder are characterized as results of an evolutionarily adaptive, but currently harmful, cognitive-behavioral program. It becomes harmful either because the current environment is different from the EEA – that is, because of a *design/environment mismatch* (compare Bolton 2008, p. 80) – or because the individual has certain further properties that stabilize this cognitive-behavioral program. I think that Beck's model is consistent with both views.

Furthermore, those factors that maintain depressive symptoms are conceptualized as being of one of the following two types: Firstly, they may be variants of generic and *evolutionarily adaptive* traits of the individual. For example, having beliefs that produce mild sadness might have been evolutionarily adaptive because it motivates people "to take stock after a devaluing experience" (Beck & Bredemeier 2016, p. 597), and to then potentially change their behavior to deal in a better way with the problem at hand.

Or secondly, they are conceptualized as maladaptive, determinate instances of more generic determinable traits that are actually *adaptive*. For example, being critical of oneself can be adaptive, since it results in questioning one's behavior regularly and thus reacting faster to mismatches between one's behavior and one's goals when compared to individuals who are less prone for self-criticism (Beck & Bredemeier 2016, p. 598). Nonetheless, an extreme tendency for self-criticism is not only not adaptive, but indeed *maladaptive*.

Importantly, this model contains Beck's first cognitive model of depression from 1967 as a part. This earlier model assumes that depressive symptoms are a consequence of "[...] the activation of a set of three major cognitive patterns that force the individual to view himself, his world, and his future in an idiosyncratic way." (Beck 1967, p. 255), the so-called "negative cognitive triad". By "idiosyncratic", the author seems to mean *negative* representations of the self, the world and the future. The negative *cognitive* representation in these three areas result in those *affective* and *motivational* responses that are typical for depression – for example, a negative view of the future is taken to result in depressed mood as well as paralysis of the will (Beck 1967, p. 256). I will discuss the exact relationship of the most recent model of MDD to this early cognitive model in the next chapter.

1.4.3 Making Sense of the Model

Here, I will focus on issues similar to those that I have already taken up in my discussion of the clinical psychological model of OCD.

Very briefly, I claimed that the model of OCD presented by Salkovskis et al. (1998) is a *causal*, cognitive psychological model that encompasses entities of different grain sizes. I tried to draw the reader's attention to the importance of propositional content in that model and to the prominence of folk-psychological vocabulary. Finally, I described the explanatory strategy exhibited by this model, showing how it results in a representation of individuals with OCD as *relatively reasonable* or *rational* by understanding their behavior against the background of particular kinds of beliefs and the context of belief-formation. Let us now see whether these findings translate neatly to this model of MDD.

The first thing to note is that this model is intended to convey the causal structure of MDD by representing how its symptoms reinforce one another and are themselves brought about and maintained by further factors.[41] Clearly, for example, the negative cognitive triad is taken to be causally relevant for negative cognitive appraisals of specific stressors, as well as for certain behavioral strategies that serve to maintain the disorder's symptoms.[42]

[41] Concerning issues of causal circularity, I think that the same points that I have already made above apply to the bidirectional causal arrows in the two depictions (fig. 1.2 and fig. 1.3) of the second model as well: They are best understood as mere shortcuts referring to *two* distinct causal relations that are relevant for different token events of the same type.

[42] Of course, this does not settle everything, since, on this reading, environmental stressors are causally relevant for negative cognitive appraisals *of these very stressors.* This does not seem correct. What seems more natural is to understand the issue here as one of a disposition and its stimulus-conditions: Individuals with MDD, according to this reading, have the tendency or disposition to interpret events as indicating that they have just lost their investment in a vital

Just like our explanatory model of OCD, this model also uses folk-psychological vocabulary. And again, the precise content of these beliefs is taken to be integral to the kind of mental disorder that the person suffers from: It is at least partially the fact that the individual has negative thoughts about himself and his future that makes him suffer from depression and not any other mental disorder.

Interestingly, in the original paper by Beck & Bredemeier (2016), the usage of terms surrounding the concept of function and functionality is mainly restricted to *evolved* functions, while sometimes, neurological functions are mentioned as well, and once in the paper, they seem to use the term with a mainly normative or evaluative meaning.[43] I would like to ignore the usage that seemingly deals with physiological dysfunction for now, since this is of little importance for my purposes. Concerning the second kind of usage, it is interesting to note that the authors appear to think that "the apparent dysfunctionality of severe depression" (Beck & Bredemeier 2016, p. 603) is illusionary if one considers the adaptive value of these syndromes in the evolutionary context. Obviously, this manner of talking only makes sense if severe depression was thought to be dysfunctional in the sense of being detrimental to survival and reproduction. Thirdly, the authors refer to an instrument that was developed in part by Beck, the so-called "Dysfunctional Attitude Scale" (Weissman & Beck 1978, p. 1). This instrument is intended as a measure for pervasive negative attitudes that are typical for depressive patients. If we take this at face value, – i.e., as a description, not as a mere *label* for this instrument – then Beck would be committed to a third, evaluative, concept of "dysfunctionality" that might be spelled out along the lines of harmfulness or the like. Thus, it would seem that this paper makes use of at least three different senses of (dys)functionality, one pertaining to evolved functions, the other to brain (dys)functions, and the third to a feature of attitudes that most probably needs to be spelled out by making use of value-laden terminology. I think that this warrants a closer look at the meaning of functional vocabulary in these clinical psychological models of mental disorders, which I will provide

resource. The environmental stressors in question now simply emerge as the stimulus-conditions for the actualization of this disposition.

[43] I take it that there is a relevant difference between the two: Understanding *function* as an etiological concept might allow for an analysis of the terms content through descriptive facts. This is contrasted by understanding dysfunction or dysfunctionality as harmfulness, which appears to imply a (non-descriptive) value-judgement.

in chapter five.

To sum up, the general explanatory strategy exhibited here (in keeping with Bolton 2008) appears to be to claim that the individual in question exhibits variants of (types of) traits that were adaptive in the past of either the species or the individual, remained stable in the individual until the present day and have become maladaptive in the very recent past. This is clear, for one, because of the very repeated reference of Beck & Bredemeier (2016) to the evolutionary adaptiveness of the depression program, and for the other, because of the assumption inherent in CBT that even maladaptive cognitions serve – or at least: once served – certain functions. One of them being the ability to make sense of the world, as we have seen in the definition of cognitive schemas.

1.5 Conclusions and Structure of the Dissertation

In this section, I will present two kinds of conclusions: I will both give an outlook on this first chapter and draw some conclusions from my two case studies and derive further questions from them. Finally, I will offer an overview of what there is to come. Let me start with the first issue.

In this chapter, I did the following: Firstly, I presented some reasons to be interested in explanatory models of mental disorders *at all*. As I pointed out, these models have received relatively little philosophical analysis (with the exception of Bolton & Hill 1996) so far, despite differing in certain ways from what I consider to be typical psychiatric or neuroscientific models of mental disorders. I then provided some reasons to be interested in two particular exemplary models, that is, one model of depression and one model of OCD. To discuss these two models in detail, it was necessary to first understand what the phenomena in question *are*. After presenting the current diagnostic criteria of OCD and MDD from the recent DSM-5, I presented a descriptive analysis of the two models of interest. As I pointed out, they have particular features that seem surprising at the outset. That is, firstly, they employ folk-psychological vocabulary, secondly, functional terminology figures very prominently in them, and thirdly, they are nonetheless intended as *causal* models.

Very roughly, the central explanatory strategy in these two models consists in identifying features of an individual (e.g., beliefs, behavioral strategies) that were – or at least seemed – adaptive when they were first adopted but became maladaptive for the individual at some later point in time. Nonetheless, these features remained stable over time and currently cause significant harm.

Having now reminded the reader of what I did before, I would like to provide some further conclusions about these two explanatory models, leading up to further questions that I intend to answer in the remainder of this dissertation.

One feature of these models is that they are purely *qualitative*. In other words: In contrast to how models in other scientific disciplines – think of, to use paradigmatic examples, physics or chemistry – are usually set up, these models do not mention relations that are formulated in precise quantitative terms. Nonetheless, the relations that are mentioned in the model are conceived of as *causal* relations.

Relatedly, these models make use of concepts that appear to be quite generic and that are vaguely formulated. This makes them applicable to a wide range of different cases. This feature, I take it, explains how every instance of heterogenous clinical types like depression is supposed to fit under only *one* explanatory model.

Note that these features make the models in question more easily applicable to individual cases in psychotherapy. Purely qualitative models are more accessible for psychotherapists and patients, I take it. As I already pointed out, using relatively generic terms is helpful for psychotherapists who deal with many individual patients on a daily basis who show extremely different syndromes.

To reiterate, I take it that these models are interesting partly because of their unclear status – they may, on the one hand, be classified as outputs of an applied or "application-dominated" discipline (compare Carrier & Nordmann 2011), since the utility of the models in question appears to be very important for how they are formulated. On the other hand, they may also be understood as basic research in an otherwise application-dominated field, since researchers in the disciplines seem to take the models in question more seriously than researchers in more paradigmatic application-dominated disciplines like engineering appear to do.[44] Thus, one question to ask here is what really is primary in the construction of these models – that is, application, or research? I will discuss this question in chapters two and three.

[44] This latter statement is based on several private conversations with researchers in the discipline.

In stark contrast to neuroscientific understandings of this condition, Beck and Bredemeier do *not* conceptualize MDD as due to underlying processes that are picked out by reference to their material structure – such as, for example, prefrontal and limbic systems within the brain (Maletic et al. 2007). Instead, although integrating biological and evolutionary perspectives, their view on the maintenance of the condition has beliefs, behavioral strategies and their dysfunctionality at heart. Similarly, the model presented by Salkovskis et al. (1998) uses only such states that are described on an intentional level.

One question to ask here is what it is that psychologists *mean* when they speak about "beliefs", about "dysfunctionality", or when they implicitly refer to rationality? And, in addition to this, *why* exactly does the language of intentionality occur in these models? This seems surprising, especially given the fact that clinical psychologists had already, before Beck, abandoned cognitive vocabulary, and in particular, vocabulary that refers to such coarse-grained notions like beliefs.

That is, after the rise and fall of behaviorism within psychology (compare, e.g. Watrin & Darwich 2012), one would have expected to see more refined cognitive or information-processing concepts – most plausibly located on the sub-personal level (Dennett 1996, p. 90-96) – represented, as they can be found in many areas of psychology. When compared to the conceptualizations of cognition and affect that are at play there, the concepts employed in my two clinical psychological models appear hopelessly coarse-grained.

In addition, when compared to the models and theories that are to be found within clinical neuroscience, one might wonder about the the benefit of employing intentional concepts, when there seem to be purely causal models available that only make reference to non-intentional, material objects. Thus, when viewed from the point of view of two of its neighboring disciplines, cognitive psychology and clinical neuroscience, clinical psychological models seem to be either hopelessly imprecise or plainly unnecessary. I would like to object to this conclusion.

In fact, issues about the model's apparent imprecision and the apparently unnecessary theoretical surplus structure only arise if we ignore the *pragmatic* goals and purposes that these models serve in psychotherapeutic practice. In other words: Their status as applied science objects in clinical psychology accounts for these apparent theoretical shortcomings.

That is, clinical psychology, unlike other disciplines that investigate mental health and mental illness, does not only develop these models to *inform* its researchers and practitioners, but also with the explicit aim of providing explanations of their patients' mental disorders to them in psychotherapy. This is suggested when, as I already alluded to above, psychologists claim that they take these explanations to be *essential* for successful therapy (Salkovskis 1999, p. S33). I assume that these models are developed with the demands of a particular social context in mind. Since the context in question is one where mental disorders are explained to laypersons, it makes sense that they would incorporate folk psychological vocabulary as well. Partly, that is, to explain their disorders in a terminology that patients understand, and partly for them to be able to intervene *themselves* on their disorder. I will investigate what these demands are and how exactly they may be taken to influence these models in chapter three.

To understand how explanatory practices in psychotherapy work and how they might influence models, it will be helpful to try and understand those concepts that figure essentially in these explanations, something that I will do in chapters four and five. Firstly, one might wonder what *dysfunctionality* means in this context. Although there has been a lot of debate about the notion of "dysfunction" in the philosophy of psychiatry, it has usually been discussed as a component of analyses of the *concept* of mental disorder (e.g. Wakefield 1992). In these debates, "dysfunctional" is often used to refer to mental mechanisms that fail to carry out their evolved function (Wakefield 1992) or to parts of the brain that fail to do what they usually do (Murphy 2006). By contrast, I am more interested in the meaning of "dysfunctionality" as it is *actually used* by mental health professionals, in particular, psychotherapists in their explanatory practices in psychotherapy. I will argue that prior analyses of this term do not do justice to how it is used in explanatory models and psychotherapeutic practice, where an individual is simultaneously represented as mentally ill, and yet, (relatively) rational.

This brings us to another issue, namely, to the topic of *rationality.* One source for the importance of concepts of rationality and irrationality in discussions of mental disorder is Albert Ellis, who developed so-called "Rational-Emotive Therapy" (*RET*) in 1955, which developed to become "Rational-Emotive Behavior Therapy" (*REBT*) later (Ellis 1995, p. 85). Like Beck's framework of

CT, this approach is based upon the assumption that cognitive processes are central in the development, maintenance, and therapy of mental disorders. In Ellis' theory, certain beliefs are identified as the root cause of mental disorders. According to Ellis, the main feature that makes a difference for one's mental health is having rational or irrational beliefs. According to him, mental disorders are ultimately brought about by *irrational* beliefs.[45]

But let me come back to reasons for investigating this concept that pertain to the two case studies presented above. As I already hinted at, these models *normalize* (compare Bolton 2008, p. 52) the condition in question. But there is more, as I see it: Namely, they also *rationalize* the patient's behavior by presenting it as related to specific mental states, more precisely, to certain *reasons.* Thereby, they make the patient's apparently erratic and irrational behavior and his emotional states appear (relatively) reasonable.[46] This representation of patients in therapy is puzzling, though, when we consider how Aaron Beck characterizes depression right at the outset of his 1967 book:

> "Depression may someday be understood in terms of its *paradoxes.* There is, for instance, an astonishing contrast between the depressed person's image of himself and the objective facts. [...] Despite the torment experienced as the result of these self-debasing ideas, the patients are not readily swayed by objective evidence or by logical demonstration of the *irrationality* of these ideas." (Beck 1967, p. 3, my italics)

This passage shows that, for the author, one of the most puzzling features of depression is the fact that these individuals cannot be understood anymore according to what one might call the "default view" of the human psyche, according to which, e.g., people's thoughts and beliefs usually represent the outside world relatively accurately,[47] they have a tendency to avoid suffering, and the like. This quote thus shows the tension arising from the perceived need to stop understanding human beings as rational agents, or, to put it in Dennett's terms, to drop from the *intentional stance* (Dennett 1971).

[45] I will ignore for the moment the question of how problematic this is before the background of theories about epistemic injustice (e.g. Fricker 2007).

[46] Note that I do not wish to claim that individuals with mental disorders *actually* behave in erratic or irrational ways; other than that, these remarks are based on my qualitative interviews, during which several psychotherapists stated that this understanding of being "irrational" or even "crazy" was how many patients of theirs understood themselves when they first came to see them.

[47] Or at least, they are in rough agreement with how other individuals represent the outside world.

In other words: Mental disorders stand in need of explanation, because, prima facie, the behavior of persons suffering from them – including certain utterances about emotional states – *makes no sense*. What does this mean? Apparently, that she does not act in accordance to her (supposed) beliefs and desires – she either acts irrationally or arationally. This allows for several different kinds of explanations (compare, e.g., Dennett 1971, Bolton & Hill 1996):

1. arguing that the behavior is caused by malfunctioning *physiological* mechanisms
2. arguing that the behavior is caused by malfunctioning *psychological* (e.g., information-processing) mechanisms and
3. arguing that the behavior is in fact (relatively) normal and rational, once the individual's belief system is taken into account – and the problematic symptoms have other origins like, for example, a mismatch of environment and design.

Psychologists explaining OCD and MDD seem to choose a combination of strategies (2) and (3): I have already pointed to the supposed normality of psychological processes that are involved in the development and maintenance of mental disorders and the idea that the individual's beliefs are relatively rational to hold, but that they nonetheless may have harmful effects. That mental disorders need not, as was assumed by some authors (e.g. Dennett 1971), necessarily be explained by dropping to a non-intentional level has already been pointed out by Bolton & Hill (1996). This raises questions about the understanding of rationality that psychologists and therapists exploit here. I will provide an answer to this question in chapter four of this dissertation.

Importantly, these models are used mainly for the purpose of explanation, both in psychotherapy and in didactic contexts. As I will argue, when psychologists explain mental disorders, they do so to achieve certain *aims*. When speaking about these explanatory aims, we should distinguish between

1. explanatory aims of the models
2. aims of using these models as part of specific explanatory practices.

While (1) refers to the target phenomenon that stands in need of explanation, (2) refers to aims that the speaker wants to achieve in providing (what looks like) an explanation. In this dissertation, I am mainly interested in the latter category, that is, *aims of explanatory practices*. While aims of the first type are directed at providing knowledge about, e.g., the causal structure of the disorder at play, aims of the second type may be directed at very different issues such as – in the case of psychotherapy – trying to elicit a particular impression that is thought to be beneficial for the patient. Although aims of explanatory practices might be present in the other clinical professions as well, *if* diseases are explained to patients – which apparently happens much more often in psychotherapy –, they have less effect on the form of the explanation that is actually delivered.[48]

In this dissertation, I will focus mainly on practical aims of explanatory practices that arise in psychotherapy. My goal is to understand how these practices work, to philosophically analyze concepts of (dys)functionality and (ir)rationality they presuppose, and to argue that these aims influence model construction in clinical psychology significantly.

What I will do in the following is to clarify how the process of theory construction and development works in clinical psychology. I will start with this in chapter two, where I provide a descriptive analysis of how these two explanatory models were constructed. There, I will stress the high importance of evidence from the psychotherapeutic context for model construction. I will argue that the form and content of these seemingly theoretical models is influenced by practical considerations arising in explanatory practices in psychotherapy.

My aim is to provide the reader with a better understanding of how clinical

[48] We may ask whether the distinction between explanatory aims and aims of explanatory practices is actually specific for the psychotherapeutic context, or whether it can also be applied to, for example, the psychiatric context. As a matter of fact, we may be in a position to make a similar distinction there, too: Very often, it would seem to be much more important to present the patient with an understanding of their mental disorder that makes them believe in the effectiveness of their medication than one that is necessarily accurate. For example, a psychiatrist might consider it appropriate to present his patient with the serotonin hypothesis (Lacasse & Leo 2005), knowing very well that it is incorrect, in order to present a simple – if probably false (Lacasse & Leo 2005) – understanding of why serotonin reuptake inhibitors work. That is, in this context, too, explanatory aims and aims of explanatory practices might diverge. Nonetheless, I believe, pragmatic aims are much more important in explanatory practices in the psychotherapeutic context. For one, this is due to the fact that explanatory practices are much more important for the success of the treatment – in many cases, that is, patients need to be provided with a good justification for why the intervention at issue is helpful, usually, because therapeutic success hinges to a large extent on whether they themselves change their behavior in particular ways. Patients who are administered a particular medication, on the other hand, usually have to do something much less complex. Furthermore, based on informal conversations with psychiatrists and several patients of both psychotherapists and psychiatrists, I would boldly state that psychiatrists explain their patient's conditions much less frequently and in much less detail than psychotherapists.

psychological models of mental disorders are constructed and employed in psychotherapy. While, in this chapter, I have merely offered a descriptive overview of two exemplary models, I will, in subsequent chapters, provide more in-depth accounts of how we might understand the terms that they utilize philosophically. There are two clusters of questions that I am interested in. Firstly, these are questions that pertain to the actual practice: How do researchers in the area and psychotherapists *really* construct and apply these models? How does the context of application influence the construction? The second cluster is about those concepts that are employed here: Which concepts of dysfunctionality and rationality make it possible to understand individuals with mental disorders as *rational*, and yet, as needing to revise particular beliefs of theirs? And what exactly do researchers and clinicians mean when they talk about a feature's *dysfunctionality*? A complete understanding of the model's application is only possible once these conceptual issues are solved.

After this first chapter, where I have given a reconstruction of the two explanatory models of interest, the **second chapter** will deal with how clinical psychology as a scientific discipline constructs and employs these exemplary models. There, I will focus in particular on the kinds of evidence that Beck and Salkovskis relied on in constructing and further developing their models. In the **third chapter**, I will ask how these explanatory models are used by mental health professionals in psychotherapy for the purposes of explaining mental disorders to patients. I will, on the basis of five qualitative interviews with psychotherapists, identify certain pragmatic aims that they appear to have when providing these explanations. The fourth and fifth chapter are concerned with the interrelated concepts of irrationality and dysfunctionality. In the **fourth chapter**, I try to answer the question of how it is possible for mental healthy professionals to simultaneously understand individuals with mental disorders as (relatively) rational in one sense, while simultaneously having it as one of their therapeutic aims to convince their patients of the need to revise certain *dysfunctional* beliefs. Once I have provided an answer to this question, the **fifth chapter** asks how to best conceive of the dysfunctionality of these beliefs, especially given that those beliefs that psychotherapists intend to challenge in psychotherapy by referring to their irrationality are precisely the dysfunctional ones. I will conclude my investigation in the **sixth chapter** by drawing the strands together and discussing open questions.

Chapter 2

The Construction and Development of Models in Clinical Psychology

2.1 Introduction

In this chapter, I will analyze the processes of construction and further development of the two exemplary models that I presented in the first chapter. These processes constitute an interesting modus of research that, to my knowledge, philosophers of science have not turned their attention to so far.

The thesis that I will argue for in this and the subsequent chapter is that the structure and content of these models is influenced by the fact that they are constructed on the basis of observations made in psychotherapy[1]. At the end of this chapter, the reader should have an idea of why I take it that practical considerations arising in psychotherapeutic practice *might* exert a considerable influence on cognitive models in clinical psychology. In the next chapter, I will describe those practical considerations I am thinking of in more detail, provide evidence for believing that specific aims of psychotherapists systematically influence the kinds of explanations they give, and argue that they *actually* formed these models.

By contrast, I will defend two different theses in this chapter. These are a relatively modest claim and a bold one. For one, I will argue that (1) the context of application is important for how these models of mental disorders

[1] Importantly, there are two ways of understanding the concept "psychotherapy". One of these senses includes all talking therapies, while the second one excludes psychoanalysis. The second sense allows for there to be psychoanalytic psychotherapies nonetheless, as these therapies are grounded in psychoanalytic theory, but are not psychoanalysis proper (Gill 1954, p. 772). For the purposes of this thesis, I will understand the term "psychotherapy" in the first, broader sense, that is, as including psychoanalysis.

developed. For the other, I will present some reasons for the more historical thesis that (2) these models were first constructed mainly – even though not *solely* – on the basis of observations gathered in the course of psychotherapeutic practices.[2]

This is related to the idea that both Beck's model of depression and Salkovskis' model of OCD are outputs of *applied* clinical psychological research, that is, research in which the context of application is primary. There are two indicators for this that I would like to already present here: (1) psychologists themselves often talk about clinical psychology as an applied scientific discipline[3] (e.g. McFall 1991), and (2) one main aim of clinical psychology is developing a *product* that allows for intervention in the world, namely psychotherapy. This suggests that clinical psychology is, to some extent, driven by considerations of *utility* (compare Carrier 2011).

This aim of clinical psychology is reflected in the following claim that Salkovskis makes in his 1985 paper, after presenting his model of OCD: "Ultimately, the utility of such a model must rest on its ability to *make a contribution to the clinical assessment and treatment of obsessional patients.*" (Salkovskis 1985, p. 582; my italics). This indicates that one main goal in formulating the model was to contribute to better psychotherapeutic treatment of the disorder.

In addition to influencing the aims of the research in question, the context in which much of the the evidence originated *is* the context of application, as will become clear over the course of this chapter.

It is important to note that I am by far not the first person noting the importance of observations from the clinical context for the cognitive theory of depression. On the contrary, clinical psychology has seen a heated debate on whether Beck's model actually qualifies as a *scientific* model (compare, e.g. Blaney 1977). Often, authors have argued on the basis of an alleged contrast between *clinical* and *scientific theory*, implying that, since Beck's theory is an instance of the former, it cannot *also* be scientific. For example, Teasdale & Barnard (1993, p. 7, my italics) state that "[...] it is, avowedly, a clinical *rather*

[2] What I will not provide in this investigation is a comprehensive analysis of the respective influences of research and therapeutic application on one another.

[3] This is a little bit tricky, since many psychologists use the term to refer to a field that also incorporates practical work (compare, e.g. American Psychological Association, Division 12 1996), in parallel with how "psychiatry" is often used. Thus, strictly speaking, I use the term "clinical psychology" as a shorthand to refer to clinical psychology research.

than a scientific theory."[4]

My interest in pointing out yet again the importance of the context of application for these models is different, though: Instead of wanting to challenge the scientific status of them, I merely want to highlight the importance of the therapeutic context for model formulation and development in order to understand what it means for these models to be the outputs of applied research processes, how they are affected by originating from psychotherapy, an essentially *discursive* practice, and finally, whether this helps to explain particular features of them that seem surprising at first glance: the prominent use of *functional* vocabulary, the focus on normalization and the reliance on *folk-psychological* reasoning in a nonetheless *causal* model.[5]

I decided to provide comparative analyses of only two versions of the two models each and sketch their respective development over time in broad strokes. This was the only way to say something substantive on how these models developed without going well beyond the scope of this dissertation. I will show that a focus on *applicability* explains those changes that these models underwent over time, even though they do not appear to be analogous at first glance: while Salkovskis' model of OCD was subsequently simplified, Beck's model of depression became more complicated. We can make sense of this by suggesting that the practical utility of these two models pulled into different directions.

It will be important to distinguish between (1) those observations that underlie model construction and (2) those observations that are made to test or lend further support to the model. I will argue that those observations that underlie model construction both in Salkovskis' and in Beck's case, are made in a clinical and usually therapeutic context. Those observations that are made to test or lend further support to the model often – but not always – come from studies

[4] Even though the authors seem to, in the following sentences, attribute the view that Beck's theory is clinical *rather than* scientific to the proponents of the theory, this appears to be factually incorrect. For example, Clark & Beck (1999, p. 55) state the following: "[...] we examine Teasdale and Barnard's (1993) claim that Beck's cognitive model of depression constitutes a clinical rather than a scientific theory. [...] we conclude that the cognitive model *can be considered an* ***applied science theory*** and so can be evaluated in terms of its ability to account for relevant clinical phenomena and experimental findings".

[5] Of course, there are different understandings on what exactly it means for something to be applied science. For example, in their book, Teasdale & Barnard (1993) argue that they themselves take an "applied science approach" (Teasdale & Barnard 1993) to the problem of negative thinking that is one symptom of depression, while thinking of Beck's model of depression as *mere* clinical theory. I would, by contrast, be tempted to understand Beck's theory as an instance of applied science, particularly because it was and is still used very often by clinicians, but having inspired a number of studies.

conducted in a more controlled environment and with larger populations.

This chapter has two parts, each of which investigates one of the two models. In the first part, I will describe how Beck's classical 1967 model of depression was constructed. I will then, in broad strokes, show how it evolved to become Beck and Bredemeier's 2016 model of depression. Similarly, Salkovskis' more recent model of OCD from 1998 has a predecessor in the model of Salkovskis (1985).[6] Thus, in the second part, I will show how the first formulation of this model originated and how it was subsequently adapted. This focus on the development of the respective first versions is guided by the idea that the later models are merely different *versions* of these respective first models.

Let me begin by presenting the construction process of Beck's classical model of depression, since his understanding of depression was later extended to other mental disorders, also influencing Salkovskis' understanding of OCD.

2.2 Construction and Development of Beck's Model of Depression

As I have already pointed out, the core of the depression model has not changed much since 1967, when Aaron Beck formulated his first cognitive model of the disorder. Nonetheless, the model expanded substantively over time.[7] Arguably, those factors that guided the construction of his 1967 model are still relevant for the more recent version. Thus, I will start my investigation of Beck's current model of MDD by first considering how his classical model of the disorder was constructed. Secondly, I will investigate how the 2016 model differs from the earlier model and sketch some of the processes and empirical evidence that can account for these changes. I will argue that the way in which the model changed makes it plausible to think that considerations of practical usability influenced the form and content of this model. Since the construction and development of it is tied intimately to the person of Beck, let me begin by presenting some background information on him.

Aaron Temkin Beck was born in 1921. He is a psychiatrist who was trained first in psychoanalysis and later came to develop CT, and later CBT. Initially,

[6] Although I speak of different models here, this is of course not quite correct, since I actually conceive of these objects of research as different versions of one of the same model. The reader is asked to forgive this imprecise usage of the term "model".

[7] Beck sometimes claims (compare, e.g. Beck 2002) that he presented six distinct, but overlapping models of depression until 2002. Here, I take these to be merely aspects of his latest model of depression (that is, the one presented in Beck & Bredemeier 2016).

he was less interested in psychiatry than in other fields of medicine, since he considered both the Kraeplinian model of mental disorder – that is, the idea that mental disorders are essentially natural disease entities (Hoff 2015) – and the psychoanalytic model as unsatisfying. Beck then got into the field by accident. He had started a neurology residency in 1949 and was then required to work in psychiatry for sixth months by the chief of neuropsychiatry due to a shortage of psychiatry residents (Weishaar 1993, p. 14-15). From 1950 onwards, due to taking up a fellowship in psychiatry at a psychoanalytically oriented hospital, he became involved in psychoanalysis (Weishaar 1993, p. 16).

Beck was board-certified in psychiatry in 1953 and was trained in psychoanalysis until 1958. A year later, in 1959, Beck became assistant professor of psychiatry at the University of Pennsylvania Medical School (Weishaar 1993, p. 17). He became an associate professor of psychiatry in 1967. During the same period of time, Beck worked as a psychotherapist in his own practice. In other words, he was simultaneously conducting research on mental disorders and treating patients with those same mental disorders. He wrote his book *Depression: Clinical, Experimental and Theoretical Aspects* in the same year (Weishaar 1993, p. 25), which also incorporates his first model of depression.

Beck is often understood as part of the *cognitive revolution* (Weishaar 1993, p. 27) that took place from the 1950s onwards in different fields of psychology and that primarily consisted in bringing mentalistic concepts back into psychological theorizing, thereby overturning the dominance of radical behaviorism within the discipline (Miller 2003, p. 141).

In my comparative analysis of the two versions of Beck's model, I will show that the model changes over time in a way suggesting that it is intended to offer a comprehensive view of the mental disorder at issue. Furthermore, these changes also indicate that the model's applicability in the therapeutic context was a factor that influenced how the model developed over time.

A central thesis of this chapter is that the main kinds of empirical evidence relevant for the construction of this model were clinical observations from the context of psychotherapy or psychoanalysis of depressed patients in the author's practice (Beck 1967, p. 209). Based on the available printed sources, this thesis is very plausible, but backing it up would require much more detailed historical investigations, which cannot be provided in this dissertation. However, the following discussion lends plausibility to, and illustrates, my thesis. I will focus particularly on observations made within psychotherapy that influenced Beck's model of depression. While the historical details of the process of model construction will not be covered here, the current chapter takes this as a given.

I think that we may understand the most important difference between those two versions of the model as a difference in how much they focus on applicability or practical utility of the model: The first version emphasizes *theoretical* considerations in representing the causal relations between the disorder's symptoms. In so doing, it focuses on underlying cognitive schemas as a predisposing factor of the disorder. By contrast, the later version of the model not only explains the structure of the disorder, but it identifies more factors relevant in its causal history. More importantly, it points to *maintenance factors*, in particular, the patient's behavior, that may be intervened upon in successful treatment or by the patient herself. These are not mentioned in the earlier model at all.

This section is structured as follows: I will begin by describing Beck's early model of depression. Secondly, I will show how it was first constructed, focusing particularly on the kinds of evidence that were relevant in this process. In a third section, I will describe how the more recent model of depression that I described in the first chapter differs from this first model. There, I will point towards factors from the context of application that I take to be at least partially responsible for this development.

2.2.1 Beck's First Model of Depression

Beck's 1967 model of depression is based on the idea that its symptoms are due to a *thinking disorder* – brought about by the so-called "primary triad" (Beck 1967, p. 255) or "negative cognitive triad" (Beck 1967, p. 255-256). The idea is, as Beck states, that "[t]he disturbances in depression may be viewed in terms of the activation of a set of three major cognitive patterns that force the individual to view himself, his world, and his future in an idiosyncratic way." (Beck 1967, p. 255).

What exactly is meant by "cognition" here? As the author writes: "The term cognition is used in the present treatment to refer to a specific thought, such as an interpretation, a self-command, or a self-criticism. The term is also applied to wishes (such as suicidal desires) which have a verbal content" (Beck 1963, p. 326). Thus, according to Beck, cognitions usually have propositional content.[8] It thus seems that the class of cognitions is slightly broader than the class of thoughts, at least if we think of thoughts as also having propositional content, an assumption that I will take for granted here. Importantly, it is used to refer to both conscious and sub- or pre-conscious mental states.

The first component is a pattern – or schema – of viewing one's self in a negative way. The second component is the cognitive pattern of interpreting the outside world in a negative manner, while the third component negatively skews one's view of the future. Beck assumes that it is the activation of these three cognitive patterns that cause the affective and motivational symptoms of depression via biasing the patient's thinking in a negative manner (Beck 1967, p. 255). This influence on emotion is supposedly due to so-called "automatic thoughts" with negative content (Beck 1967, p. 321-326), that is, thoughts that occur in the individual as a reaction to particular stimuli. The content of these thoughts is determined by the cognitive schemas they result from. They thus express negative value judgements about one of the three areas of interest and supposedly influence the patient's emotional life by leading to emotions such as sadness, hopelessness and the like. In several of his writings, Beck describes this as the cognitive triad in depression being "primary" (e.g. Beck 1967, p. 255–257).[9] In a nutshell, the idea appears to be that depression occurs in case an individual exhibits these three negative cognitive patterns – usually due to specific kinds of prior experiences in life – and (2) these negative cognitive patterns are *activated* by a specific event.

According to this model, the concept of "information-processing bias" must be understood as a relatively objective one: According to it, the information-processing of individuals suffering from depression is *objectively* skewed, resulting in more false beliefs about the world than other people hold on average. This becomes clear when Beck states that "[a] crucial characteristic of these cognitions [that is, the verbalizations of the depressed patients] is that *they*

[8] I am offering this reconstruction of Beck's statement, because we may want to distinguish verbal languages from nonverbal languages (like sign languages) and allow for sign languages is to have propositional content that is represented in thought.

[9] Interestingly, it is discussed quite a bit in the psychology literature whether this idea of cognitive primacy in mental disorders is really about cognitive factors *causing* the disorder, or whether it is rather about conceptual primacy (Weishaar 1993, p. 63). Since I cannot solve this issue here, let me continue by using the perhaps less suggestive term of cognition being *more basic*.

represent varying degrees of reality distortion." (Beck 1967, p. 233, my italics) From this, he concludes that a so-called "thinking disorder", a "bias against themselves" (Beck 1967, p. 234) is present in *all* forms of depression.

2.2.2 Constructing the Model of Depression

According to Marjorie Weishaar, a close working colleague of Beck who has collaborated with him on several occasions and authored a biography on him, Beck describes his process of theory construction as follows:

> "He begins with observations, often as much about himself as his patients, [...] develops ways of measuring these observations, formulates a theory if the observations are validated by a number of cases, designs interventions that are congruent with the theory, and continues to assess whether the theory is confirmed or disconfirmed over time and through further experimentation." (Weishaar 1993, p. 22)

This indicates that Beck's procedure of model construction is based essentially on observations made in relatively uncontrolled settings. That is, two crucial kinds of evidence are his self-observation and observations made in the context of application.

Which kind of context is that, precisely? That is, if Beck really used evidence from psychotherapeutic sessions, what did these sessions look like? At the time of setting up his earlier model, Beck carried out psychoanalysis with his patients.

In his landmark paper on depression, Aaron Beck describes the approach he took as follows:

> "Face-to-face interviews were conducted during the periods of time when the depressions were regarded as moderate to severe in intensity. The author was active and supportive during these periods. Formal analysis was employed for the long-term patients except when they appeared to be seriously depressed; the couch was utilized, free association was encouraged, and the psychiatrist followed the policy of minimal activity." (Beck 1963, p. 325)

Let me unpack the information that is condensed into a few sentences here. Firstly, it is useful to point out what he means by "formal analysis". This denotes a kind of psychoanalysis that is administered by a professional to a patient in a somewhat professional or formal setting. It thus is distinct from psychoanalysis as applied to oneself (Beiser 1984).

Usually, psychoanalytic sessions take place in the analyst's consulting room for several times a week, and with a session length of approximately 50 minutes. "Using the couch" means that the patient is lying on his back while the analyst sits behind him, out of the patient's sight. This supposedly frees both patient and therapist from being distracted by the other person's reactions to their statements and thus enables *free association*, in which the patient reports everything that occurs in his stream of consciousness, with the therapist only giving prompts. That is, she asks questions or offers interpretations. When Beck says that he "followed the policy of minimal activity", he refers to the ideal of psychoanalysis that the analyst as a person with his own individuality and emotional reactions should stay in the background as much as possible, such that both the analyst and the patient can focus as much as possible on the patient's mental processes (Milton et al. 2011, p. 5-6).

As a psychoanalyst at the time, Beck started out with a particular psychoanalytic model of depression that he tried to test – and to validate – empirically. He started out with an investigation of psychotically depressed soldiers, whose thoughts and ideas seemed to suggest that they had self-punitive wishes (Beck & Valin 1953, p. 352). This led him to believe that depressive states are indeed due to so-called "inverted hostility", a hypothesis that was common among psychoanalysts at the time (Beck 1967, refers, among others, to Freud and Abrahams). The proposition of one brand of psychoanalysis on the origins of depression was that individuals with depression had hostile emotions such as anger that was originally directed against others and becomes directed against themselves, leading to a so-called "need to suffer" (Beck & Hurvich 1959, p. 51). It is a relatively common observation that depressed patients engage in something that can be understood as "self-tormenting" (Freud 1957, p. 250) by, for example, engaging in self-criticism, having suicidal wishes, losing libido and the like (Beck 2008).

Beck tried to test this hypothesis empirically in his so-called "dream studies" (Beck 2019, p. 17), a series of studies that investigated the dream content of depressed individuals. His motivation to provide empirical evidence for the correctness of psychoanalysis had to do with the fact that, having received psychoanalytic training, he had undergone psychoanalysis himself. Beck was as he remembers, “totally committed to the theory and therapy” (Beck 2019, p. 16), but dissatisfied with the lack of scientific basis of psychoanalysis (Weishaar 1993, p. 17). Thus, we may conclude that a kind of research that aimed at providing empirical evidence for the correctness of psychoanalytic propositions was not widespread at the time.

In his studies, Beck intended to find out whether depressed patients would show more so-called “masochistic” dream content than healthy individuals (Beck 1967, p. 170). Masochistic content was considered to be any content that somehow devalued the dreamer. This was considered to be “masochistic”, since the dreamer, as Beck & Hurvich (1959, p. 51) state, “‘makes’ himself the recipient of criticism, rejection, or other types of discomfort”, thus leading to unpleasant emotional experiences. In the background of this stands the motivational model of psychoanalysis, according to which the symptoms of mental disorders are due to unconscious conflicts of the individual that are different from what can be observed on the level of the person’s behavior. According to the theory, this different content of the underlying conflict should be accessible via the patient’s dreams, since these were – at least in classical psychoanalysis after Freud’s *The Interpretation of Dreams* from 1899 – thought of as one of the main ways for the unconscious to express itself. In particular, it was thought that dreams would serve the purpose of fulfilling wishes that the dreamer had but could not fulfill in real life (see Freud 1982).

When testing this hypothesis with depressed and healthy individuals in his private practice, Beck found that, indeed, the so-called “neurotic depressed” patients showed more dreams with this particular kind of content than nondepressed individuals (Beck & Hurvich 1959, p. 51).

Another prediction of the psychoanalytic theory of inverted hostility was that themes like guilt or hostility should be prominent in the dream content of depressed individuals. Since Beck could not find a particularly strong presence of such themes, he felt to need to conduct further studies to further put this theory to the test (Beck 1967, p. 171).

According to Salkovskis (1996), Beck’s conceptualization of depression really changed once the author conducted further studies in which he assessed the impact of failure and success in specific tasks on depressed individual’s self-esteem. The psychoanalytic model predicted that depressed people would react negatively to success, because it would contradict their self-punitive wishes. But the results were quite the opposite: Depressed individuals actually reacted positively to success, that is, show higher self-esteem and better performance on the next task. He found that they would, in fact, react even more positively to positive feedback than non-depressed individuals. This seemed to indicate that depressed individuals did not actually *wish* to fail, but that their self-image was adjusted depending on whether they experienced failure or success.

At the beginning of the 1960s, when he noticed that his experiments failed to back up the hypothesis of inverted hostility, he went back to his initial dream studies, searching for a simpler explanation of his findings than the one provided by psychoanalysis Beck (1967, p. 185). He suggested that the content of patient's dreams could also be interpreted as "contain[ing] the same themes as the patients' conscious cognitions [...]" (Beck 2008, p. 1), namely thoughts that contain a negative value-judgement of the self. In a personal communication with Salkovskis, Beck points out that he noticed in therapy that these dreams could also be interpreted as expressions of the waking concerns of the patient. For him, this constituted a *simpler* understanding of depression. Beck then started doubting the whole motivational model of psychoanalysis (Salkovskis 1996).

When Beck's research indicated that the psychoanalytic theory made the wrong predictions, he abandoned the idea of an underlying, unconscious conflict being responsible for depressive symptoms. Instead, he hypothesized, partially on the basis of clinical evidence from his own therapeutic sessions, that depressive symptoms and depressive dream content may actually be due to a "thinking disorder" (Beck 1967, p. 269), in which specific patterns of thought are responsible for a negative bias in the interpretation of the patient's everyday experiences (Beck 1967, p. 255).

For him, the decisive evidence for the superiority of the new model over the psychoanalytic model was the *therapeutic effectiveness* of interventions that were based on it: According to him, the reappraisal and correction of the patient's misinterpretations resulted in a much quicker decrease of depressive symptoms in psychiatric patients than psychoanalysis (Beck 2008, p. 2). That is, the *utility* of therapeutic methods that were based on this model was decisive evidence in favor of it.

In my understanding, the difference between cognitive models and psychoanalytic models of mental disorders is best understood as hinging on what is thought to be *at the core* of, or to ultimately cause, the mental disorder in question. While psychoanalysis posits that what is ultimately responsible for the symptoms of mental disorders are specific unconscious *conflicts* that need to be resolved in order for the symptoms to disappear (e.g. Milton et al. 2011), cognitive models posit that mental states or structures with particular cognitive content *bring about* or *maintain* the symptoms in question (e.g. Beck 1967, p. 239). This also leads to different approaches to treatment, of course. Having merely described how Beck came to develop his model of depression, I will now point out why I believe that the context of application was an invaluable source of evidence in this construction process.

As we have seen, Beck worked as both a researcher and a therapist. This makes it highly likely that these two contexts influenced one another, and in fact, Beck states this explicitly.

More precisely, the author points out that one of his main hypotheses – that "highly charged *dysfunctional attitudes*" (Beck 2008, p. 2) are responsible for the negative bias that can be found in depression – was based primarily on "clinical observations *supported by* research" (Beck 2008, p. 2; my italics).

What exactly does it mean for a hypothesis to be based upon clinical observations that are *supported by* research?

Let us take a closer look at the apparent meaning of terms like "clinical observation" and "research". In my understanding, "clinical observation" refers to unguided observations made during the course of therapeutic sessions with his patients. This kind of observation stems from the point of view of someone who is himself deeply invested in the search for possible causes of illness, in the interpretation of the patient's utterances and, in particular, in efforts to intervene upon those factors identified as harmful or illness-inducing. By contrast, it seems that the term "research" is used here for a kind of activity that involves the *guided* observation of larger groups of individuals in the clinical context as well as experimental work in the laboratory. Thus, the main kind of empirical evidence that was relevant for the *construction* of Beck's model stems from the psychotherapeutic context.

The influence of clinical observation on Beck's theory becomes clear when considering that the concept of so-called "automatic thoughts" that is central to his model of depression, is based upon the case of one particular patient from 1959. This patient apparently criticized Beck angrily when free-associating in therapy. When asked about his feelings, the patient reported that he felt guilty because he had particular kinds of self-critical thoughts. This gave rise to Beck's idea that a secondary train of thought, present in the mind beyond the more easily accessible, conscious thoughts, would be responsible for certain kinds of emotional reactions to be found in individual people (Diffily 1991, cited after Weishaar 1993, p. 19). Usually, the individuals are not aware of these automatic thoughts until asked about them by a therapist. They can thus be described as *pre-conscious* (Beck 1967, p. 20).

Furthermore, as I pointed out above, Beck relied not only on the introspective reports of his patients, but also on *self-observation* or *personal experience* in the formulation of his theories. For example, he thinks that by experiencing moderate depression, he got to understand this disorder better (Weishaar 1993, p. 22).

Thus, much of the evidence of relevance in the model's construction process came either from the psychotherapeutic context – and thus, to a large extent, introspective reports from patients – or from his own introspection and self-observation.

How did prior *theoretical* work influence the construction process of this model?

Beck readily admits that he was influenced by psychoanalytic theory in developing his model of depression. According to him, he only asked for "meanings" (Diffily 1991, p. 25, cited after Weishaar (1993), p. 51) in psychotherapy because of his psychoanalytic training. Asking for meaning heavily influenced his model of depression.

The influence of psychoanalytic theory on Beck's theorizing is also present in his etiological model: There, early experiences occupy a central stage, similar to how they are understood within psychoanalysis (Milton et al. 2011, p. 17). According to Beck's model, *traumatic* events occurring early in life predispose individuals to developing depression later. This seems to be ultimately rooted in the idea from psychodynamic theory that particular kinds of events happening in someone's childhood can influence the individual to such a degree as to make the development of mental disorders later in life much more likely (Milton et al. 2011, p. 22).

Nonetheless, there is a substantial difference between the psychoanalytical conception of the unconscious and Beck's understanding of it. In psychoanalysis, unconscious processes can only "be inferred from the effects" (Milton et al. 2011, p. 17). In Beck's framework, it is assumed that at least specific preconscious entities, most importantly, automatic thoughts, are in principle observable by introspection. Cognitive schemas – the more fundamental entities – are not directly observable, and extremely are hard to bring into awareness. Thus, the difference between the two is one of degree: they differ in how closely tied observable processes are to the theoretical processes in question.[10]

Similarly, in Beck's theory, the idea of cognitive schemas is combined with the view that there are two processing systems in human beings, one of them fast, unconscious and sparing of resources, the other slow, conscious and demanding of cognitive resources. Beck claims that, in its original form, this idea goes back to Freud (Beck 2008).

One of the most important influences on Beck was George Kelly, in particular, his concept of "constructs" that Beck used for some time. This concept is based on the idea that human beings function like scientist insofar as they intend to predict and control their environments, and that they understand their experience in terms of certain preconceptions. Thus, they make use of *constructs*, that is, cognitive structures that they create and then apply to reality (Kelly 1991, p. 3-7). Beck later abandoned this term, because he understood the cognitive structures he was interested in as not necessarily bipolar and started speaking of "cognitive schemas" (Weishaar 1993, p. 20).

To recapitulate: much of the empirical information relevant for model construction is obtained in the *context of application*. This information is supplemented by the results of more empirical investigations that do not take place directly in the therapeutic context. A large part of the theoretical information of relevance here stems from psychoanalytic theory.

Let me now ask the same questions for the recent model of depression by Beck & Bredemeier (2016).

[10] Although this distinction is extremely blurry, I think that it is actually the best way to make sense of the self-understanding of many cognitive and cognitive-behavioral therapists. That is, the theoretical entities that they posit seem to be more closely linked to actual observable processes than those of psychoanalysis. (Or they are, in practice, usually thought of as easier to operationalize.)

2.2.3 The More Recent Model of Depression

Beck's model evolved considerably over the course of almost fifty years from his first formulation to the most recent one (as described in Clark & Beck 1999, Beck 2008, Beck & Bredemeier 2016). Nonetheless, I take it to be most reasonable for the purposes of this investigation to restrict my focus to the differences between Beck's 1967 and 2016 model of depression.

Beck's more recent model of depression is intended to bring together different perspectives on the phenomenon. It can nonetheless be understood as a clinical psychological and cognitive model, because the most important factor is *cognitive*: That is, biological, genetic and other factors predispose an individual for depression in virtue of increasing the probability for individuals to develop *depressogenic* schemas, cognitive biases or thought patterns. Just consider that the only factor mentioned *both* in their account of the etiology of MDD and the maintenance of the disorder are precisely these *depressogenic beliefs*. Thus, the causal factor that is effective in maintaining the disorder are the beliefs in question, *not* the physiological or biological – including evolutionary – factors.

In the paper of Beck & Bredemeier (2016), the attitudes in question are referred to as "dysfunctional attitudes", in keeping with a paper by Weissman & Beck (1978) where the authors speak repeatedly of "dysfunctional attitudes" and "dysfunctional beliefs" as underlying those characteristic distortions that can be found in depressive individuals. For now, we may understand "dysfunctional beliefs" as *harmful* beliefs. Although I will argue in chapter five that dysfunctionality needs to be understood in a slightly more complicated manner, this may suffice for the moment.

One might wonder to what extent the more recent model of Beck and Bredemeier is nothing more than an *extension* of Beck's earlier model of depression described above. This question is important to me primarily because I will argue that the context of application was important for the construction of this model and *still* is of relevance for its form and content.

Although the more recent model of the disorder's *etiology* (as depicted in fig. 1.2) and *maintenance* (as depicted in fig. 1.3) is more comprehensive than the original, the classical model of Beck already incorporates the pathways from the famous *negative cognitive triad* (Beck 1967, p. 255-256) to negative judgements of the self – that is, negative automatic thoughts (Beck 1967, p. 273). Furthermore, the pathway from *negative automatic thoughts* to *cognitive and emotional symptoms* was part of his early understanding.

But which changes did the model undergo? I will start with changes in Beck's model of the disorder's *etiology*. In his 1967 model, Beck describes the vulnerability to depression as due to the three cognitive schemas mentioned above. Furthermore, he specifies them as consisting of "generalizations [the individual] has made on the basis of his interactions with his environment" (Beck 1967, p. 275) that are inactive in the individual until a specific stimulus event occurs. When active, the main effect of these patterns is to negatively bias the individual's self-concept. Thus, it seems that the cognitive components that are mentioned in Beck's and Bredemeier's more recent model (compare fig. 1.2) are virtually identical with those Beck proposed already in 1967: According to his recent model, traumatic early experiences lead to information-processing biases and to depressogenic beliefs – the latter also being known as the *negative cognitive triad*. It is noteworthy that in the recent model, information-processing biases *interact* with, but do not necessarily occur as *a consequence of*, the negative cognitive triad. The factors that are missing from the earlier formulation are the biological factors or physiological factors – that is, *genetic risk factors* and *biological stress reactivity*. These supposedly increase the vulnerability to depression as well.[11]

Thus, the classical model of depressive etiology shares a surprising number of features with the recent model of the disorder. I think that these observations support my view that the recent model of depression is merely another, more mature *version* of the classical model. That is, its central assumptions remain the same, and further details are added.[12] While the recent model contains

[11] I use the term "supposedly" here, because the importance of genetic factors for the predisposition for depression has already been called into question by other researchers in the field. For example, in a private conversation, a professor of clinical psychology told me the following (anonymously, and in German), referring to the model of Beck & Bredemeier (2016): "So, this is a little bit, 2016, it is already three years old, or something like this, there are still working groups that propagate this, but actually, there is a big meta-analysis from '17 that has shown this to be primarily an artefact. Back in the day, people thought that this polymorphism interacts with stress and then, in the interaction, something like depression develops. One does not believe this anymore. Yes? So, especially in the domain of these genetic or neurobiological factors, we have to add a big question mark, and insofar, this box has to be put out of the model."

[12] As a matter of fact, I am not alone in thinking so – in a private conversations, one professor of clinical psychology and psychotherapy stated something very similar. More precisely, he said this, referring to the model of Beck and Bredemeier from 2016 (I translated his statement from German into English): "So, this model is much older, and every ten years or so, another box is added."

further causal factors, its explanatory core remains unchanged.

The thought motivating the construction of the recent model of depression is that, according to Beck, the first model of depression – and later versions of it – had been tested sufficiently, lending support to the proposed causal factors. In Beck & Bredemeier (2016), the authors point to many studies that back up this model. Furthermore, in empirical studies, additional causal factors that seem to be relevant for the predisposition, development and maintenance of depression have emerged. These factors were also incorporated into the new model. This becomes clear when Beck states that "[...] the cognitive model of depression buttressed by years of systematic research has grown to maturity [...]" (Beck 2008, p. 4). This is why, in that specific paper from 2008, "it seems timely and appropriate to compare it with the burgeoning findings in neurogenetics and neuroimaging." (Beck 2008, p. 4).

The model was found to fit well with many empirical studies that were conducted on the phenomenon of depression. For example, in Clark & Beck (1999), the authors review over 1,000 studies on depression that are of relevance for the cognitive model and cognitive theory of the phenomenon (Beck 2008). Additionally, the model was repeatedly adapted over time to fit further empirical findings on depression.

It is an intriguing question what exactly it means for these models to be "[...] buttressed by years of systematic research [...]" (Beck 2008, p. 4). The kind of evidence that Beck responds to in further developing his model of depression is of many different kinds. There are, firstly, those studies that actually *test* parts of the model by first deriving predictions from it and then testing those through, e.g., *prospective* studies in which certain features at one point in time predicts the occurrence of depressive symptoms at a later point in time (see, e.g. Harkness & Lumley 2008, where the authors found early life stress to be positively related to cognitive vulnerability to depression and to later depression) or the like. There are several group comparisons of depressed individuals and nondepressed controls (see, e.g. Clark et al. 1998, where psychiatric inpatients with depression were compared with chronically medically ill depressed patients and nondepressed controls). Secondly, there are studies that investigate how well CBT works for patients, and the positive results of these studies are often interpreted as providing evidence for the correctness of the model (for example, Beck & Bredemeier (2016) quote Cristea et al. 2015, a meta-analysis that provided evidence for the primacy of cognitive change in symptom change in depression through investigating CBT, other psychotherapies, and medication). Thirdly, there are many studies that are *consistent* with the model of depression or that have results which support some of its assumptions, without themselves being proper *tests* of these models (for example, Beck & Bredemeier, in their 2016 paper, refer to Rao et al. 2010, who showed that early life stress results in smaller hippocampal volume that may be a physiological correlate of increased vulnerability to depression).[13]

Beck's expanded model has changed from a solely cognitive model to a cognitive-behavioral model. In the recent model, *behavioral* strategies play an important role, in contrast to their absence from the classical model of depression: For example, *adaptive behavioral strategies* are mentioned as factors that reduce the number of stressors in the environment. This is due to the influence of large amounts of experimental and theoretical work on the importance of behavior in the development, maintenance, and treatment of depression (Beck cites, among others, Hammen 2006, in which the author, among other things, summarizes stress generation research, that is, research dealing with how depressed individual's behavior leads to higher stress than the behavior of healthy individuals).

The work of Seligman (1972) on learned helplessness in other animals also influenced Beck's current model of MDD by way of suggesting that the depressive reaction might be of evolutionary origin (Beck & Bredemeier 2016, p. 604). Beck thus points to the similarity in the depressive reaction among different species – most importantly, the reduction in the individual's activity.

He assumes that it was *evolutionarily adaptive* to react to loss of investment with depressive symptoms. He suggests that the depressive reaction is caused by the absence of a *resource* that usually ensured that the individual's needs were met – not by the absence of positive reinforcement *per se*. Furthermore, this depressive reaction is brought about precisely *because*, in the environment of evolutionary significance, it *conserves energy*.

This similarity in reducing one's activity in reaction to particular kinds of stim-

[13] As I have already pointed out, the evidence base that Clark & Beck (1999) cite in support of the cognitive theory of depression contains over 1,000 studies, and in later papers, for example, Beck & Bredemeier (2016) and Beck (2008), many more are added to the list. Thus, the studies that I quote here can only serve as individual examples of a general trend. For practical reasons, it was not possible for me to review all the studies that the authors refer to.

uli thus strengthens the plausibility of the view that the depressive reaction has the function of conserving energy. Just think of Seligman's famous experiments with dogs that were repeatedly exposed to "uncontrollable traumatic events" (Seligman 1972), that is, electric shocks, and, in reaction to this, eventually stopped trying to avoid this situation even when there were possibilities for escaping the situation.

The idea that autonomic and immune responses and their resulting behaviors play a crucial role in the symptomatology of depression comes from a parallel between normal so-called "sickness behaviors" and the kinds of behavior that are characteristic of depression. The term "sickness behavior" refers to kinds of behavior that are shown by individuals of different vertebrae species when they contract a disease, including lethargy, reduction in food intake, reduction in grooming, and the like (Hart 1988, p. 123). These kinds of behaviors are also commonly associated with the depressive syndrome in human beings. There is experimental evidence to the effect that negative thoughts can result in a heightened secretion of cortisol, thus leading to an activation of the immune system. Immune system activation then supposedly results in the sickness behaviors of interest (Beck & Bredemeier 2016, p. 607). If sickness behaviors have the function to promote energy conservation – a thesis that seems to be quite well confirmed (e.g. Johnson 2002) –, and if the analogy between these behaviors and depressive behaviors goes through, then the depressive symptomatology really is about the conservation of energy.

I take it that mentioning both *underlying factors* such as dysfunctional beliefs and *maintenance factors* such as specific dysfunctional behavioral strategies reflects the intention for these models to be useful for clinical practice: When treating someone's mental disorder, the most common procedure is to try and change *both* the underlying dysfunctional beliefs of the individual *and* to intervene on her problematic behavioral or cognitive strategies that partially maintain the symptoms. This suggests that one of the reasons maintenance factors are included is because they can be used by the therapist to construct a psychotherapeutic intervention, and to help an individual patient to identify factors to intervene upon herself to reduce her symptoms. This last strategy is plausibly dependent also on the therapist's ability to pick out causal factors of a particular type. That is, the therapist needs to be able to pick out factors that the individual can *in principle* intervene upon herself or that she can at least intervene upon with the help of the therapist. In this context, it would not be of much help to identify neural factors that are hard to intervene upon or look like they are hard to intervene upon.

In a nutshell, Beck's model of depression has changed from a solely cognitive model to a cognitive-behavioral model of the disorder, in which behavioral factors are represented as maintenance factors of the disorder. This plausibly has something to do with the need for intervention that arises in the context of application. As some psychotherapists have emphasized in personal communications, while thoughts and beliefs *can* in principle be intervened upon, it is much easier to intervene on someone's *behavior*. We might understand these changes over time as an instance of simplification, but also of idealization – maybe even in the sense of actual *distortion* – (Frigg & Hartmann 2017) for the purpose of usefulness in therapy.

The inclusion of maintenance factors into explanatory models distinguishes cognitive and cognitive-behavioral models from most disease models in general medicine and neuropsychology. Furthermore, there is a distinction to be made between such models from CBT and psychoanalytic models (at least on an orthodox interpretation), where the symptoms can be eliminated only by intervening on the root cause of the disorder, that is, an underlying conflict.

Thus, it seems that the changes from the classical model of MDD to the recent model of the disorder underscore the influence of the context of application on this model. In particular, the development from a rather theoretical or conceptual model focused on the internal structure of the disorder towards a model that incorporates maintenance factors seems to indicate that considerations of applicability influence which factors are incorporated into the model.

With this interim conclusion in mind, I will now discuss the construction and further development of my second exemplary model: Salkovskis' model of OCD.

2.3 Construction and Development of Salkovskis' Model of OCD

Paul Salkovskis understands himself as a clinical psychologist, but also as a cognitive-behavioral therapist and mental health professional (Salkovskis 2019). Born in 1956 – and thus, being 35 years younger than Beck – he graduated in 1979 from Kings College Institute of Psychiatry, Psychology and

Neuroscience in London, where he also worked under Jack Rachman. After qualifying as a clinical psychologist, he worked a full-time job as a clinical psychologist in a psychiatric clinic. Salkovskis also spend time with Aaron Beck during that period. According to him, it was the intersection of these factors that resulted in his paper from 1985 (Paul Salkovskis, personal communication).

In 1985 – and *after* writing his first paper on OCD (Paul Salkovskis, personal communication) –, he began working at Oxford as a Research Clinical Psychologist (*Vita of Professor Paul Salkovskis* 2017).

By the time Salkovskis was trained as a psychologist at university, first CT and later CBT became more and more established. Nonetheless, according to him, some parts of psychiatry and clinical psychology were still dominated by psychoanalysis on the one hand and radical behaviorism, on the other in the 1980s (Salkovskis 2019).

Salkovskis has conducted research on anxiety disorders in general, and on panic disorder, agoraphobia, OCD, health anxiety and specific phobias in particular. He is professor of Clinical Psychology and currently the director of the Oxford Institute of Clinical Psychology (*Vita of Professor Paul Salkovskis* 2017).

He clearly thinks that these models should be useful in the context of application. Salkovskis points out repeatedly that individualizing his model of OCD is essential for CBT (e.g. Salkovskis 1999, p. S33). He seems to claim that the model itself should be used within psychotherapy to explain mental disorders to patients.

Additionally, in his 1985 paper on OCD, he argues that the utility of explanatory models of mental disorders relies on whether they contribute to a better treatment of patients.

This focus on the usability of models of mental disorders in psychotherapy is, I take it, important for the way he sets up his first model of OCD and for how that model further develops over time.

Salkovskis formulated his first cognitive model of OCD when he was still working as an NHS clinical psychologist in a psychiatric hospital in Leeds. His applied work with patients – which was at least partially *therapeutic* work – strongly influenced his theorizing (Paul Salkovskis, personal communication).

Beck's work also heavily influenced him. In fact, he refers repeatedly to Beck in his papers on explanatory models of OCD (e.g., in Salkovskis & Warwick 1985).

To make sense of this, it is important to know that Beck's model was, after Beck first formulated it to explain depression, subsequently used to explain other mental disorders as well. While Salkovskis adopts some key ideas of Beck's cognitive model he seems to consider Beck's understanding of OCD as incomplete (Salkovskis & Warwick 1985).

In my comparative analysis of the two models of OCD by Salkovskis, I will aim to show that the later version of his model is simpler and more idealized than the earlier version of it. I will argue that this supports my hypothesis that the model has, over time, changed to become better applicable within clinical practice.

This section is structured similarly to the last section on Beck's model of depression. That is, I will start by describing Salkovskis' early model of OCD and its construction. In a second section, I point out how the more recent version of this model differs from the first, identifying factors that might account for these differences.

2.3.1 Salkovskis' Early Model of OCD

Paul Salkovskis presented the first formulation of his cognitive model of OCD in 1985. This model, in keeping with the fact that CBT slowly became more established at the time, includes both cognitive and behavioral factors that are thought to be relevant for the development and maintenance of obsessive-compulsive symptoms. Important for Salkovskis' first formulation of the cognitive model of OCD was the distinction between intrusions and negative automatic thoughts contained in the model. More generally, he draws on the idea that intrusive thoughts are, first of all, involuntary cognitions that occur in healthy individuals as well, but only mentally ill individuals interpret them as indicators of danger. They thus emerge as *stimuli*, rather than as problematic *reactions* to particular cognitions (on that matter, Salkovskis follows Rachman 1971, 1976). Relatedly, the occurrence of intrusive thoughts is not understood as problematic in itself – following the empirical results of Rachman & de Silva (1978). What is problematic is the negative evaluation of these intrusions that is taken to come in the form of negative automatic thoughts.

According to Salkovskis & Warwick (1985), potential stimuli for intrusive thoughts – for example, specific kinds of situations – become *triggering* stimuli for automatic thoughts if they are not avoided. In agreement with Beck's understanding of automatic thoughts in the context of depressive disorders, he claims that automatic thoughts that result from a particular interpretation of the intrusions led to the problematic mood disturbance. Automatic thoughts, in contrast to intrusions, are consistent with the individual's set of beliefs.[14] This mood disturbance, in turn, is taken to activate a specific schema whose content concerns themes such as loss, threat or blame, thus raising the probability of further problematic automatic thoughts to occur. Furthermore, when this mood disturbance is perceived as implying the individual's responsibility for harm to himself or others, it results in either obsessive actions or in attempts to escape the situation. These obsessive actions have several consequences: For one, they result in the perception of not being punished – which stands in contrast to what the individual takes to happen if she does not act at all. This experience of a reward further increases the probability of perceiving oneself as responsible for taking action once intrusive thoughts occur. In addition, neutralizing behavior brings about an increased acceptance of the initial intrusion. The cognitive mechanism proposed here is that by acting on the intrusion, the individual accepts its content as valid (Salkovskis & Warwick 1985). Furthermore, neutralizing behavior is hypothesized to bring about more triggering stimuli in the future. By reducing the individual's discomfort, the expectancy of further mood disturbances rises.

One of Salkovskis' main motivations in developing his new cognitive-behavioral model was that the treatment of patients with methods that were based on the psychoanalytic models has led to "conspicuously poor outcome" (Salkovskis 1985, p. 572). While the number of psychoanalytic treatments administered to patients were already in decline in the 1980s (compare, e.g. Gifford 2008), due in part to the increasing prominence of cognitive and behavioral methods for the treatment of mental disorders, the author describes cognitive methods as largely "atheoretical" (Salkovskis 1985, p. 572). Behavioral approaches are,

[14] Obviously, this cannot be the whole story, as it seems to involve the strongly implausible assumption that the individual's set of beliefs is consistent in itself. Nonetheless, the distinction between beliefs that are perceived by the individual as alien to himself (or "ego-dystonic") and beliefs that are perceived as close to the self (or "ego-syntonic") still makes sense. It might be refined by framing it in terms of the individual's most highly valued *attitudes* rather than in terms of his or her *beliefs*.

by comparison to cognitive approaches to treating OCD, “more traditional” at the time (Salkovskis & Warwick 1985, p. 243). In a nutshell: For Salkovskis, psychoanalytic models provided a theoretical basis for therapy that was not matched by cognitive or cognitive-behavioral models of the disorder. But psychoanalytic models suffered the immense defect that they could hardly be used to successfully treat the disorder in question. This motivated Salkovskis to develop a new model.

In addition to that, previous models of the disorder did not distinguish properly between the causal processes underlying OCD and those underlying other anxiety disorders. A model proposed by Beck tried to make sense of the symptomatology by claiming that the thought content of an individual with OCD would revolve around thoughts of “doubt or warning”. Salkovskis criticized this since it did not allow for a distinction between individuals with OCD and individuals with other anxiety disorders. He thus set the task for himself to develop a model of the disorder that identifies factors that are *specific* for OCD (Salkovskis 1985, p. 571).

How did Salkovskis proceed in developing this model? He describes the construction process as follows:

> “The formulation outlined above was arrived at as a result of careful consideration of a large number of obsessional patients, and is illustrated below by two examples of quite different patients. Both patients were interviewed about the content of their intrusive thoughts, and then asked to try and focus on any thoughts subsequent to the intrusions as they occurred, particularly if these were associated with discomfort.” (Salkovskis 1985, p. 576)

Note how patients were asked to provide reports based on introspection that were then used as evidence in the construction of the model. Thus, clinical evidence from his practical work was a primary source of evidence for this model. According to Salkovskis, the two cases mentioned in his first paper on OCD and other instances he discusses there were taken from clinical work that he carried out during this time (Paul Salkovskis, personal communication).

As indicated above, one important experimental finding underlying this model was the result of Rachman & de Silva (1978) that most healthy individuals experience intrusive thoughts with similar content to the intrusive cognitions of individuals suffering from OCD.

Theoretically, his model is based upon Rachman's *behavioral* model of the disorder, according to which "[...] obsessional thoughts are noxious conditioned stimuli which have failed to habituate, and which are maintained by the mechanisms involved in two-process learning" (Salkovskis 1985, p. 573). That is, both the theoretical and the experimental work of Rachman was an important point of reference for Salkovskis in setting up this model.

The term "two-process learning" refers to the idea that "two independent hypothetical constructs, habituation and sensitization, interact to produce the net response to repeated stimulation." (Groves & Thompson 1970, p. 421). That is, when an organism is repeatedly confronted with the same stimulus, the intensity of its reactions over time can either decrease – that is, it *habituates* –, or it can increase over time – that is, it *sensitizes*. Whether an organism habituates or sensitizes to a particular stimulus is determined by both how frequently the stimulus occurs and by how easily excitable the organism is at that particular point in time. According to the author, obsessional ruminations usually lead to sensitization because they have a special significance for the individual. In other words, the individual is particularly excitable for this particular kind of noxious stimulus. The difference between healthy individuals and individuals with OCD then amounts to the fact that someone with the disorder tends to react with higher levels of arousal to such noxious stimuli or that their intrusive thoughts are more intense (Rachman 1971, p. 231-232).

Rachman's model is silent on the question of *why* these individuals are particularly excitable to these specific types of stimuli. Salkovskis' model of OCD from 1985 can be understood as an attempt to answer the question of *why exactly* some individuals react more strongly to intrusive thoughts than others – and why they fail to habituate to them. That is, they react more strongly because of particular dysfunctional beliefs, and they fail to habituate to intrusive thoughts, because they avoid the stimulus situation through safety behaviors or neutralizing actions. The important factors in question are, just like in Beck's model of depression, specific kinds of schemata that are activated in these individuals.

Let me shortly come back to Beck. An important influence on Salkovskis was Beck's cognitive model (Salkovskis & Warwick 1985). As was already discussed above, Beck's model assumes that one of the main causal factors in the production of depression are cognitive schemas, consisting of specific patterns of thought. These cognitive schemas give rise to particular kinds of automatic thought about internal or external events. In parallel to this, Salkovskis' model assumes that negative automatic thoughts play an important role in OCD, but, in contrast to prior theorizing, not because they are *identical* to intrusions, but because negative automatic thoughts are problematic *interpretations* of intrusive thoughts.

Let me offer some concluding remarks here. We have seen in this section that Salkovskis drew on several sources in setting up his model of OCD. These were, on the one hand, theoretical sources like Beck's cognitive model or Rachman's theory of OCD. On the other hand, he used experimental findings such as the discovery of Rachman & de Silva (1978) that healthy individuals *also* experience intrusive thoughts. But again, as Salkovskis states himself, he also made use of information gathered within the psychotherapeutic context, that is, when treating patients with OCD. Thus, besides drawing heavily on a model that was itself constructed on the basis of evidence from the context of application, the context of application is directly relevant for his model as well.

2.3.2 The More Recent Model of OCD

When comparing the first model of OCD (Salkovskis 1985) with its recent version, the first thing to note is that, when considering their respective graphical representations (see figure 2.1 and figure 1.1), the second model appears to be less complex than the first one: It contains fewer causal factors – nine in comparison to thirteen or fourteen[15] –, and the processes depicted are less complex as well: While the first explanatory model mentions causal as well as potentially disrupting factors that have the effect of masking the symptoms in someone who suffers from OCD, the second model only mentions causal relationships that hold in case someone *actually* experiences all of the disorder's symptoms. In fact, it arguably idealizes the disorder somewhat, since not every patient with OCD carries out compulsive actions. This reduction in complexity is noteworthy in part because Beck's model of MDD seems to

[15] This depends on whether one counts only those factors that are connected via arrows – which, according to my understanding, represent causal factors –, or *also* those factors that may disrupt certain causal processes.

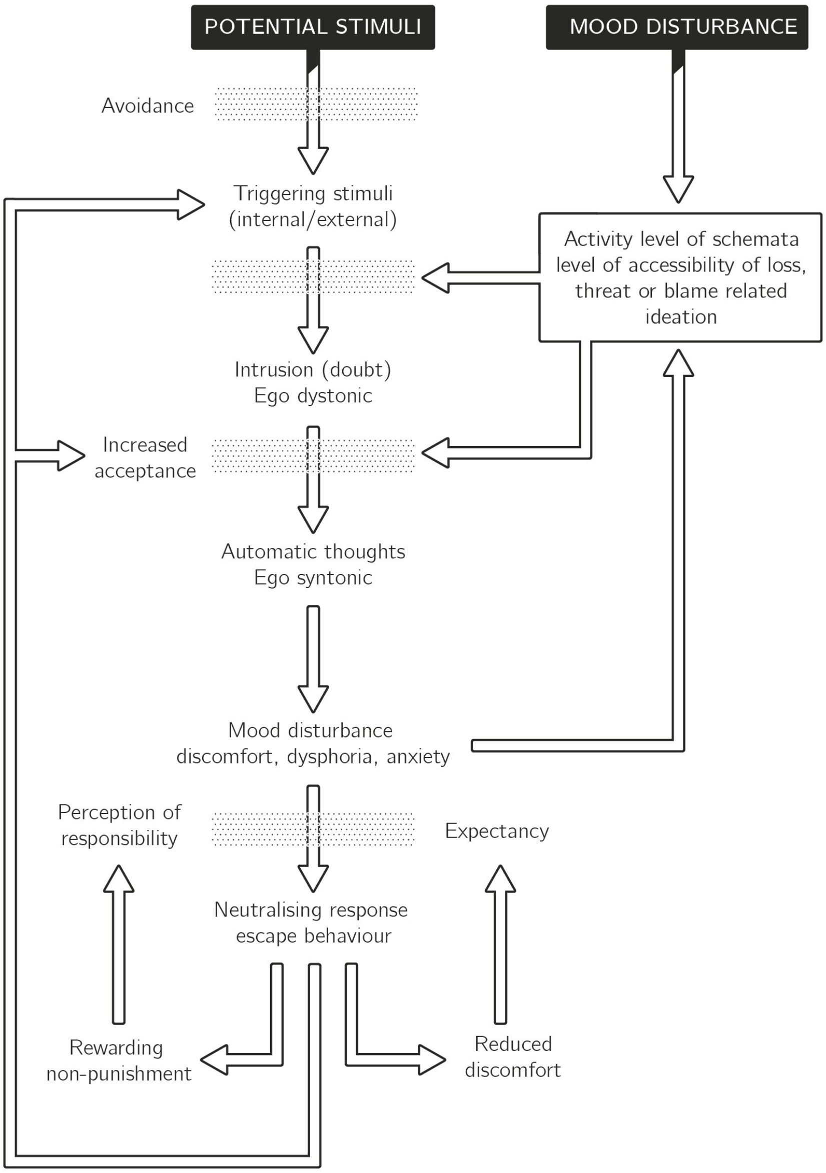
POTENTIAL STIMULI
MOOD DISTURBANCE
Avoidance
Triggering stimuli (internal/external)
Activity level of schemata level of accessibility of loss, threat or blame related ideation
Intrusion (doubt) Ego dystonic
Increased acceptance
Automatic thoughts Ego syntonic
Mood disturbance discomfort, dysphoria, anxiety
Perception of responsibility
Expectancy
Neutralising response escape behaviour
Rewarding non-punishment
Reduced discomfort

Figure 2.1: Cognitive model of the origins and maintenance of Obsessive-Compulsive Disorder, put forward by Salkovskis (1985).

have evolved in the opposite direction by adding further causal factors to the first model. It is also interesting because the second model of OCD – again, other than Beck's more recent model of MDD – does not mention causal factors from other disciplines such as physiology or genetics, but only seems to provide the reader with a simpler version of the first model. Nonetheless, it is this second version of the model that has made its way into textbooks of clinical psychology (Wittchen & Hoyer 2011, p. 1009-1010). Thus, it is fruitful to ask: Why exactly did the author change his model in this way?

Thus, in a nutshell, the differences between these two versions of Salkovskis' model are, when compared to the development of the model of depression, more substantial.

The two models appear to account for different phenomena: The 1985 model seems to focus on how the disorder's symptoms hang together causally. It shows how certain cognitions and behaviors give rise to further symptoms of the disorder. In other words, it shows why, in a specific situation, someone would display the syndrome that is typical for OCD. Consequently, it says little about the disorder's *etiology.* By contrast, the model of Salkovskis et al. (1998) intends to account for *both* the disorder's etiology and, more importantly, for its maintenance. This is the case *even though* this model does not say much about the disorder's etiology. For this reason, only the more recent model of OCD mentions early experiences that lead to particular kinds of dysfunctional beliefs that may become activated at a later point in time. Interestingly, the earlier model does refer to cognitive schemata that are activated by disturbed mood. We may understand this as parallel to the understanding of the later model, according to which particular kinds of assumptions and general beliefs of the individual that have been dormant for some time become activated by specific events. Thus, the later model includes both etiological and maintenance factors of this disorder.

Concerning the part of Salkovskis' recent model that deals with OCD's *maintenance*, the issue is more complex, when compared to those changes that I observed for Beck's model of depression. Although the model from 1998 clearly takes the 1985 model as its basis, the causal factors mentioned there are neither a super- nor a subset of those mentioned in the earlier model. Even among those factors included in both models, the causal relations that are mentioned differ to an almost puzzling degree. But let me start with the good news: Most of the causal factors that are mentioned in the 1998 model of OCD have already been mentioned in the classical model, even though their labels have sometimes changed. Even the causal factor *attention & reasoning biases* that does not explicitly occur in the earlier model, nonetheless is there implicitly: Those biases occur when, as is represented in the earlier model, mood disturbances (brought about by automatic thoughts) increase the activity of schemata that have content related to loss, threat or blame.

The other causal factors mentioned in the earlier model can quite easily be mapped on those factors mentioned in the later one. Since the terminology differs somewhat, it is helpful to shortly describe which factors appear to be analogous to one another: *Intrusions* of course map onto the *intrusive thoughts, images, urges and doubts* of the later model. It gets more interesting from here on: I take "automatic thoughts" to have roughly the same meaning to the later "misinterpretations of significance". Quite clearly, the "mood disturbance[s]" of the earlier model map onto the "mood changes" of the later model, whereas "neutralising response[s]" are clearly analogous to "neutralizing action[s]". Finally, the "counterproductive (safety) behaviors" of the later model may be understood as analogous to both earlier *avoidance* of potential stimuli and to *escape behaviors* in the situation at hand.

I take one of these terminological changes to be relevant for my argument: The only substantive change is the one from "automatic thoughts" to "misinterpretations of significance". In changing this label, the more recent model of OCD explicitly mentions *what goes wrong* in the reasoning process of someone with OCD. That is, a relatively neutral term is exchanged with one that implies a judgement about the correctness of this output. This focus can be understood as being due to an increased interest in using this model in the context of application: it directly shows both mental health professional and patient where to intervene.

Concerning the causal *processes* that are depicted in these two models, there are several things to say: According to the first model, disturbances in the individual's mood bring about neutralizing behavior and escape behavior and the activation of problematic cognitive schemata. Disturbances in the individual's mood indirectly increase – in the old model – the probability for further intrusive thoughts. All of these effects are, in the more recent model, ascribed to *misinterpretations of significance* and the – allegedly skewed – *perception of responsibility.*

I would like to suggest that this change in the proposed causal relations might partially be due to the fact that the more recent model of OCD is explicitly intended for usage within psychotherapy, that is, as a means to explain this mental disorder to the patient. It seems that, by putting misinterpretations of significance at the center of the disorder, the audience should get the impression that *cognitive* factors are at the heart of OCD, that is, certain false beliefs. This seems to fit one of the goals of therapy: to convey the idea that, while certain relatively stable features of the patient may be causally relevant for her troubles, there are factors that help maintain the disorder that she can intervene upon herself or in therapy. The usage of talking therapy, and more specifically, CBT for treating OCD is implicitly justified by this model.

A further important difference is that the earlier model of OCD relies more strongly on Beck's classical model of depression than the more recent model, by referring to automatic thoughts and cognitive schemas as relevant causal factors in the explanation of the disorder. By comparison, the later model gets rid of talk about schemas and instead refers to assumptions and general beliefs as well as to particular attention and reasoning biases. This change in vocabulary might be due to the wish of formulating a simple explanatory model that can easily be used in psychotherapy. This makes sense, since Salkovskis (1999) underlines the need to present an explanation of the mental disorder to the patient that takes his model as its basis.

2.4 Conclusions

In the preceding two sections, I analyzed how two explanatory models of mental disorders were constructed and further developed over time. In comparing an earlier to a later version of each model, I focused especially on changes in these conceptualizations that might be due to pressures from the context of application. In the following, I will revisit the most important findings from this investigation.

Salkovskis' model of OCD was simplified over time – in the sense of a reduction of causal factors included in the model –, while Beck's model of depression became more complex. I have suggested that both of these changes might also be due to the influence of particular pragmatic goals in the process of model construction: There, the model's usability in explaining mental disorders to patients plays a big role. A model is only usable if it is sufficiently simple. As Salkovskis 1985 version of his model is extremely complicated, while Beck's 1967 version of his model is more intelligible, it makes sense that they would develop differently.

A further point I made above is that different kinds of evidence are important in model construction and in the further development of these models.

We have seen above that clinical evidence from the context of application was of particular importance for both the initial formulation of Beck's depression model and for the first formulation of Salkovskis' model of OCD. Both authors explicitly state that their models were, to a large extent, based on individual clinical cases of patients that they treated in psychotherapy. In the case of Beck, we may even point to specific statements of patients within psychoanalysis that led to the development of particular concepts that now are at the core of the cognitive model of depression. Thus, the importance of the psychotherapeutic context for these construction processes seems hard to deny.

I think that the respective developments of these two models of OCD and MDD can be understood as indicating the importance of the context of application. This has different consequences for each model, to be sure: Firstly, Beck's model develops from a primarily *conceptual* model of depression that mainly aims to represent the underlying structure of the disorder to a model *explicitly* including factors that are considered important in the development, and, more importantly, maintenance of the disorder. This shows the importance of the model's usability for developing of psychotherapeutic treatments, since maintaining factors are one central factor that – in addition to dysfunctional beliefs – psychotherapeutic treatments intend to change in order to help the individual. Beck's classical model of depression, which does not mention any maintenance factors at all (Beck 1967). Secondly, Salkovskis' model also shows changes over time that might very well be understood as guided by practical considerations: Over the course of time, his model became less complex. Furthermore, it shows changes in terminology that arguably have to do with the pressure of usability for therapeutic intervention. In combination with his insistence that good psychotherapeutic treatment needs to start with an explanation of the disorder in question (Salkovskis 1999, p. S33), we may understand this development as due to a stronger focus on usability of this model in the actual psychotherapeutic context.

Since the development of effective psychotherapy indeed is a chief goal of clinical psychology, this discipline is built upon the implicit assumption that effective treatment of mental disorders through talking therapy is possible. This has a number of interlinked effects. That is, if the effectiveness of psychotherapy is assumed at the outset, and psychotherapy is partially characterized by its contrast to somatic medicine, the clinical psychologist's models will necessarily show some very particular features. These are due to the fact that there is only a limited number of *objects* that psychotherapy with its focus on repeated interpersonal interaction can reasonably intervene upon *at all.* Furthermore, there is an even smaller number of objects that psychotherapy can reasonably intervene upon *better than* medical treatments. I would like to make the relatively bold claim that considering psychotherapy effective at the outset places several constraints on the explanatory models that are developed. For example, it makes it much more plausible to include mental states that are believed to be introspectively accessible to patients.

In the development of these models, other kinds of evidence come in, particularly, evidence from controlled studies (compare, e.g., Clark & Beck 1999). A substantive part of the evidence that is used to give credence to these models is also evidence obtained by outcome studies that investigate the effectiveness of therapeutic interventions that are based on cognitive theory (compare, e.g., Clark & Beck 1999, p. 400). The underlying reasoning seems to be that, if the intervention works, then the underlying model must be correct. But this inference does not go through, since there may be third factors responsible for the superiority of one kind of treatment over another. This has led several researchers to conclude that, while psychologists seem to know *that* CBT works, no one knows exactly *how.* That is, the mechanisms of change are yet unclear (compare, e.g., Flynn & Warren 2014). Nonetheless, even if we may be skeptical about the quality of *some* of the evidence that is mentioned in favor of these models, the sheer amount of it (compare, e.g., Clark & Beck 1999, Beck & Bredemeier 2016) might ease those worries that one might have about the context of application influencing these models in a way that undermines their correctness.

There are several questions that are raised by the analysis and discussion in this chapter that I will tackle in the remainder of this dissertation.

Firstly, as I have pointed out before, these explanatory models are – in accordance with Salkovskis' recommendation – often used by psychotherapists in order to explain a patient's idiosyncratic symptomatology to the patient. This makes it plausible to assert that *utility* and *understandability* are guiding values in the development of these models, and those changes that we have seen in both Salkovskis' model of OCD and Beck's model of depression over time suggests that this might actually be true. But what does it mean in practice to explain mental disorders to patients? In the next chapter, I will develop on how exactly – and with which practical goals in mind – these models are used in psychotherapy to explain mental disorders to patients.

Secondly, I have argued that one main source of evidence for both models – that is, so-called “clinical evidence” – derives directly from the therapeutic situation, or more precisely, from a psychotherapeutic situation that involves a conversation about the patient’s thoughts, beliefs and emotions as a defining feature. This extends to almost all kinds of psychotherapeutic methods. When regarded in this context, the use of folk-psychological concepts in explanatory models is not at all surprising, provided that folk-psychology is what the untrained layperson uses in order to talk about her inner states. Psychotherapy, in turn, almost always starts with the individual’s subjective or life-world perspective on her symptoms.[16]

In the next chapter, I will extend this line of reasoning further, arguing that those noteworthy features of explanatory models that I have identified in the first chapter map very well on particular aims that therapists have when explaining mental disorders. This suggests that explanatory models might be influenced by considerations arising from the therapeutic context.

Thirdly, and finally, if evidence from the context of application is as important for the development of these explanatory models as I have tried to argue, what exactly does it mean for these models? Are there particular features of the context of application that influenced the form and content of these models? In the next chapter, I will argue that this is indeed the case.

[16] This is how many of my interviewees described their process of construction an explanation of their patient’s disorder in the psychotherapeutic context.

Chapter 3

Using Explanatory Models in Psychotherapy

3.1 Introduction

In this chapter, I will be concerned with how mental health professionals explain their patients' disorders in therapy. More precisely, this chapter is about what Cooper (2007) calls *individual case histories.* I will ask how such individual accounts of the development and maintenance of someone's mental disorder are constructed on the basis of the explanatory models in question. In answering, I will describe how three specific practical aims influence the explanation that is formulated on the basis of both the model and the clinical information. I will discuss whether being first constructed and used on the basis of evidence from psychotherapy also influenced the form and content of these models. My thesis is that being based in psychotherapeutic practice actually influenced these models, and that this is due to three practical aims that these explanatory practices serve. My analysis is based on the results of six qualitative interviews with cognitive-behavioral therapists.[1] The interested reader may consult the subsequent section about methodology for further information.

As I have pointed out in the first chapter, the explanatory models of interest here are mainly used in two contexts: firstly, in the *context of research*, where they are intended to further the expert's understanding of the disorder, and, secondly, in the *psychotherapeutic context*, where explaining this disorder to the patient is held to be crucial for successful psychotherapy. For example,

[1] This is also why, in this chapter, I will very often claim that particular facts seem to hold "at least" for CBT. I add this proviso not because I want to discredit other kinds of psychotherapeutic interventions, but because, for reasons of space, I cannot extend these claims to other therapeutic orientations here.

Paul Salkovskis (1999, p. S33) writes: "[...] an individualised version of the model [...] is an *essential part* of the process of therapy [...]" (my italics). I am mainly interested in the context of application in this chapter, since it seems – both from my experience as a psychology student and from what conversations with psychotherapists and researchers in clinical psychology have taught me – to have a much more intricate influence on what is presented as an explanation and which standards of good explanations are operative in the respective context. Furthermore, this context has so far been neglected in the philosophical literature

I think that there are two questions to ask when considering Salkovskis' statement: Firstly, why exactly would explaining her mental disorder to the patient be *essential* for psychotherapy? That is, what is the *function* of explaining these disorders, and why is this function considered important enough to make these explanations count as a necessary part of therapy? And secondly, what does it mean to *individualize* such a model? Naïvely, one might think that one has individualized such a model just in case one has mapped each *type* causal factor to a token causal factor one takes to be operative in the patient. An "individualized version" of the model would then consist in representations of these token causes plus representations of the token causal relation(s) that supposedly hold between them, presented in a narrative form.

Consider a simplified model, according to which a patient's panic attack is caused by their experience of a specific physiological symptom that is interpreted in a problematic way, that is, as potentially dangerous for the patient.[2] On our naïve understanding, an individualized version of this model looks roughly like this: The experience of chest pain, in combination with the belief "chest pain means that I am having a heart attack" causes the patient to experience severe anxiety, and eventually, panic. Let us call this naïve view of how individualization works "simple individualization".

While this does, to my knowledge as a psychology student, fit how these models are individualized when students are taught about clinical psychology, mental disorders and their treatments, it is not quite how they are usually individualized within psychotherapy, when disorders are explained to patients.[3] Espe-

[2] This is an extremely cut down description of the model of panic first put forward by Clark (1986). This model is now standard within clinical psychology.

[3] Which might be quite different from an explanation of the patient's disorder that a therapist formulates for herself. These may, in many cases, be closer to simple individualizations, but they

cially more complicated explanatory models like those that I have discussed here seem to require a different mode of tailoring them to the patient, his cognitive abilities, his needs and his narrative about his life.

In informal conversations with psychotherapists in training[4], several of them pointed out that these explanations needed both to make sense *in the patient's individual narrative* of his life, and they needed to *facilitate interventions*. In many cases, the patient's individual narrative stands in tension to the kind of explanation one gets when individualizing the model in the sense presented above: In my interviews, one participant referred to *Narcissistic Personality Disorder* (*NPD*) (for the diagnostic criteria, see appendix A.7) as an extreme example of this. Here, patients often believe that the causes of their problems are located in the external world. According to the therapist, by contrast, the real cause very often is that the patient himself had particular, exaggerated ideas about of how much attention from other people he was entitled to. But at least in the beginning of, and sometimes throughout the therapeutic process, the latter account of the patient's problems could not be given without risking that he would discontinue therapy.

More generally, formulating an explanation in such a way that it uses vocabulary similar to the patient's in describing his experiences and not contradicting his narrative too extremely was understood by several of my interviewees as a way of taking the patient seriously as an epistemic agent.[5] Among other things, taking the patient's reports seriously is taken to be the basis for a good therapeutic relationship, at least in CBT, where both a good therapeutic alliance and so-called "collaborative empiricism" are central tenets of the method. Collaborative empiricism refers to the questioning of the patient's dysfunctional beliefs by the therapist and the patient *together*, taking the patient's statements as primary evidence for the rationality or irrationality of

will, I take it, take into account and represent much more information about the context, for example.

[4] Importantly, these are not identical to the qualitative interviews that I am referring to here – they are, so to speak, an informal pilot study that I conducted to generate some preliminary hypotheses about psychotherapeutic practice.

[5] I will come back to this issue when discussing further observations from my qualitative interviews, which, among other things, indicated that there might be an interesting difference – if the relevant interviewees of mine are right, that is – between parts of medical practice and (at least) cognitive-behavioral therapeutic practice when it comes to the issue of epistemic injustice. (Please be aware that the reason I am only speaking of CBT here is the fact that I have limited my interviews to individuals who practice this form of therapy.)

holding these beliefs (e.g. Tee & Kazantzis 2011).[6]

I thus decided to investigate how psychotherapists utilize these two perspectives, that is, the seemingly more objective "causal" and the supposedly more subjective "life-world" perspective (compare Bolton & Hill 1996), when explaining mental disorders to their patients.[7] For now, I will neglect the normative question whether it is legitimate to do so. My guiding idea is that both the incorrectness of the naïve conception and the sometimes simultaneous use of these two perspectives by psychotherapists results from the operation of particular practical aims of explanatory practices in psychotherapy.

The data from my qualitative interviews will be used to reveal exactly how the intuitive understanding fails to represent actual practice. I will show that those noteworthy features of explanatory models that I have already discussed in the preceding chapters enable mental health practitioners to satisfy their practical goals when engaging in explanatory practices. Given the observation that models of mental disorders are constructed on the basis of evidence gathered in the therapeutic context, which I argued for in the preceding chapter, a plausible explanation for this fit between model features and aims of explanatory practices is that the latter *do*, in fact, influence the models' content and form.

Let me remind you of the features that I allude to here. Firstly, these models are intended as *causal* models. Secondly, they employ the *folk-psychological vocabulary* of beliefs and desires, thus allowing for the patient's subjective perspective to be taken into account. As I see it, they thus employ two perspectives at the same time: the *causal third-person perspective*, and the *first-person, life-world perspective*, where reasons for particular actions and experiences are presented. Thirdly, functional vocabulary is at the heart of these models: Talk of the "dysfunctionality" of particular beliefs is at the core of CBT. Finally,

[6] This is also clear when considering that core techniques of CBT are referred to as "Socratic questioning": The idea here is that the patient is aided through particular kinds of questions to find the solutions for his problems himself (Braun et al. 2015).

[7] Importantly, the observation that these two perspectives are operative in psychiatric and psychological explanation of mental disorder is not new but has been made by Bolton & Hill (1996) first. Nonetheless, their focus, while fascinating, is different from my own in that they take the explanations of psychiatry and clinical psychology at face value, trying to develop and account of intentionality that allows for intentional explanations of human behavior to be a proper part of scientific efforts. By contrast, I ask why clinical psychologists and psychotherapists would, not only for theoretical reasons, but also for pragmatic reasons that arise in the context of application – that is, in particular, in psychotherapy –, need to have *both* intentional and causal vocabulary in their explanatory models of mental disorders.

these models *normalize* the mental disorders in question in the sense of Bolton (2008). That is, those mental processes underlying the symptoms in question emerge as (on a dimension with) statistically relatively normal, sometimes as *functional*, human processes. This manner of understanding mental disorders stands in contrast to a prominent view in the philosophy of psychiatry, according to which mental disorders are harmful dysfunctions of evolved mechanisms (Wakefield 1992).

Thus, the models developed in clinical psychology and used in psychotherapy do not adhere to Wakefield's account of what mental disorders are. This is interesting even if it does not count as evidence against his view of what mental disorders *really are*. That is, it might be the case that mental disorders actually *are* harmful dysfunctions, but psychotherapists and researchers in clinical psychology just conceptualize them as the result of normal psychological processes, perhaps doing so for practical reasons.

This chapter is structured as follows: In a first section, I will describe the methodology of my qualitative interviews. In a second section, I will present those aims that I take to be operative in explanatory practices in psychotherapy. I developed my account of these aims on the basis of informal conversations with therapists in training and the literature that influenced Aaron Beck in setting up his first model of depression. In a third section, I will present what I think are good indicators for the correctness of these hypotheses both from my qualitative interviews and from the literature on CBT. Section four deals with the question whether it makes sense to think that these aims of explanatory practices exert some influence on explanatory models of mental disorders. Having put forward some reasons to think so, section four deals with a potential problem that may arise because (1) patients in therapy *react* to being given particular kinds of explanations for their mental disorders and (2) these reactions are systematically evoked, influenced and exploited by mental health professionals for therapeutic gains. The phenomenon that might occur here is similar to Hacking's *looping effects.* Finally, a last section draws these strands together, offering conclusions and noting further issues of interest.

3.2 Methodology

My analysis is based on the results of six qualitative interviews with psychotherapists.

These interviews were semi-structured expert interviews that I conducted with six experienced cognitive-behavioral therapists in Germany. To make sure that the interviews would cover the same topics in a similar way, I prepared an interview guideline, in keeping with those principles mentioned in Galletta (2013). Roughly two thirds of the questions were identical among all interviews, and one third differed between them. This latter third of questions were either adapted to the particular interviewee on the basis of my knowledge of their specific line of work, their publications, or they were questions that came up after reflecting on the results of prior interviews. In this investigation, I use my interview data mainly for two purposes: In this chapter, to inform my analysis of how models of mental disorders are applied by psychotherapists in practice. In particular, I use it to inform my ideas about those aims that therapists have when explaining mental disorders in therapy. In the subsequent two chapters, I use it in investigating the concepts of "dysfunctionality" and "irrationality". There, I intend to provide an analysis of those terms that fits the actual practice while simultaneously being free from contradictions, simple and fairly general.

For the latter purpose, it was sometimes necessary to *reconstruct* what my interviewees told me. That is, while I set up my account such that it is consistent with *most* of the empirical data, I needed to sometimes assign more weight on specific statements than to others – on occasion, this was for the simple reason that the practitioners' statements would contradict one another. In other words, I am providing a *rational reconstruction* (Lakatos 1970) of what these therapists told me.

With my interviews, I aimed to generate information about the cognitive processes and practices that influence the construction and psychotherapeutic application of the two explanatory models that I discuss here.

The participants of my interviews were recruited via email. I aimed to mainly interview therapists who appeared to have a particularly strong interest in the theoretical foundations and social implications of their work and who already practiced psychotherapy for several years. I identified such individuals mainly by attending conferences on clinical psychology and psychotherapy in Germany. Since only a fraction of those therapists I contacted were willing to participate in these interviews at all – from the 21 individuals I contacted via email, only six told me they were willing to participate (yielding a response rate of roughly 30%) –, it was necessary to ask some therapists that I had been acquainted with before.

When those individuals I contacted confirmed that they were interested in participating in an interview, they were sent an information sheet via email, providing them with further information on the study (see appendix B.1). At the beginning of each interview, they were handed a printed version of this sheet and potential remaining questions were discussed. They were informed verbally about the aims and structure of the interview. Finally, each participant gave written consent for participating in the study (see appendix B.2). Importantly, participation was completely voluntary and participants had the opportunity to withdraw at any point in time without negative consequences of any kind.

The interviews were recorded on a Philips VoiceTracer recorder and transcribed afterwards. Both the audio files and the transcripts were stored on a password-protected computer in pseudonymised[8] form, that is, without mentioning the name, address or any other kind of information that could be used to identify the participant, but in such a way that re-identification of the participant in question was possible by using an additional list. I agreed to only give the data to third parties, including publication, in anonymized form. For this reason, quotations from my interviews will be given without names.

Overall, I interviewed three male and three female therapists, resulting in a balanced – but not representative[9] – gender ratio. Three of these interviews covered the exemplary model of depression I am interested in here, three covered the model of OCD. The length of the interviews varied between 48 minutes and three hours and 41 minutes, with a mean of one hour and 23 minutes. That is, the longest interview was clearly an outlier. When analyzing the data, I identified common factors, for example, efforts to *normalize* the patient's dis-

[8] "Pseudonymisation", in the *GDPR* (the *EU General Data Protection Regulation*), "is defined as the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information" (Voigt & von dem Bussche 2017, p. 15).

[9] This can be inferred from the fact that the German Bundespsychotherapeutenkammer (2019) has, among its members, roughly 70% women and 30% men.

order, in the therapist's answers. Having done that, I searched for exemplary statements that represented these similarities well enough. Problematically, the therapist's answers differed greatly both in the stance they took towards their therapeutic practice, with some therapists describing their everyday work in lots of concrete detail, and others pointing out their reasons for proceeding like that. They also differed greatly in the level of detail they would go into when doing so.

Since the interviews were conducted in German, all the quotes that can be found in the course of this and other chapters are my translations.

For these reasons as well as the fact that the number of interviews I conducted is quite limited, I consider it best to understand the data generated there as providing *indicators* for the correctness of the hypotheses that I presented before. They do not, in and of themselves, constitute conclusive evidence.

3.3 Aims of Explanatory Practices

On the basis of informal conversations with aspiring psychotherapists, I developed the hypothesis that explanatory practices in therapy serve the function of ensuring that patients are more likely to continue and finish talking therapy by being both motivated and properly prepared for treatment. Furthermore, they serve a very first therapeutic function, that is, to alleviate harm. This results in three practical goals of these practices. I take it that a patient is more likely to continue and finish structured talking therapy if these goals are satisfied.[10] Generally, the aim of these explanatory practices seems to be that patients should not feel at fault for developing her disorder, while she should nonetheless take over responsibility for counteracting her symptoms. This leads to three sub-goals that I will name now and describe in more detail below:

1. attribution of limited responsibility
2. attribution of agency
3. presenting possibilities for intervention to the patient

[10] Importantly, while explanations may be helpful for therapeutic progress, they neither constitute *sufficient* nor necessary conditions for the intervention's effectiveness.

Firstly, the patient perceives herself as *not at fault* for developing and still suffering from her disorder. “Not being at fault” can mean several things here: (1) The agent *could not choose* whether to adopt or not to adopt certain features that were causally relevant for developing the disorder in question, and thus, she is not responsible for them, or (2) although the agent was able to control those factors that later turned out to be harmful, she *did not know* about their harmfulness, and thus, cannot be blamed for adopting them (Rudy-Hiller 2018). Since neither of these two factors are easily changed at the current point in time, the individual cannot be blamed for still suffering from the disorder.

Factors of type one are usually *predisposing* aspects that are involved in the development of the disorder – think of the patient’s genetics, personality structure or other deep-seated, dispositional factors. Factors of type two are often beliefs that *made sense* for the individual to adopt. This means that considerations of a belief’s *rationality* and *justification* are relevant for this, too. Finally, to make the patient understand that it really *is* hard to act otherwise in their situation – instead of things just subjectively appearing to be hard[11] – at least some of them are described as substantively different from the features of healthy individuals, and furthermore, as relatively *deeply embedded* in the individual.

What exactly do I mean by a feature’s being more or less deeply embedded in an individual, though? In my understanding, a feature is more deeply embedded in an individual the harder it is to *control* or *change* by the person. For the purposes of illustration, think of the difference between two kinds of actions or behaviors: for one, take the – ceteris paribus – strongly automatized action of using the clutch when driving a car. For the other, consider the action of solving a particular mathematical equation. For most people, using the clutch when driving is more deeply embedded into their action repertoire, such that it becomes hard not to engage in it, when they are in a situation that resembles the one in which they usually use the clutch in relevant ways. Just consider a person who is asked to drive an automatic car, when not being used to it. Solving a mathematical equation is not as deeply embedded in most people’s action repertoire, making it much more easy to stop doing so.

[11] This is extremely tricky to frame in the right terminology, since there is a tension between representing the patient as able to take action against her disorder (which is another aim of explanatory practices, as I point out below), and being substantially different from other individuals.

Some features that are involved in the development and maintenance of mental disorders are deeply embedded into the individuals in this sense – think of certain harmful beliefs which these individuals hold over a long period of time. Other features are even more deeply embedded – think of personality features or genetic factors that they arguably have even less control over, if at all. One of the key tasks of psychotherapy is to identify those features of the individual that are not too deeply embedded to be subject to change within the therapeutic process. Even those will be relatively deeply embedded – if they were not, the individual would probably not be in need of psychotherapy in the first place.[12]

If things go as planned, an explanation that satisfies these criteria results in the patient representing herself as not fully blameworthy for having *developed* and for still suffering from the disorder – which may be an understanding of herself that is substantially different from her previous one.

Secondly, she perceives herself as possessing *agency*. This means that she does not conceptualize herself as broken, malfunctioning, or crazy. According to several of my therapists, this is a relatively common understanding or fear that patients have when beginning therapy. Instead, the agent understands herself as still able to act. This has two components, namely, (1) being in a position to understand her experiences and behavior from the intentional stance (compare, e.g., Dennett 1971), and (2) being able to understand herself as relatively normal – that is, understanding her mental processes as statistically normal or as relatively *functional* (both components are called "normalization" in Bolton 2008).

Thirdly, she perceives the therapeutic approach as a useful tool to *intervene* on the maintaining factors of her disorder. In my understanding, without the beliefs of still having a considerable amount of agency *despite* one's mental illness and of CBT being a useful tool to intervene on the maintaining factors of the disorder, CBT cannot get started at all. This is because this therapeutic approach relies on understanding and changing the patient's experience and

[12] One might wonder about factors that are not deeply embedded in the sense of being *in principle* hard to control or change. Think of an individual's *inactivity* in depression. Being inactive is extremely hard to change for someone with depression, plausibly because the individual's emotional state motivates her to remain inactive. I will classify this kind of behavior under the heading of "deeply embedded", because it is, given the patient's context, extremely hard to change. Furthermore, the causes of this state are not under the agent's control: if Beck & Bredemeier (2016) are right, the depression program is hard-wired into human brains.

behavior from the intentional stance, and in the patient changing *his own* beliefs and behavior *deliberately.* It seems hard to intervene on one's own mental processes and behaviors if one does not believe in the effectiveness of doing so. This is particularly the case for mental disorders, which usually are present in the individual over an extended period of time *precisely because* acting otherwise – not engaging in neutralizing behavior despite suffering from OCD, being socially engaged despite suffering from depression – has become extremely difficult for the individual, just as I pointed out above.[13]

Interestingly, these three aims fit quite well with how Samuel Kraines describes the goals of explanations in psychotherapy in his book *Mental Depressions and Their Treatment* from 1957. Kraines was a psychiatrist and researcher, more precisely, Clinical Assistant Professor of Psychiatry at the University of Illinois. One of the readers of Kraines' book was Aaron Beck, who further developed his description of a *thought disorder* in depression (Beck 1967, p. 228). It thus is instructive to take a closer look at Kraines' statements about the aims of explanation in psychotherapy.

Kraines writes that "the physician must also "explain" to the patient the nature of his illness and outline for him the therapeutic procedure." (Kraines 1957, p. 406-407). Beck explicitly refers to this: "Kraines and Campbell both stress 'formulation to the patient', i.e., explaining to him the nature of his illness, emphasizing that he will be cured, [...]" (Beck 1967, p. 313). But how exactly does the author set up this formulation to the patient? Kraines claims the following:

> "Such explanation, which serves both to reassure and to elicit cooperation, must be adapted [...] Patients do not understand 'what has happened'; they fear some dread disease, insanity, harming someone; they feel hopeless and helpless. Many patients are greatly relieved to learn that their illness has a physical basis, that it is a common illness and

[13] It is a fascinating question where exactly this difficulty lies. As one of my interviewees pointed out, on a particular perspective, acting otherwise is, if one takes a certain perspective, actually *easy.* He said this: "I never had anyone sitting here who said: 'I cannot drive a car, I'm so scared, and now you have to teach me how to learn driving a car!', or something like that. And this is the case for most behaviors, to make clear to oneself: Wait, this thing that is the new target behavior, or also, if it is about stopping a behavior. Not smoking. This is extremely easy. So, it is easier than smoking. If I smoke, I have to go buy cigarettes, I have to own a lighter, so, do you see? I have to do something. Not smoking? At first sight, regarded this way... [...] We have much more potential control over the behavior than over our inner life. And the inner life makes it hard. The craving. The fear. Whatever. The thoughts. The emotions. This will be the challenge. Not the behavior."

> they are not the only ones so disturbed [...] Steadily accompanying all these explanations, designed to help the patient understand, are:
> 6. The reassurance of recovery
> 7. The reiteration of the self-limiting nature of the illness, and
> 8. The stimulation of hope" (Kraines 1957, p. 407)

There are many things to unpack here: For one, this statement shows the importance of normalization that I have stressed above: The route the author takes is to normalize via pointing out how *common* depression is. If depression is such a frequently occurring illness, then someone who suffers from it can plausibly not be too far from statistical normality in his mental processes. Kraines also hints at something he will claim later on, namely that instilling *hope* in the patient is a central aim of explaining depressive disorders. I think that this indicates the importance of the second aim of explanatory practices. That is, patients are supposed to understand themselves as still possessing agency, instead of being *insane* or helpless in the face of their disorder. Furthermore, outlining of the therapeutic process to the patient seems to serve the purpose of showing that psychotherapy is a useful tool to intervene on the disorder in question.

Finally, I take it that even the aim to present the patient as not being at fault for developing the disorder is implicitly alluded to in this quote – at least, this is how I read Kraines' claim that patients are relieved by knowing that their illness has a physical basis.[14]

Another piece of evidence for the importance of rationalization in psychotherapeutic explanatory practices comes from his claim that explanations in psychotherapy should be formulated in such a way that the patient feels that the physician or therapist "[...] 'understands' that there are 'justifiable' reasons for the patient's anxiety [...]" (Kraines 1957, p. 413), since this supposedly leads to a reduction of the symptoms.

Furthermore, Kraines mentions at one point that "Certain things you do and feel can retard or accelerate recovery." (Kraines 1957, p. 409), thus effectively pointing out possibilities for intervention that the patient has and the mental health professional is supposed to inform her patient about.

[14] If the illness has a physical basis, this means both that it can be treated (in this case, with medication or other physiological therapy), but it also means that the patient is not simply acting in a wrong manner, and thus, not responsible for developing the disorder. That is, pointing to the physical basis of the illness means pointing to a very deeply embedded cause of the illness.

Thus, these quotes indicate that I might be on the right track with those three aims of explanatory practices in psychotherapy that I have indicated above.

But there is more to take away from his book, when it comes to how explanatory practices in therapy should be set up. They are of particular interest for the question of how explanatory models are actually individualized in psychotherapy. When giving an example of an explanatory account of depression, the following is something that the therapist might say, according to the author:

> "The thing for you to remember is that this exhaustion can and will be overcome. You will need patience and you will need to cooperate. It won't be easy; it will take time; BUT YOU WILL RECOVER." (Kraines 1957, p. 410, emphasis in the original)

Given that, in reality, not every single individual recovers from depression, this statement is remarkable. We may conclude that either Kraines wants to allow the therapist to bend the truth somewhat, or that he has a flawed understanding of the course and variance of the disorder. Since he would only be justified in claiming what he does in case *every single individual* in fact recovers from depression – which he hardly can be ignorant of –, the first option appears more plausible.[15]

He adds: "The transforming agent is not the scientific accuracy of the explanation *but the patient's acceptance of it.*" (Kraines 1957, p. 413, my italics) At the very least, it seems that the patient's acceptance of an explanation is a necessary condition for it being helpful in the therapeutic process. The quote suggests that there may be cases where the "scientific accuracy" of an explanation and its acceptance by the patient are in conflict.[16]

[15] Even though Kraines does in fact describe the course of depression as *naturally leading to recovery* (Kraines 1957, p. 19). But clearly, the natural course of a disease or disorder does not need to be instantiated in every single case (Porta 2014, p. 193-194).

[16] I have, in this paragraph, assumed that the statement "You will recover." is actually (at least in part) a *prediction*. In effect, one may ask whether it should, following Hampshire & Hart (1958) not better be understood as an announcement of the therapist's decision to do whatever he can to make the patient recover, or as declaring the decision to work on the patient's recovery as a team. Even though this may be part of what is going on here, I take it that this statement *also* has an (at least supposedly) *predictive* component. I conclude this partially from Kraines' explicit distinction between the scientific accuracy of an explanation and its effect on the patient, which suggests that he does think of the statement here as a prediction of the future course of the patient's disorder. Furthermore, the status of the therapist as part practitioner, part authority on the facts on the matter (that is, the disorder) implies that his statement will at least be understood by the patient as partly predictive. I owe this remark to Dietmar Hübner (private conversation).

Kraines goes on to discuss *why* the truth of these explanations is less important than other factors: "More important than the element of truth in many such medical explanations is the fact that the patient, having a plan for the therapy of his emotional state through the 'cure' of his physical disorder, is enabled to make a social adjustment and does not give up hope." (Kraines 1957, p. 415) This is very similar to what I have suggested before, namely that these explanations are supposed to put the patient in a position from which she considers herself able to intervene on her disorder.

These statements by Kraines fit quite well with the fact that several of my interviewees pointed out that they not only sometimes deliberately presented simplified explanations of the mental disorder in question, but also sometimes presented a more positive image of the chances of getting better than the actual research results would suggest. This seems to indicate that, in the therapeutic context, the scientific accuracy of an explanation might be of minor importance. However, Beck, in his discussion of Kraines, points out that such reassurance should be used with caution concerning how it might be perceived by the patient. In particular, the client might feel invalidated if the therapist presents a very positive outlook on the further course of his disorder (Beck 1967, p. 316).[17]

At a later point, the author becomes more explicit concerning the effect he wants to achieve in the patients by giving these kinds of explanations: According to him, "the majority of patients are *comforted, sustained, and encouraged* by such a straightforward explanation of the illness and such positive reassurance of their recovery" (Kraines 1957, p. 410). It seems plausible to suggest that these effects are intended to bring about suitable conditions for psychotherapy, and eventually help the patient recover from his or her mental disorder.

Furthermore, when pointing out that the first important element of good psychotherapy is *understanding*, the author claims the following:

> "*Any* explanation – evil spirits, complexes, or disturbed neuronal circuits – is *equally reassuring* to the patient. The depressed patient feels alone, strange, different; if he is 'understood' by another, the very sharing of

[17] Interestingly, Beck does not (at least explicitly) advocate caution when it comes to the actual *correctness* of the explanation in question. It might be, though, that this is because it is taken for granted anyway.

> his symptoms makes him less alone, less strange, less different." (Kraines 1957, p. 413)

I think that we should not take Kraines too literally here: As I understand him, he does *not* want to claim that *any* possible explanation is equally reassuring to *any* patient. Clearly, for some people, explaining a mental disorder as possession by an evil spirit – an attempt to explanation that some priests still make (Hanwella et al. 2012) – is not as reassuring to the patient as an explanation that points to a hormonal imbalance in the brain. Thus, how reassuring an explanation is *to a particular patient* will depend on her social context, her background beliefs, and the like. As I understand the author, what he is getting at here is that the first helpful factor about giving an explanation of a mental disorder is that such an explanation makes what happens to the patient appear not just random and inexplicable, but as something that can be made sense of, and dealt with. I think that something similar is implicit when today's psychotherapists aim to present CBT as a useful tool for intervention for changing the patient's symptoms.

The author then adds to this another criterion. That is, optimally, the physician "sympathetically 'understands' that there are 'justifiable' reasons for the patient's anxiety" (Kraines 1957, p. 413). This supposedly reduces feelings of guilt and self-condemnation. Kraines even claims that the therapist should identify "understandable cause[s]" (Kraines 1957, p. 413) of the symptoms. These different mentions of *justification*, *understanding* and *reasons* seem to suggest that it is ultimately also about showing that the patient is still relatively rational, and thus, not "crazy" in the sense of experiencing symptoms that cannot be made sense of and predicted from the intentional stance.

In my understanding, these reassurances are ultimately about trying to demonstrate to the patient that (1) there is an internal, understandable logic to the disorder – because not understanding something that happens in one's own mind or body causes intense insecurity and suffering in most individuals – and that (2) the patient's experience and behavior are relatively rational.

Let me now present the results of these qualitative interviews, focusing on their import concerning the suggestion that these practical aims are operative in clinical practice.

3.4 Psychotherapeutic Explanatory Practices

Did the data from these interviews fit my hypotheses? In this section, I proceed as follows: I will begin by presenting data that indicates that the three goals I mentioned above are really operative in explanatory practices in therapy. Going into the details of the respective statements, I will present those strategies that therapists seem to use to accomplish these goals.

As a precondition for what was to come, I started each interview by asking whether – and if so, how – psychotherapists made use of explanatory models *at all* when explaining mental disorders to their patients. I was interested in particular in the most recent versions of the two models I have discussed in the beginning of this dissertation. All practitioners reported that they either used the model in question, a simplified version of it, or parts of it. There were different kinds of uses, though: They either used it to explain the disorder to the patient, explain it to themselves, or to keep in mind potential factors that they might intervene upon in treatment. In this section, I will focus on explanations that are given to the patient. I will, in the last section of this chapter, shortly come back to the other two kinds of uses.

Firstly, how does the naïve conception fare? Most therapists told me that they used only those parts of the model which they had found helpful for patients in the past, that their explanations were simplified in comparison to the actual models, focusing on those factors that can be intervened upon, and that they normally used them to *add* to the patient's self-understanding, not to *challenge* it — with the exception of cases where that self-understanding was harmful to the patient, including being detrimental to the therapeutic progress.

As I already stated, I take it that the main aim of explanatory practices is to ensure that the patient begins and successfully finishes structured talking therapy, which is facilitated if the therapist brings about two things in the patient: She (1) does not feel at fault for developing her disorder, while she (2) *does* take over responsibility for counteracting the disorder's symptoms. We can find this dual way of thinking about explanations of mental disorders represented in the following statement of one of my interviewees:

> "In any case, the fact that there are biological influences or factors, I do not conceal that. [...] I always say that, also [there is a] genetic vulnera-

> bility, and [the] reward system [is altered], here: amygdala. Yeah? Ehm, but at the same time, I would, if someone comes at me argumentatively: 'Yes, but I cannot change anything', I would say: 'Well, doing yoga can also change your brain.' Yeah? So, that there is some elbow room. [...] Also, that it is an illness. A diabetic is not at fault as well."

The therapist interviewed here points out quite clearly that, on the one hand, she understands depression as an *illness* insofar as the patient is *not at fault* for falling ill. On the other hand, she claims that certain actions can be taken such that the disorder's symptoms can be reduced. It is instructive that she compares depression to diabetes, that is, a *chronic medical illness*. In the example of diabetes, we tend to think that individuals are not to blame for developing this condition, but they are nonetheless responsible for acting in such a way that they keep their problematic dispositions from actualizing, if possible.

In the following, I will look at the three sub-goals of this overarching aim in more detail, focusing on the strategies that psychotherapists employ in order to invoke this particular self-representation in the patient.

3.4.1 Attribution of Limited Responsibility

Firstly, what is there to my hypothesis that the patient should understand herself as not responsible for having developed and for still suffering from the mental disorder in question? There were several statements of my interviewees that indicated that this was indeed the case. The most paradigmatic of those was the following, which was uttered by one of my participants when talking about how she presented the patient's disorder in therapy:

> "[...] it is not only his [...] behavior, which led to [the disorder], and now, he simply changes it, but that he also has a *predisposition* for [developing depression]."

Before commenting on how this is related to my thesis, let me quickly say something on a tension that arises when we compare this statement to the explanatory models from the first chapter. If we believe the words of the therapist, then, according to the patient's own default narrative, his behavior brought about the disorder in question. This seems puzzling at first glance.

In our explanatory models, the patient's behavior is usually described as a *symptom* or a *maintenance factor* of the disorder. In particular, what seems confusing is that the therapist appears to claim that *the same* kind of behavior that led to the disorder might *now* be changed. How should we best understand this?

I think that this reference to behavior as a *cause* of the mental disorder should be understood as implicitly hinting to the fact that maintenance factors of a mental disorder *can* constitute relevant factors in developing the *full* clinical picture: For example, in depression, avoidance of social situations – according to the model, a maintenance factor – strengthens depressive symptoms. Similarly, in the case of OCD, compulsive behavior heightens the probability for further intrusive thoughts and misinterpretations of these intrusions.[18] That is, particular behaviors may both be causally relevant for the development *and* the maintenance of a mental disorder. In effect, what the therapist here wants to distance herself from is the view that mental disorders are nothing but problematic and relatively superficial *habits* (compare Kinderman & Cooke 2017) that the patient consciously decided to engage in. This would make him somewhat blameworthy for developing and continually suffering from his disorder.

The idea that the patient's behavior is responsible for his suffering seems to be implied by understanding mental disorders as, nothing more than certain unfortunate behavioral and thought patterns.[19] The psychotherapist interviewed here seems to suggest that instead, the patient also has relatively stable features that predispose him for developing this disorder. I think that this predisposition can be understood as a *temporally* relatively *stable feature* of the individual, akin to, for example, personality factors (e.g., the *Big Five*, compare McCrae & Costa Jr 1999). Now, it seems that such features are relatively

[18] They are, I take it, referred to *merely* as maintenance factors in the model simply because their status as maintenance factors is more relevant in practice than the fact that they may *also* be causally relevant factors in the development of the disorder.

[19] It is important to note that, strictly speaking, this implication does not go through. Furthermore, importantly, Kinderman and Cooke understand themselves as mental health advocates who do, as I understand them, not want to suggest that individuals with mental disorders are fully responsible for developing and still suffering from their respective conditions. Nonetheless, I think that it is helpful to bear in mind that not only talk about mental problems as mental disorders can have problematic and potentially stigmatizing effects, but that talk about mental disorders as mere unfortunate habits may also run the risk of making people with mental disorders feel that they are either fully responsible for developing their conditions or not justified in their suffering. This, I think, also helps to explain why their guideline for journalists was met with some resistance also by patients themselves.

deeply embedded in the individual, thus exerting a certain control over her actions that makes it harder for the individual to act otherwise. Very often, the individual either (1) adopts them for good reasons – which is mostly relevant for the case of beliefs – or (2) did not consciously decide to adopt them – which applies mostly to predisposing factors like genetic conditions or the like. In both cases, the patient is less responsible for acting due to such features of hers than she is for other actions or behaviors of hers that are (due to) less deeply-embedded features. An important indicator for this is that the therapist explicitly speaks out against the idea of "simply chang[ing]" the behavior in question.

But why should representing the patient as not being at fault for his condition be what therapists are aiming at when presenting the patient as being predisposed for their disorder? There are several reasons for thinking this, one of them being that one therapist I interviewed explicitly said that what he tried to tell his patients was, among other things, that "what you are having, that's not your *mistake*". That is, he tries to invoke the perception in the agent that suffering from their mental disorder does not mean that they have acted wrongly. In my understanding, what therapists usually do to dismantle this impression is to show that the patient was relatively rational in adopting the beliefs in question.

To conclude, I think that the first quote shows that at least some cognitive-behavioral therapists consider it an important part of their explanatory practices in therapy to point to stable illness-disposing features of the patient. In my understanding, this strategy is intended to achieve several aims. For one, the patient should understand himself as not fully blameworthy for having developed this condition. Additionally, representing the patient as robustly disposed to develop this disorder also justifies engaging in long-term psychotherapy and might help the patient to develop reasonable expectations for the kind and the speed of progress that may occur in therapy.[20]

[20] This is, in fact, a whole topic in its own right. I asked several therapists about whether they considered vocabulary along the lines of "health" or "healing" appropriate in the context of therapy, and while opinions were divided on the question whether one should have the patient's health as the overarching goal of therapy at all – and whether "alleviation of symptoms" was not a much better aim to go for, given that, for one, there is this stable disposition for particular mental disorders, and for the other, that perfect mental functioning was unattainable. For example, one therapist said the following in my interviews:

> "You know, 'healing'? That's not the word for me in psychotherapy. It's not the word. [...] I don't heal. I alleviate harm, and in fact, this is also capacity building. [...] There are

3.4.2 Attributing Agency

What about the idea that patients need to understand themselves as possessing *agency* – understood as the self-conceptualization of still being in a position to act? Elsewhere, I have referred to this as "de-pathologizing". As I hinted at above, this goal can be divided into two related, but distinct, sub-goals. Both have already been described by Bolton (2008). He uses the term "normalization" for both of them and does not focus on therapeutic practices, but psychological research. Bolton distinguishes four forms of this:

> "(1) it may regard the abnormal as within the normal range of functioning
> (2) it supposes that abnormal emotions may appear as more appropriate, more understandable, more like the normal case, when the person's experience of the situation is better understood and taken into account
> (3) it emphasizes that much of what presents as symptoms of abnormal functioning are in fact strategies for solving problems, strategies that are reasonable within their own terms
> (4) psychology typically emphasizes that patterns of behaviour are learnt, and that dysfunction may arise when behaviours that are reasonable in the context in which they were originally acquired are applied in different contexts, but again the psychological processes involved are not qualitatively different from those operating in the normal case." (Bolton 2008, p. 16-20)

Both (1) and (2) roughly map on what I call "normalization", whereas (3) and (4) are more closely related to what I have been calling "rationalization". I will come back to the question of how well therapeutic practice aligns with this conception of normalization at the end of this section.

Firstly, there is the goal of *normalizing* the patient's experience in the sense of pointing out that her mental faculties are actually working very similar to the mental faculties of healthy individuals, or, alternatively, are carrying out their actual functions. It thus serves to provide counterevidence to many

> areas of medicine, where one, I think, can speak of 'healing', one can speak of a wound being healed. [...] So, "healthy". So, in our, in the psychotherapeutic context, right? Only as a joke. Right? [...] Only as: 'Now he's healthy. Haha."'

Although fascinating, I will sadly not be able to cover this issue in the following.

patients' belief that they may have a "broken brain" (compare, e.g., Andreasen 1985). Normalizing often runs counter to the idea that there is something *physiologically wrong* with the patient – as it would be if, for example, a mechanism in the brain would not carry out its evolved function.[21]

Secondly, there is the goal of *rationalizing* the patient's behavior and emotional experiences. With "rationalization", I mean showing that patients actually have (relatively) *good reasons* for behaving and feeling as they actually do and for adopting those beliefs that are at the core of their disorders. Importantly, rationalizing the patient's experience and behavior counters the belief that the patient may be *crazy*, may act without valid reasons, and may thus differ in fundamental ways from other rational agents.

Why does it make sense to distinguish these two goals? It seems that rationalizing the patient's experience and behavior does, to some extent, also amount to normalization. It would at least seem very odd for a human being to think and act rationally without also having relatively normally working mental processes. By contrast, normalizing the patient's experience and behavior does not necessarily also rationalize it: She may very well have normally working mental faculties without acting from (relatively) good reasons. This is the case at least if classical accounts of rationality get it right, according to which most healthy people very often act irrationally or make irrational decisions (e.g. Tversky & Kahneman 1974).

Normalizing someone's mental processes refers to a *statistical* standard: That is, what is normal is defined by what an average person does or how her mental faculties operate. Rationality is arguably more demanding and more normatively laden in the sense of requiring that someone, to count as rational, must have (relatively) good reasons for thinking or acting as she does.

I will now tackle the question whether these two sub-goals plausibly influence explanatory practices of mental disorders in psychotherapy, one after the other.

[21] In my interviews, several participants referred to the belief that mental disorders are fundamentally just physiological disorders as the "medical model" of mental disorder. But since talk about the "medical model" is hopelessly vague and ambiguous – as has been pointed out by, for example, Lilienfeld et al. (2015, p. 9) –, I will try to stay clear of this terminology, instead trying to say more precisely what the usage of this term amounts to in each particular instance.

Normalization

Let me start by discussing the first. Do psychotherapists actually try to *normalize* the patient's experiences when providing them with explanations of their mental disorder? As I have already pointed out, normalization provides an alternative understanding of the agent's experiences and behavior that is opposed to her usually negative self-evaluations as weak or even morally deficient because she is mentally ill, and thus, not functioning *normally*. The particular phenomenon that individuals with mental disorders do not only suffer as a *direct* consequence of their symptoms, but also as a consequence of devaluing themselves or worrying *because* they have symptoms is often called "secondary disturbance" or "symptom stress" (Joshi & Phadke 2018, p. 77). Symptoms stress is usually reduced substantially once symptoms are presented as outcomes of relatively normal mental and behavioral processes of the individual. As one psychotherapist said:

> "It [the model] can be suitable for this, too. *We are all in the same boat*, what you are having, that's not your mistake, your deficit. You're not crazy, [...] in your case, something has run out of control, you usually have too much of something, and too little of something else. And this dysbalance, it causes you to suffer. This produces tension, pressure, weird experience. [...] Across the board, people say: 'I understand this. This is great.'."

This indicates that there may be truth in my suggestion that explanatory practices in psychotherapy often focus on showing that the patient's mental faculties work just fine. As this interviewee said, it is merely a *dysbalance* between different processes that causes the symptoms. To see why this is interesting, consider Jerome Wakefield's account of mental disorders as *harmful dysfunctions*. According to him, someone suffers from a mental disorder just in case a mental or brain mechanism fails to perform its evolved function (Wakefield 1992) and this brings about harm. In contrast to this, in the case of dysbalances between different kinds of systems, it may very well be the case that all relevant neurocognitive mechanisms[22] actually perform their

[22] I am speaking of "neurocognitive mechanisms" here in order to avoid two things: (1) needing to discuss whether there can be purely cognitive mechanisms and (2) needing to discuss whether Wakefield's account is confined to neurological or other mechanisms. Both issues are interesting, but only of minor relevance to the issues addressed in this chapter.

evolved functions, but nonetheless cause intense suffering. One example of this is *Social Anxiety Disorder* (*SAD*), which Wakefield does not view as a mental disorder: According to him, all mental mechanisms that we know to be involved in this condition actually perform their evolved function (Wakefield et al. 2005). In SAD, the mechanism that causes individuals in today's society to suffer supposedly has the function to ensure that the individual would not loose social status in a group by prompting a fear reaction when confronted with a situation where one would perform behaviors that could result in being negatively evaluated by other members of one's own group (for the diagnostic criteria, see A.4).

Nonetheless, there is enormous suffering in patients with SAD, and arguably, their responsiveness to stimuli that potentially signify social threat is overly strong, when evaluated against what is adaptive in today's society. Thus, the different systems could count as dysbalanced.[23] In this case, someone would be diagnosed with a mental disorder – and, I would argue, actually *have* one – based on the fact that she is suffering, even though the mechanisms that cause this suffering carry out their evolved functions.[24]

To come back to the quote above, I consider it interesting that, with the last sentence, the therapist openly justifies his explanatory practice by referring to the fact that his patients find it "great" or helpful. This fits quite well with the attitude that Kraines appears to have about the respective importance of patient well-being and an explanation's correctness.

There is another statement that I would like to mention here, because it highlights the intimate connection of normalization and function concepts:

> "I think the most commonly used, probably in all psychotherapeutic practices, especially used by myself, is the vicious circle of fear, that also, in principle, consists of these elements. Thought, emotion, physiological change, ehm, and I have such a standard lecture over 15 minutes, where

[23] I am of course aware of the fact that this is nowhere near constituting a good argument against Wakefield's account, but merely shows how our premises differ: While he is more willing to give up on the status of SAD as a mental disorder than on his account of what mental disorders are, I am more willing to give up on his account of what mental disorders are than on the status of SAD as a mental disorder.

[24] Clearly, for someone who takes the harmful dysfunction view as primary, my reasoning here simply begs the question. This is particularly because, for me, one central criterion of adequacy for an account of mental disorder is whether it agrees with established diagnostic practices in the field, and, in particular, gets the right kind of result in (what practitioners consider to be) paradigmatic cases of mental disorders.

> I explain precisely what happens in the body concerning physiological changes, ehm, when someone is afraid. Because I think... especially in the case of fear, yes, I don't know, it de-catastrophizes a lot, if one can imagine that as a *meaningful, logical, helpful* bodily event."

This shows quite clearly that, for this therapist, the aim of pointing to the vicious circle of fear is to point out that anxiety is not harmful in itself. Quite the opposite, anxiety is presented as a bodily reaction that carries out a particular *function* – this is how I understand his description of fear being "meaningful, logical [and] helpful". Importantly, this way of normalizing by pointing to particular underlying functions of the individual's symptoms is not normalizing via *statistical normalcy*, but normalizing via *functionality.*[25]

I think that the main strategy for achieving this goal is to point out – in consonance with the two explanatory models I have presented before – how the workings of normal psychological processes, often processes that actually serve a biological or evolved *function*, account for the symptoms that the patient is suffering from.

Rationalization

I will now continue with a discussion of *rationalization.* Concerning this aim of explanatory practices, one should note that it is *also* about the justification of the agent's experience and behavior: It evaluates the agent's experience and behavior as *reasonable* and *permissible* when held against a particular standard.

Above, I have pointed out that explanations that are given in the therapeutic context take mental contents of the patient into account. A patient can only *really* understand why he developed and still suffers from a particular disorder when understanding the beliefs at its core. Such an understanding for himself does, I take it, not only serve the function of providing the feeling of being in control, but it also carries a sense of *justifying* the agent's actions. For example, one of my interviewees said the following when talking about how she constructs individual explanations:

[25] One might even point out that, actually, this de-pathologization via pointing to the functions of particular mechanisms that bring about the patient's symptoms is located somewhere on a dimension between normalization and rationalization: It is not, one might claim, solely about statistical normalcy, but it is also not about being *reasonable* (that might be too strong of a requirement). Instead, it is about the patient's psyche still *functioning properly.*

> "Learning history, how does it look like, which experiences has she had, how did she grow up, so achievement orientation is usually another fact [...] Or also, parenting style, so: What was important? Hedonistic principles, but also performance orientation or so, is the person able to enjoy things or not? Which would be another resource, or is she very performance oriented, which would be a risk factor for depression or anxiety. [...] I would say, it can also be a cognition, or schemata, that are activated, like: 'I am only worthy of love if I perform well.', 'I am not allowed to fail.', 'I should always be capable.', or 'I am only worthy as a person, if...'."

This shows the importance of referring to certain *beliefs* of the patient that are taken to be (partial) *reasons* for that patient's actions – which may be either open behaviors or thought patterns – or that allow us to easily derive the agent's *motivating reasons* (compare Alvarez 2017). That is, without the belief in question, the patient's behavior would not make sense. But given that particular belief, it is possible to present the actions of the patient as the conclusion of a practical syllogism scheme. I will expand on this point in the next chapter.

In this way, therapists make sense of those actions of the patient that may be hard to understand at the outset.[26]

Furthermore,the importance of such factors already emerged from the discussion of the explanatory models as such, since these explicitly refer to dysfunctional *beliefs* as the sources of the agent's problems. Just consider that, in several cognitive-behavioral models of mental disorders, the individual's beliefs are referred to as *reasons* for her behavior and emotional experience. Remember my reconstruction of the explanatory model of OCD: There, certain dysfunctional assumptions emerged as bringing about misinterpretations of intrusive thoughts as dangerous. These assumptions are not only causally

[26] I think that these explanations are, in fact, *action explanations* (Wilson & Shpall 2016). This makes it necessary to understand the compulsive behavior of someone with OCD an action. I think that action explanations provide an understanding for why the compulsive behavior in question was carried out at first, but that this intention gets lost, the more and more automated the individual's behavior becomes in response to the threatening stimulus. This also fits Salkovskis' claim that compulsive behavior can become so automated in someone with OCD that the negative emotional state that motivated the agent to engage in this behavior in the first place, ceases to occur. Similarly, the intention might not be present in the individual every time she carries out the compulsive behavior. Note also that, for this to make sense, we must allow for pre-conscious or unconscious intentions.

related to the symptoms, but they also *rationalize* them. If someone actually believes that intrusive thoughts *are* indicative of the behavior she will carry out – trying to kill one's colleagues, let's say, with the example given before –, then it *makes sense* for her to think that she may be a danger to others. Given that the individual does not want to bring harm to her colleagues, refraining from going to work seems entirely reasonable. Similarly, on this model of OCD, neutralizing actions emerge as relatively reasonable, given the individual's wish for herself and others to be safe in combination with believing that neutralizing actions actually reduce the danger inherent in the situation.[27]

Among other things, I think that those two strategies that psychotherapists pursue in their therapeutic practice are very similar to those aims that Bolton (2008) describes. Nonetheless, my account makes a substantial contribution to an understanding of models of mental disorders over and above his account of normalization. It shows, for one, how exactly normalization is carried out in psychotherapeutic practice, and for the other, the distinction between rationalization and normalization is one that can fruitfully be exploited for understanding therapeutic practice better. I will add more details to that in the subsequent chapter.

3.4.3 Presenting Possibilities to Intervene

Thirdly, do psychotherapists aim to make the patient believe that she – by using techniques from CBT – can *intervene* on her disorder? On this matter, one of my participants, taking the perspective of the patient, said the following:

> "[...] another possibility, why such a model is of course important is, somehow, to say: 'Okay, and *what will I do differently in the future*? [...] and this means, if I go into the same kind of situation next time, I can make some deductions on the basis of the vicious circle of fear. For example, I can try not to catastrophize in my head, [...] I can remain in the situation, because I have learned something about habituation [...]' "

What I find very interesting is the focus on the patient needing to intervene by

[27] In some cases – namely, when the compulsive behavior in question does not *really* reduce the alleged danger in the situation and the individual knows this to be the case –, the rationalization in question works differently, namely, via the individual wanting to feel less anxious, and carrying out specific kinds of behavior because of their effect on her emotional state.

himself: In the example that the therapist presents, someone suffers from panic attacks, and these are explained by the so-called *vicious circle of fear*, the idea that panic attacks are the result of a negative interpretation of normal bodily anxiety symptoms (Clark 1986) that I have already described above. This process is known as "catastrophization", because a relatively minor trigger may be sufficient to lead to a full-blown panic attack. Since it is easiest to intervene on the cognitive appraisal, the patient is usually asked to try to change it.

Now, when another therapist talked about why she usually did not place a huge emphasis on those physiological or biological factors that co-occur with the psychological symptoms of mental disorders, she said the following:

> "If I emphasize too much that there is also a biological correlate, this very often results in a feeling of 'then I cannot change anything anyway' [...] Now, if it is only somatic, and it is these transmitters up there, *how am I supposed to intervene*?"

The last sentence must be understood as a question that the patient asks himself when being confronted with explanations that emphasize biological correlates. This statement shows that using folk-psychological vocabulary also contributes to the aim to present the patient with possibilities for intervention: The patient can only intervene on his disorder if it is explained to him at least partially by referring to factors that he can observe, either being open behavior or introspectively accessible mental states. Thus, if psychotherapists want their patients to be able to intervene, their explanations should not be formulated wholly in sub-personal terminology.[28] Only then is he able to develop coping strategies to deal with his symptoms in his everyday life, since this terminology aligns well with how most of us think about our mental lives and the mental lives of others. The quote above also shows that this psychotherapist assumes that patients have a *need* to feel able to intervene in order to alleviate their symptoms.

We can thus infer that one of the positive features of such clinical psychological models of mental disorders, for psychotherapists, is the fact that they give

[28] Of course, one could argue that it would indeed be possible to explain the disorder, using only sub-personal concepts and then linking those sub-personal concepts to possible interventions on factors that can be manipulated by the agent. It does seem, however, as though this way of framing the matter – even if there were good explanatory models available for doing so – would not have the same effect on the patients (at least if my interviewees are to be believed).

patients the impression that they have some control over their symptoms. This also shows that the aim of presenting the patient's disorder as caused by something different than a broken brain hangs together with the aim of presenting CBT's tools as useful for intervening in the disorder.

The therapist's strategy for achieving this aim would thus amount to presenting the patient with epistemically accessible features that are (at least partially) under the agent's control.

To conclude, I have shown in this section that particular kinds of goals influence and shape those explanatory practices that occur in psychotherapy. This does not yet, however, imply anything substantive about the features of explanatory models of mental disorders. Let me tackle this issue now.

3.5 Practical Aims and Features of Explanatory Models

Is it plausible to assume that, in addition to explanatory practices in psychotherapy being influenced by certain practical goals, these goals also influence the content of the *explanatory models* that are used in constructing these idiosyncratic explanations? Put differently: Do practical aims, by influencing explanatory practices in therapy, also affect the explanatory models themselves? In arguing that they do, I would like to go back to an observation from the last chapter, namely, that the explanatory models in question were first constructed on the basis of evidence obtained in the psychotherapeutic context. Furthermore, I will point out that these practical aims fit surprisingly well to the noteworthy features I have discussed above, making it plausible to assume that the models actually have these features *because* these aims influence explanatory practices in psychotherapy. In other words: Not only the idiosyncratic applications of explanatory models, but also the models themselves are, as I will argue, influenced by these practical goals.

To recapitulate: As we have seen, Beck's 1967 model of depression, which my two exemplar models are based on, was developed on the basis of information gathered within talking therapy. Talking therapy usually starts with the patient's self-understanding or self-conceptualization. For this reason, it will almost inevitably employ personal-level, folk-psychological vocabulary. This makes the usage of folk-psychological vocabulary in these models appear nearly trivial. Furthermore, Aaron Beck explicitly refers to Samuel Kraines' *Mental*

Depressions and Their Treatment in his first book on depression. Kraines, as we have seen, explicitly advocates for explaining mental disorders to patients in a way which ensures that particular practical goals are achieved.

How do the therapist's strategies exploit the features of these models? Let us start with the usage of folk-psychological vocabulary in these models: Folk-psychological vocabulary is related both to the pragmatic goal of conceptualizing the patient as possessing agency despite his mental disorder and to the goal of making him understand that he may be in a position to intervene on his symptoms. The use of folk-psychological vocabulary contributes to reaching this goal by contributing to the *rationalization* of the patient's experiences.

Furthermore, there is the reference to particular relatively stable, usually *dispositional* features of the agent. I presume that this is used to limit the attribution of responsibility to the patient. That is, it is used in part to draw the line between the patient merely having "unfortunate habits" and having a more "deeply seated" condition. These features are usually referred to as "dysfunctional" beliefs or behavioral strategies.

Finally, there is the *normalization* of mental disorders by way of emphasizing the similarity of those mechanisms that are productive of the problematic symptoms with the mechanisms that are operative in healthy individuals. Additionally, the function of these mental mechanisms is often emphasized. In my understanding, this is also connected to the aim of conceptualizing the patient as possessing agency.

When considered in isolation, the convergence between practical goals of explanatory practices and features of these models is surprising. It becomes less surprising once we take seriously the possibility that the two might influence one another – that is, that those aims that were relevant in the context of application might have exerted a considerable influence on what seemed to be good explanatory practices. That is, a *good explanation* would have partially been characterized by the extent to which explanatory practices achieved these practical aims. If we now assume that explanatory practices accumulated in the form of Beck's model of depression, then this would explain why the content of this model seems to fit well to these practical explanatory aims.

I consider it plausible to conclude that the kinds of models that are employed in clinical psychology today arguably have the noteworthy features discussed above because

1. They were constructed in the context of application, which partially explains the presence of folk-psychological terminology,
2. Within clinical psychology, there is a preference for models that can serve particular practical aims, and
3. The models are sufficiently operationalized as to be relatively well testable – which supposedly distinguishes models from the psychoanalytic tradition.

The last point on this list may also help us understand why the cognitive and cognitive-behavioral models of mental disorders are more widely used in psychotherapy and more thoroughly researched today than explanatory models that are based on psychoanalytic theory. This is not obvious, since psychoanalytic models were also constructed in the context of application, contain folk-psychological terminology as well, and can also be used for achieving the practical aims in question. They are, however, harder to test, as has been pointed out repeatedly in the literature (compare Grünbaum 1984). A further factor of importance is that there is some consensus in the discipline to the effect that CBT currently is the best treatment we have (e.g. David et al. 2018). That is, it allegedly is more effective for treating mental disorders than both psychoanalysis and even psychoanalytic treatments – that is, treatments that have emerged from psychoanalysis (compare, e.g. Shedler 2010). Furthermore, the evidence base for psychoanalysis itself is thin (Fonagy 2003).[29] This may have lead psychotherapists and researchers in clinical psychology to stop holding these explanatory models to be true, as well.

As one professor of clinical psychology and psychotherapy pointed out to me in a private conversation, he believed that psychoanalytic models were less relevant than cognitive-behavioral models and they were primarily studied by researchers who reflected on the model's historical importance for psychology. By contrast, he used cognitive-behavioral models in clinical practice, in research and teaching. My own experience as a student of clinical psychology and many personal conversations with psychologists and psychotherapists

[29] This issue is subject to much debate, with many psychoanalysts arguing that their method can for principled reasons not be evaluated with the same kind of evidential standard as CBT (Fonagy 2003, e.g.).

aligns with this view. Furthermore, those handbooks that I used as a student either did not mention psychoanalytic models of mental disorders at all, or, if they did, only to point out their historical relevance for more recent developments in the discipline.

For the purpose of clarification, let me very quickly summarize the argument I have developed over the course of this and the preceding chapter:

($\mathbf{P_1}$) Explanatory practices in psychotherapy are intended to achieve particular practical goals.

($\mathbf{P_2}$) Several explanatory models of mental disorders have originally been constructed in the psychotherapeutic context.

($\mathbf{P_3}$) Explanatory models of mental disorders are based on evidence obtained in the context of application, that is, psychotherapy.

($\mathbf{P_4}$) Within clinical psychology, there is a preference for models that can serve particular practical aims within psychotherapeutic practice.

(**C**) Those pragmatic goals that guide explanatory practices in therapy also influence the form of those models that are widely accepted in the discipline.

Clearly, the conclusion does not follow deductively from the premises. Instead, the argument is abductive: Those noteworthy features of explanatory models that I have identified here map onto the explanatory aims that are operative in clinical psychological practice. As I have pointed out, this is surprising and in need of explanation. I take it that one good explanation for it is that there is an influence from clinical explanatory practice on those features of the models. Assuming such an influence to be present becomes more plausible when considering that these models of mental disorders are first constructed in the context of application.

Quite roughly, it seems plausible to me to suggest that the kind of therapy also shapes the respective explanatory model, not only the other way around. This holds especially for the contrast between classical psychiatric – that is, medically-oriented – and psychotherapeutic treatments of the disorder.

Let me now consider a further issue that arises when dealing with different models of mental disorders that are used in treating patients and that may be of particular relevance for psychotherapy. This is the issue of feedback effects, which is relatively similar to Ian Hacking's *looping effects*.

3.6 Feedback Effects in Therapeutic Practice

As we have seen, psychotherapists aim explicitly at changing the self-understanding of their patients when explaining mental disorders. Their idea appears to be that, by explaining the patient's disorder in a particular way, his self-conceptualization will be modified, which will change his behavior and experience and alleviate the symptoms. Note that the explanatory practice would arguably have this effect even if the psychotherapeutic process would be discontinued after giving this explanation.[30] If true, this process is strikingly similar to a phenomenon that is widely discussed in the philosophy of psychiatry, namely so-called *looping effects of human kinds*, first described by Ian Hacking (1995). It is particularly interesting when considering the fact that different self-conceptualizations of the patient may lead – as some of my interviewees have pointed out when contrasting medical models of mental disorders with clinical psychological models – to different kinds of behavior and experience. These may, again, have differential effects and thus be incorporated into new explanatory models of the disorder. Now, it seems like the different explanatory practices that are based on these very models actually lead to systematically different kinds of behaviors and experiences. For these effects to be conceptualized in parallel to looping effects, these differences in behavior and experience need to be a result of the explanation in question that *goes beyond* the effects of the actual therapeutic intervention. Let us thus try to see whether such effects may play a role in the construction and use of these explanatory models.

In his well-known paper, Hacking characterizes looping effects as follows:

> "To create new ways of classifying people is also to change how we can think of ourselves, to change our sense of self-worth, even how we remember our own past. This in turn generates a looping effect, because people of the kind behave differently and so *are* different. That is to

[30] This is an empirical question that can hardly be decided on merely analytical or conceptual grounds. But I think that the fact that psychotherapists could point to differential effects of medical models of mental disorders in their patients is a good first indicator to adopt the working hypothesis that such effects may indeed occur. This is especially true since psychiatric and medical treatment of individuals with mental disorders often does not involve much more than (often implicitly) presenting the patient with a physiological model of her disorder and prescribing a particular medication (I take this from several private conversations with psychiatrists, psychotherapists and patients as well and my own experience as an intern in psychiatric clinics). Note that this of course *need* not be the case.

> say *the kind changes*, and so there is new causal knowledge to be gained and perhaps, old causal knowledge is to be jettisoned." (Hacking 1995, p. 369, my italics)

In other words: Looping effects occur whenever the act of introducing a new way of classifying someone or something as being of a certain kind results in systematic changes in the individuals so classified, such that, eventually, the kind itself changes.[31] Usually, this change is brought about because the individual in question starts to think about herself differently. I think that it is instructive to consider what the underlying mechanism is supposed to be, which I would like to do by giving an example. But first, let us be clear what exactly we are talking about here. For the purpose of clarification, it is important to keep three kinds of objects apart: The *linguistic term*, the *concept* and the actual *kind*, which, I take it, is a feature of someone or something.

One example for a kind that historically created looping effects is – to go beyond the example of *multiple personality disorder* discussed at length by Hacking (1998) – the kind *deaf-mute* that was used to classify deaf individuals (Söderfeldt 2013). Historically, the term "deaf-mute" was used for a long time to classify deaf people who used sign language[32] or who were not able to use an oral language. This term is laden with negative connotations, resulting in the individuals falling under the concept being treated differently, but also in them conceptualizing themselves differently than before being so classified. Arguably, this led to changes in *what it is to be deaf*: For example, it seems plausible to assert that at least part of the often worse socio-economoic status of many deaf people in comparison to the hearing (Emmett & Francis 2014) might be due to being conceptualized as deaf-mute, and thus, being understood as incapable to communicate and having a deficit that needs to be fixed. This may lead to a self-fulfilling prophecy: By both being treated as though they were unable to communicate and understanding themselves as having a communication deficit, many deaf people would indeed not be suited to pursue particular careers. Thus, in Hacking's terms, there was new knowledge about

[31] Clearly, on this understanding, kinds are dependent upon their instances, nothing more.

[32] One might ask whether these deaf individuals *really* had sign languages at the time – especially knowing that, at least after the so-called "Milan conference" in 1880, there was a trend to not teach deaf individuals sign language, but to use the so-called "oral method" (Moores 2010). But, at least if the personal report of deaf individuals who went to school during a time when the usage of sign language was still frowned upon are to be believed, these prohibitions were rarely effective in keeping these individuals from using sign languages.

people to be gained that fell under the label of "deaf-muteness" that was not there before the kind was introduced: They indeed *were* not in a position to take up particular kinds of careers, precisely *because* deaf individuals understood themselves – and were understood by others – as unable to properly communicate and stupid.[33]

This example can only serve as a first illustration of Hacking's idea, but it suffices for what I want to point out here.

I think that we are able to observe something similar to looping effects in psychotherapeutic explanatory practices. Because the phenomenon that I describe is merely similar, but not identical to Hacking's concept, I will call it "feedback effect" in the following. As I sketched above, there seem to be changes in explanatory models that are caused by the patient's reactions to the specific explanatory practices employed in therapy. One feedback loop that may occur here roughly amounts to the following: (1) the patient is depressed, and, thus, reaches out for help, (2) he is diagnosed with the disorder, that is, the kind concept is applied to him, but importantly (3) his depression is explained to him with a cognitive-behavioral model from clinical psychology, which leads to (4) the patient changing his behavior as a result of having been given this particular kind of explanation (instead of a medical explanation), since he feels, for example, less blameworthy for having developed his condition and thus, (5) by acting in accordance with this model, he provides the mental health professional with what appears to be further evidence for the model's correctness.

If such feedback effects occur in psychotherapeutic practice, they may have *both* positive and negative effects – which stands in contrast to how Hacking's looping effects are usually understood, that is, as having only negative consequences. A positive result of such feedback loops in psychotherapeutic practice might be that the individual's change in behavior will usually mean that the patient better exploits those options for positive changes that exist. But there

[33] This is both very rough and extremely simplified. There of course are particular kinds of careers that were historically harder or impossible for the deaf or hard of hearing to pursue. But what I am thinking of here are rather careers which, for example, necessitate a university education. There is no compelling reason why deaf people should be less capable of studying at universities – in principle –, but they do so, in fact, much less frequently than the hearing (Garberoglio et al. 2017). This is partially due to pragmatic obstacles like access to sign language interpreters (which appears to be still problematic, according to private conversations with deaf individuals in Germany) – which are problematic in their own right, but not my main focus here –, but may in part still be due to particular kinds of stereotypes and conceptualizations of deaf individuals.

may also be negative effects of such feedback loops. For example, they might result in behavioral changes that merely *appear* to fit into psychological models of their disorders, while the patient may actually have been better served by a treatment that focuses on the medical side of his disorder.

Why would one think that such feedback effects actually occur in explanatory practices in clinical psychology and psychotherapy? Firstly, several psychotherapists I interviewed suggested that something like this was the case: Some of my interviewees complained about how their patients often started psychotherapy with specific "medical models" in mind, especially when they had received medical care before. According to my interviewees, this often influenced the patients self-conceptualization in a way that made them more passive in their behavior, expecting mental health professionals to *heal* or *cure* them through the use of medication or other – usually medical – procedures. This was, according to my interviewees, due to the fact that these patients had previously been in contact with explanatory models that located the cause(s) of their disorders or the disorders themselves on the biochemical or physiological, instead of psychological or intentional, level.

One example of such a model is the so-called "serotonin hypothesis", according to which the symptoms of depression are ultimately caused by a lack of serotonin in the brain (compare Lacasse & Leo 2005). This model suggests that, to cure depression, what needs to be done is to fix the amount of certain neurotransmitters in the brain. If this is true, then it is at least not obvious at first glance how behavioral or cognitive interventions would be suitable for treating depression.

Thus, specific feedback effects appear to occur here at least in the sense that certain explanatory practices influence the kind in question via leading to changes in the behavior of the individuals falling under the kind: Explaining depression as due to a lack of serotonin changes the patient's behavior, and thus, may give rise to or further reinforce the idea that depressive people are not in a position of control over their disorders. Accordingly, several therapists commented on how it was harder to work with patients who understood themselves as victims of a purely physiological condition who bore no responsibility for their disorder at all.[34]

[34] Please bear in mind that I have no intention to suggest that there are not at least some mental disorders that are, indeed, purely physiological or medical conditions. The fact that psychotherapists find it hard to work with patients adhering to a particular understanding of their individual mental

In addition to this, one psychotherapist provided further evidence for my suspicion that both the patient's self-conceptualizations and the explanatory models of mental disorders may change, due to differences in the individual patient's behavior that were, again, brought about by a shift in her understanding of the mental disorder. According to him, the effects of explaining depression according to either the medical or the psychological model were substantial. One might even suggest that different explanations of the same disorder may yield differences in the patient's self-understanding, and thus, in her behavior that are substantial enough to yield distinct forms of depression. In this case, the respective ways of conceptualizing the patient seems to elicit positive feedback-effects through the self-conceptualization of the patient, resulting in the production of what looks like more evidence for the correctness of the respective model in the interaction of patient and therapist.

Note that there are differences between *concepts* of mental disorders and *explanatory models* of these disorders: Take the human kind *Major Depressive Disorder*, for example. Arguably, the *concept* of depression is – roughly – encoded by the diagnostic criteria in the DSM-5.[35] Nonetheless, there is a multitude of different explanatory models of the disorder on the market, even if we limit ourselves to clinical psychological models: For example, there is the cognitive-behavioral model of MDD on the one hand, and several different psychoanalytical models, on the other. Now, these models of the disorder arguably have developed in the past without the *concept* of MDD necessarily changing as a result. It gets a little more complicated if we consider the other direction, though: If the symptoms of a mental disorder change, then a change in the explanatory model might need to follow, but this is not necessarily the case.[36]

disorder does not mean that these disorders are not, in fact, purely physiological conditions, although it is indicative of the fact that at least some mental disorders are not purely physiological conditions.

[35] This is of course not *obvious*, since one might also assume that really, the diagnostic criteria that are mentioned in the DSM really only function as *indicators* for the underlying mental disorder. But remember that, according to the DSM-5's characterization of mental disorders, they are really sets of symptoms, that is, *not* the underlying dysfunctions. Now, if mental disorders really are to be understood as sets of symptoms, then it makes sense to assume that they are precisely those sets of symptoms that are mentioned in the DSM's diagnostic criteria. Furthermore, even if the diagnostic criteria do not strictly *make up* the concept of depression, they still restrict the concept of depression to particular symptoms, whereas the explanatory models mention both symptoms and further factors. This is, in fact, sufficient for my purposes here.

[36] Just think of the recent change in the DSM-5 in how depression was conceptualized: Complicated grief is no longer seen as a reason for someone to not get the diagnosis of depression after experiencing depressive symptoms for longer than two weeks. This is a substantive shift in the diagnostic criteria, but it does not need to change the explanatory model of depression, since that

As we have seen, those feedback effects brought about by the sole availability of a medical model of depression might result in making the disorder more difficult to treat with talking therapeutic methods, since these methods require the patient to actively change her behavior or thought. The same problem does not occur in the other direction: Since the patient does not need to take action herself for her medication to work, it might not be ideal to only explain her mental disorder in intentional terms, but it would not lead to the same kind of problem that is arguably caused by explaining someone's mental disorder according to a medical model and then trying to make that same patient intervene on particular maintaining factors of her disorder.

This indicates that the usage of medical models ultimately changes the model itself by influencing the individual patient's self-understanding, which then influences her behavior. Now, it seems that, if clinical psychological models of the disorder also lead to changes in the patient's behavior, a feedback effect would occur here, too. But let us try to make this more explicit.

Let's say that someone's depression is explained as due to specific dysfunctional beliefs that are caused by particular early life events. The exact result of feedback effects depends upon the prior conceptualization of the mental disorder by the individual. In the case of depression – at least if we believe my interviewees[37] –, it is highly likely that she conceptualized her depression as a medical illness in the sense of a neurochemical imbalance in the brain. On this understanding, there is not much the patient can do: Instead, she must rely on the expert to treat her disease. If the conceptualization of her disorder changes later and she understands it as caused by particular early experiences and problematic beliefs, she will probably regard herself as more able to influence this disorder. According to my interviewees, this may already prompt changes in her behavior.

model only deals with the explanation of the occurrence and maintenance of depressive symptoms.

[37] For example, one of my interviewees said the following:

> "We will talk about diagnoses at some point, because, for example now, in the documentation of the first session, I have to record a diagnosis. [...] At the same time, I don't like diagnoses. Thus, I rather have a systemic perspective, to say: A problem is when someone says "I have a problem", or "there is a problem". [...] Sometimes I think, or sometimes I also say that, because it differs, because it's a contrast to those experiences which they have made with other people. Not everyone, but not few have already talked to a physician about it, with their general practitioner, or they have been somewhere else, where they have been confronted with the classical medical model. And have been treated accordingly. And then there is someone who takes his time, who asks, and says: "I want to get to know this, tell me about it. And so on. And this is – many people like this."

3.7 Conclusions and Further Directions

Let me reconsider the first question that I posed above. That is, why are explanations of mental disorders taken to be *essential* for successful therapy? I think that we have seen in this chapter the many ways in which explanatory practices are important for successful psychotherapy. That is, these practices serve certain practical goals of the therapist. They allow for a representation of the patient that makes it more likely for that individual to continue psychotherapy, but also to work on her problems on her own. They also give possibilities for the individual to intervene, and thereby, reduce her symptoms. Strictly speaking, without these explanatory practices, there is no rationale given for the interventions that are carried out in the psychotherapy. Furthermore, they supposedly stimulate hope in the patient by representing him in a certain way that makes him less blameworthy for developing and maintaining the disorder.[38]

As I stated above, these explanatory models may, *in addition* to being used for the purpose of explaining the mental disorder to the patient, be used to (1) explain the mental disorder to the psychotherapist herself and (2) for the therapist to keep in mind potential factors that she might intervene upon in treatment. To be able to intervene by means of such a model, it seems that the therapist would need to believe the model to not only be useful for evoking particular positive effects in the patient, but she would also need to believe these models to be – at least to a certain extent – factually accurate.

These two aims pull into different directions. Firstly, it is important to keep in mind why the therapist would explain the patient's mental disorder to herself *at all*: According to my interviewees, the main reason for trying to understand the patient's disorder, and most importantly, the interrelations between the different symptoms is to derive potential interventions. The conceptualization that therapists have in mind of the patient's mental disorder does not always coincide with the explanation that she presents to the patient.

What those therapists I interviewed often mentioned as differences between their personal understanding of the patient's disorder and the explanation

[38] Whether working against one's symptoms is the very same thing as working against one's mental disorder is debatable. Several patients would, for example, claim that even though they have failed to actually get rid of their disorder, they have managed to reduce their symptoms such that they are able to lead a happy, fulfilled life.

they provided was the respective explanation's *level of precision*: That is, their own conceptualizations would be much more complex than what they presented to the patient. To my mind, this kind of difference isrelatively unproblematic. Things become more complicated with other dimensions that some of the therapists mentioned, concerning how positive or negative they presented the patient's situation. Similarly, some therapists mentioned that there are kinds of disorders – something I have come across several times was the example of NPD – that required not presenting the patient with the full picture of his condition, mainly because that condition implied a negative value-judgement about the patient's character. It was interesting to see how strongly the therapist's opinions diverged on whether it would be morally justifiable to present the patient with what we might call a "sugarcoated" view of their disorder: While some of them considered this to be absolutely unacceptable, other therapists openly admitted doing this, since it supposedly has a positive effect on patients. Although this is a fascinating issue in its own right, it is not one that I will be able to pursue in the remainder of this investigation.

Secondly, for the purpose of planning interventions, some psychotherapists I interviewed used specific versions of the explanatory models discussed before. In other words: Although there is pressure from practical goals from the context of application, there also is some pressure to *correctly represent* the disorder – arguably, for interventions to be possible on the basis of this model, they need to depict at least *some* actual causal relations.

Thus, this aim seems to pull in the direction of taking a somewhat *realistic* stance towards these models: Planful intervention in the world seems to require a somewhat realistic depiction of the mental processes that are to be intervened upon. I think that there are several ways of dealing with this pull towards a correct depiction of processes that are actually "out there" which secures my point that there still is some substantial influence on the form of these explanatory models that researchers in the discipline should be aware of.

Although it is true that there needs to be *some* correctness assumed for these models to be also used for the purpose of intervention, this still allows for a lot of wiggle room. That is, one may be skeptical about just how realistically these models need to represent the psychological processes that are relevant in the development and maintenance of mental disorders for them to be used this way. Although clinical psychologists surely *intend* to depict those causal processes that are “out there”, they may actually represent them only very crudely. As a matter of fact, several of my interviewees seemed to have a position somewhere along these lines. One fact that several of my interviewees referred to was that different treatments of mental disorders often have a similar effectiveness, even though they are based on wildly differing theoretical models of the disorder (e.g. Lambert & Vermeersch 2002). Furthermore, studies in clinical psychology indicate that more of the variance in outcome – exactly *how much* is unclear, with estimates varying between 30% and 70%, respectively (Imel & Wampold 2008, p. 255) – between different psychotherapists can be accounted for by more generic factors that are not specific to the treatment of interest, but that these are common to different therapeutic styles. These so-called “common factors” are factors like the therapeutic relationship or the therapist’s personality (Imel & Wampold 2008). From this, one may draw the skeptical conclusion that, really, the mechanisms of change operative within the psychotherapeutic process might not be represented by these models at all. This may suggest that these explanatory models of mental disorder are relatively far off when it comes to the correct representation of the relations between symptoms and underlying causes.

In the beginning of this chapter, I alluded to the fact that therapists often try to explain their patients’ disorders not only to them, but also to themselves. When the psychotherapist does the latter in an attempt to derive potential interventions from the model, she is not bound anymore by the patient’s individual narrative of his disorder. As we have already seen, there are some mental disorders – in particular those that implicitly carry negative evaluations of the individual’s personality – where the explanation that is offered to the patient and the explanation which the therapist gives to herself differ substantially. Nonetheless, in most cases, the main difference between the therapist’s own, “hidden” explanation of a patient’s disorder and the explanation presented to her client is merely on the *level of detail.*

One fascinating finding emerging from my interviews was the fact that, at least if my interviewees are to be believed, *epistemic injustice*, a concept introduced by Fricker (2007), may be less pronounced in this brand of psychotherapy than in other health professions (for an analysis of epistemic injustice in healthcare more generally, see Carel & Kidd 2014). Fricker distinguishes between hermeneutical injustice and testimonial injustice, and both may be operative here. *Hermeneutical injustice* occurs whenever someone's experience cannot be understood, either by themselves or by others, because there is a lack of concepts that adequately capture that experience. To give an example, it is plausible to think that the marital exemption for rape – the idea that rape can, by definition, not occur in marriage – has lead to hermeneutical injustice for many women in several countries (Fus 2006). *Testimonial injustice* occurs when someone's testimony is disbelieved or ignored because of the social group the individual belongs to (Fricker 2007).

For example, one interviewee explicitly stated that he actively tried to take seriously the patient's folk psychological understanding of their disorder, which, according to him, set him apart from many psychiatrists and other more medically-minded colleagues. According to him, these colleagues usually tried to impose their medical or psychiatric models of mental disorders on their patients without taking into account the patient's understanding of his disorder.

Furthermore, when asked whether she sometimes bent the truth when explaining mental disorders to patients, one therapist pointed out that she did not actually understand herself as an *expert* for the patient's condition. As she put it:

> "I am authentic there, because I think, this is nonsense, if I have another conceptualization than the patient. I also think, so for myself, I cannot go authentically into the relationship, because in that case, I act *as though I knew more than the patient. But I don't.* [...] So, this is his life. I cannot act as if I knew it better. Because, this also has something to do with power. So, I play the expert for something, that I do not know about, I do not live this life."

I find it particularly fascinating that this therapist framed the problem she saw when understanding herself as too much of an authority as an issue of *power*.

This seems to imply that epistemic injustice would be less pronounced in psychotherapy than in more medically-oriented treatments of these patients. I think this can be understood as a consequence of the much higher importance of the patient's experience of her life-world for psychotherapeutic treatments than for somatic or psychiatric treatments. One question to ask here is whether this makes clinical psychology, and psychotherapy, more specifically, more subjective, and thus, "less scientific" in a problematic way.

By contrast, some therapists claimed that they sometimes found it appropriate to not tell their patients the whole truth. They admitted to doing this in two cases: For one, when they knew that telling patients the whole truth about their mental disorder would probably result in the patient discontinuing therapy. For the other, this occurred when they were reasonably certain that presenting the patient's condition in a more positive light would lead to a faster improvement. Interestingly, when I afterwards asked other interviewees openly about this, most of them emphatically denied to do this. I take this to mean that there is a conflict between those values that most psychotherapists hold – being truthful and authentic in the therapeutic relationship, for example – and their practical aims that are intimately connected to the overarching aim of therapy: improving the patient's condition.

In this chapter, I have presented three aims that are operative in shaping explanatory practices in psychotherapy, namely, presenting the patient with possibilities for intervention, reassuring her of not being responsible for suffering from his disorder, and finally, attribution of agency. These three aims are based on taking two perspectives on the patient simultaneously: On the one hand, when the therapist points out possibilities for intervention, she does this on the basis of conceiving of her client as a patient who has certain harmful features – especially dysfunctional *beliefs* – that are in need of correction. On the other hand, when she presents the patient as not responsible for developing his mental disorder, she conceives of him as having good reasons for adopting particular harmful beliefs or having features that have *functions*. The attribution of agency, similarly, works via normalization and rationalization. This poses the question how exactly this dual representation of the patient as simultaneously (relatively) rational and normal as well as dysfunctional is possible. As I see it, this is a *conceptual* question that needs to be answered by carefully considering the concepts of (dys)functionality and (ir)rationality that are employed in psychotherapeutic explanatory practices. This is what I will do in the next two chapters, starting with the concept of rationality that is operative in psychotherapeutic practice.

Chapter 4

Concepts of (Ir)Rationality in Psychotherapeutic Practice

4.1 Introduction

One noteworthy feature of explanatory practices in CBT is that they represent the patient in two ways that appear to stand in tension to one another: that is, as (1) rational and relatively normal in developing dysfunctional beliefs in the first place (compare, e.g. Beck 1995, p. 15) and as (2) irrational in still holding these beliefs (which is the basis of therapeutic techniques such as cognitive disputation, as presented by Wittchen & Hoyer 2011, p. 555). Understanding the patient in both these ways simultaneously is central for achieving the aims of CBT, because it allows to represent the patient as having agency, while nonetheless identifying particular beliefs of hers as in need of change.

My thesis in this chapter is that, to make sense of phenomena like this one, we should disentangle two concepts of rationality. These are, firstly, the concept of *theoretical rationality* and, secondly, the concept of *pragmatic rationality.*

Even though CBT does not explicitly rely on the notion of rationality anymore, I take it that in therapeutic practice, psychotherapists *do* implicitly rely on judgements of rationality *and* irrationality. This is not trivial, though. In fact, several of my interviewees claimed that the notion of irrationality is not important for CBT and that therapists nowadays employ judgements of dysfunctionality instead. In the next section, I will argue that both notions are important for analyzing and understanding psychotherapeutic practice.

Thus, the notion of dysfunctionality *also* plays an important role in clinical psychology and psychotherapy. In the subsequent chapter, I will provide an analysis of the concept of *dysfunctionality* as used in these practices. Very roughly, my take on the respective roles of these two notions is that, while dysfunctionality allows us to understand what therapeutic interventions *target*, (ir)rationality allows us to understand how psychotherapeutic interventions, including explanatory practices and certain disputation techniques, *work*.[1]

Let me quickly say something on how this chapter relates to the earlier ones: In the first chapter, I identified the explanatory strategy that is at the heart of two models of mental disorders. I pointed out that it relies on assuming that someone with the disorder has particular beliefs that she once adopted for good reasons. She kept these beliefs, and at some point, they started causing harm. I expanded on this issue in chapter three. I claimed that *normalizing* and *rationalizing* a patient's experience and behavior are important parts of psychotherapy that serve certain practical aims. I argued that, for CBT to get off the ground at all, the patient needs to understand herself as *able* to change something about her symptoms. Representing the patient as both in some sense rational and in another sense irrational makes it possible to point out to her that she still has *agency*, and to convince her that it nonetheless makes sense to get rid of particular beliefs of hers.

Methodologically, what I do in this and the following chapter is partially based on *conceptual analysis* and partially on *conceptual engineering* (Plunkett & Cappelen forthcoming). It is akin to conceptual analysis insofar as I try to accommodate the intuitions of practitioners about the usage of their terms. Simultaneously, I intend to make my two accounts as simple, general and encompassing as possible. My method thus bears similarities to conceptual engineering, more precisely, the kind of conceptual engineering that deals with the improvement of existing concepts (Plunkett & Cappelen forthcoming, p. 3-4). More specifically, it is akin to Carnapian *explication* insofar that I take a given, more or less inexact concept and substitute it with a more exact one (Carnap 1959, p. 12). This latter concept should then comply better with theoretical virtues like coherence, simplicity, scope, and the like. Furthermore,

[1] The underlying observation is very similar to one that I already referred to in the last chapter by (Bolton 2008, p. 16-20), who writes that "psychology typically emphasizes that patterns of behaviour are learnt, and that dysfunction may arise when behaviours that are reasonable in the context in which they were originally acquired are applied in different contexts, but again the psychological processes involved are not qualitatively different from those operating in the normal case.", only that I add to this the observation that those beliefs that are described as dysfunctional (where Bolton speaks of "dysfunction") currently are also described as irrational by mental health professionals.

part of what I am doing in this particular chapter also exploits the method of rational reconstruction that I introduced earlier.

My reconstruction, while it should be coherent with as much of psychotherapeutic practice as possible, need not be compatible with every single usage of "(ir)rationality" by psychologists and psychotherapists. My analysis is to be understood as an attempt to provide concepts that are slightly more precise than those employed by practitioners.

Adopting this method is motivated by several observations: Firstly, the usage of these two terms has changed over time. Secondly, it differs between different therapeutic schools, for example, between CBT and RET, which Albert Ellis introduced towards the end of the 1950s (Ellis et al. 2010, p. 23)[2], and lastly, the usage of this term is not consistent among cognitive behavioral therapists or extremely broad, thus subsuming many things under it that should be kept apart. My analysis should both be consistent with large parts of the therapeutic practice and enlightening as a reconstruction of it. At the same time, my improved concepts must be able to do the same work that those old concepts carry out.

The analysis of actual therapeutic practice that I offer is based both on statements of my interviewees about the meanings of these terms and on the reconstruction of my two exemplary models from the beginning. Furthermore, I will use my own, pre-philosophical understanding of these terms as a resource. I generated this understanding when studying clinical psychology, and, more importantly, when working as an intern in psychiatric hospitals.

This chapter is structured as follows: In the second (and next) section, I will start by providing evidence for my earlier claim that therapists do, indeed, use judgements of rationality in therapy. Then, I will appeal to the practical syllogism as providing a basic model of an action explanation and I will argue that we should (to a first approximation) understand psychotherapeutic rationalizations of seemingly disturbed behaviors typical of mental disorders like OCD and depression along the lines of that model. I will use this scheme to argue that, in the course of the therapeutic process, one and the same behavior

[2] A full discussion of how "irrationality" is used in RET could probably fill several other books, which is why I will not go into the details of this here. It should suffice to say that the use of the term is quite different in RET than in CBT, and that RET seems to have engaged more with the philosophical foundations of the term, as becomes clear when considering the book of Ellis et al. (2010), for example.

can be described as rational in one sense and irrational in another.

Section three will take a closer look at two kinds of objects that are also involved in practices of normalization and rationalization in psychotherapy, namely emotions and reason-processing faculties. What is exploited in these practices are *appropriateness* and *normality*, respectively. Both notions have success conditions that differ in interesting ways from the conditions under which a belief or cognition counts as rational.

In section four, I will finally develop my two accounts of rationality for cognitions. My two notions are the notion of theoretical rationality and the notion of pragmatic rationality that require different forms of justification. The concept of theoretical rationality represents whether someone's beliefs are well grounded in the evidence available at a particular point in time. Another concept of pragmatic rationality represents whether an agent's beliefs are helpful in attaining her explicit goals. Furthermore, I take it that the two notions of rationality can be used to describe the same belief as rational at one point in time – usually, this is the time of belief formation – and irrational at a later point.

In section five, I will use the practical syllogism model of rationalization to provide a more fine-grained analysis of the various ways in which therapists structure their therapeutic practice around notions of rationality and irrationality. I will argue that my two notions of rationality allow for a more precise way of understanding the therapeutic practice of challenging specific harmful beliefs of patients.

4.2 Ascribing Rationality and Irrationality

Let me begin this section by making good on what I promised to do in the introduction: that is, to present reasons for my conviction that indeed, psychotherapeutic practice still implicitly relies on judgements of irrationality, over and above rationalizing and normalizing their patient's emotions and behavior. Furthermore, I will offer a few first hints on why the concept of irrationality should be kept apart from dysfunctionality. I will elaborate on this issue in the next chapter.

In her monograph on Cognitive Therapy, Judith Beck (1995, p. 3) says something that is quite instructive for someone who is interested in the role of *rationality* in psychotherapy:

> "In a nutshell, the cognitive model proposes that dysfunctional thinking (which influences the patient's mood and behavior) is common to all psychological disturbances. When people learn to evaluate their thinking in a more realistic and adaptive way, they experience improvement in their emotional state and in their behavior. [...] If you then examined the validity of this idea, you might conclude that you had overgeneralized and that, in fact, you actually do many things well. [...] For lasting improvement in patients' mood and behavior, cognitive therapists work at a deeper level of cognition: patients' basic beliefs about themselves, their world, and other people. Modification of their underlying dysfunctional beliefs produces more enduring change."

This is a first indicator shows that CT depends on interventions that take judgements of a thought or an inference's rationality or plausibility as their basis. Even though what therapists *do* tackle, according to Beck, are dysfunctional beliefs, these beliefs are tackled by pointing out to the patient that they are *unrealistic*.

To gain a very first grasp of what such a notion of rationality and irrationality might look like, consider the following characterization from an introduction to REBT:

> "To describe a belief as 'irrational' is to say that:
>
> 1. It blocks a person from achieving their goals, creates extreme emotions that persist and which distress and immobilise, and leads to behaviours that harm oneself, others, and one's life in general.
> 2. It distorts reality (it is a misinterpretation of what is happening and is not supported by the available evidence);
> 3. It contains illogical ways of evaluating oneself, others, and the world: demandingness, awfulising, discomfort-intolerance and people-rating."
>
> (Froggatt 2005, p. 2)

Clearly, many different things are run together in this characterization. For one, considerations about the *harmful effects* of these beliefs as well as considerations about *how reasonable* it is to hold them are part of this understanding of irrationality, but also, it *both* refers to the belief's correctness *and* to irrational ways of reasoning.

I take it that these should be distinguished from one another to allow for a more precise understanding of *both* what allegedly goes wrong in the patient and of how psychotherapy operates *exactly*: It should at least *in principle* be possible to be rational or have good reasons to act in a manner that harms oneself. But this is ruled out for *conceptual* reasons by this analysis. That is, it runs together the feature of counteracting an individual's goals and causing her harm – two things that do not always map onto each other. Just think of someone who has the explicit goal to earn a specific amount of money. To achieve this goal, she has to choose between a set amount of professions, none of which satisfies a deep-seated need that she may have. Let's say that she has the need to work with other people – i.e., in a helping profession. In this case, the agent's explicit goals and her needs are in conflict, such that acting on her rational belief that she needs to work in a certain profession actually causes her harm.

Furthermore, this distinction is important in understanding certain processes in the treatment of mental disorders. Someone who enters psychotherapy as a patient often perceive it to be necessary to reconsider her priorities and life goals. As I will argue later, in cases like these, the harm-inducing beliefs only *become* unreasonable or irrational once the patient adopts goals that are consistent with her well-being.[3]

Above, I offered some reasons to believe that *not only* does the notion of rationality underlie models of mental disorders and psychotherapeutic practice, as the strategy of *rationalization* indicated that I described in detail in the last chapter. Instead, judgements of irrationality *also* implicitly underlie important aspects of therapeutic practice. Furthermore, I pointed out that the notion of irrationality must be distinguished from dysfunctionality. In the subsequent section, I will ask how therapists do, in practice, *rationalize.* In doing so, I will ask what the *objects* of rationalization are. Furthermore, are an agent's actions *rationalized* within psychotherapy, and if so, how? And what about his beliefs and emotional reactions?

[3] It is a fascinating question whether aligning one's goals with one's well-being is actually *rational*, no matter what one's goals really are. Later in this dissertation, I will argue that this is actually the case for most patients in therapy, and that this fact is actually used in psychotherapy to bring about change.

As I have argued before, I take it that explanatory practices in psychotherapy usually aim to represent the patient's thoughts and behaviors as *relatively* normal and rational. Think of compulsions, which, at the outset, appear to many people as rather senseless behaviors. Even in the DSM-5, compulsions are described as "not connected in a realistic way with what they are designed to neutralize or prevent" (American Psychiatric Association 2013, p. 237). When understood according to the model of OCD (Salkovskis et al. 1998), even actions of this kind emerge as relatively rational, since patients have reasons for engaging in these behaviors. This suggests that there are two perspectives on this phenomenon in play here, and they are both important for psychotherapeutic practice.

My thesis is that, when psychotherapists explain mental disorders in psychotherapy, they suggest that the patient's actions are rational in the sense of instantiating a practical syllogism schema.

To see this, we might ask: Which conditions do we usually take an action to be rational? This is actually not *quite* the right question to start out with. As I have pointed out, I am interested in how this concept is utilized in therapeutic practices. In these practices, the behavior of a patient is often *rationalized*, that is, presented as the result of a perfectly rational instance of practical reasoning.

Thus, I will focus on the conditions under which an action counts as *rationalized*. That is I will present the conditions under which we, as laypeople, consider someone's actions to be reasonable, given the individual's beliefs and intentions. Alternatively, we may ask for the conditions under which we usually think that someone's action has been explained to us. To get a grasp on this matter, it is instructive to take a look at philosophical action theory. I take it that the process of rationalization of an individual's apparently nonsensical *actions* consists in creating a practical syllogism for an action akin to the following schema, taken from Nordenfelt (2007, p. 90):

(P_1) A intends to bring about G.
(P_2) A believes that he is in [context] C.
(P_3) A believes that he will not bring about G in C unless he performs F.
(P_4) A is capable of and unprevented from performing F in C.

This practical syllogism schema rationalizes the agent carrying out an action of type F. How do explanatory practices in psychotherapy instantiate this practical syllogism schema? Let me illustrate this with the imagined case of a psychotherapeutic patient, call him Pete, who suffers from OCD. Pete has intrusive thoughts about throwing other people from bridges and is convinced that he has to perform particular idiosyncratic rituals – let's say, multiplying the even numbers from 2 to 42 in his head – in order to make sure that he does not actually carry out this behavior. These rituals are a *maintaining factor* of the disorder, which makes them both important to *understand* and important to *challenge* them in CBT. When we consider the explanatory model of Salkovskis et al. (1998), the syllogism schema seems to be instantiated as follows:

> (P_1) Pete intends to bring it about that he does not throw anyone from a bridge.
>
> (P_2) Pete believes that he is in a context where the intrusion about throwing someone from a bridge increases the likelihood of him actually throwing someone from a bridge.
>
> (P_3) Pete believes that he will not bring it about that he does not throw someone from a bridge in a situation where he had an intrusion about throwing someone from a bridge, unless he multiplies the even numbers from 2 to 42 in his head.
>
> (P_4) Pete is capable and unprevented from multiplying the even numbers from 2 to 42 in his head.

It is instructive to conceptualize explanatory practices and some therapeutic strategies in CBT with the help of this schema. It demonstrates that these practices are actually akin to how we, as human beings, usually try to make sense of other people's behavior. That is, it shows the similarity of this kind of reasoning to folk-psychology (compare, e.g. Ravenscroft 2019). According to this schema, if we take specific background beliefs of the individual into account – in particular those that are referred to in premises two and three – his behavior is, in fact, rational. This is importantly different from how these individuals are often perceived by themselves or others at the outset: that is, as acting in weird, even crazy, ways.

As becomes clear when considering Salkovskis' recent model of OCD, psychologists and psychotherapists usually tackle both the belief that is represented in premise two and the belief represented in premise three. According to Salkovskis' model of OCD, the belief described in the second premise – that the individual is in a context that involves a high probability for him to harm himself or others – is incorrect. This belief can be challenged by pointing out that intrusive thoughts do not actually bear the meaning that the patient takes them to have. That is, they do not actually represent an intention of the agent, and therefore, they are *not* indicators of danger or harm. Since intrusive thoughts appear in healthy individuals as well with roughly the same content and frequency (compare Rachman & de Silva 1978), they *cannot* mean what the patient takes them to mean. The belief described in the third premise, in turn, is challenged by pointing out that the patient does not have to do anything to bring about a situation where there is comparatively little danger to himself or others, since *there is nothing in the initial situation that would have increased the danger in the first place.* Both of these are encoded in the explanatory model of OCD of Salkovskis et al. (1998) that I have discussed in detail in the first chapter.

In addition to rationalizing the agent's *actions* by presenting them as plausible consequences of the patient's intentions and particular relevant beliefs, therapists also partly rationalize and normalize her *beliefs* and *emotional reactions.* Very roughly, I take it that the agent's beliefs are *rationalized* by pointing out that they are *either* coherent with the best evidence or because they allow the client to reliably achieve particular, consciously held, goals of hers. Her beliefs can also be *normalized* by showing that they are produced by perfectly statistically normal or functional cognitive processes. In parallel to that, most emotions emerge as *plausible* reactions to the situation at hand. The point here is that the patient's particular appraisal of the situation would lead to this emotion in many, if not most, people. Take the person with OCD who believes that her intrusive thought is indicative of danger. It seems absolutely normal to react with anxiety to this interpretation. In fact, most people would probably react like this.[4] This leads to conceptualizing emotions as normal

[4] If we wanted to frame matters this way, we might say that not the individual's mental faculties – understood as dispositions – are in disarray, but her beliefs about certain facts of the matter. This also allows for the opposite case to occur. That is, think of someone who has emotions that are appropriate to a given situation, but possesses abnormal emotion-processing faculties. We might think that individuals who suffer from *Antisocial Personality Disorder* (see American Psychiatric Association 2013, p. 659) may, in some situations, fit this description.

and *understandable* when the patient's particular appraisal of the situation – sometimes brought about by idiosyncratic behavior, like focusing on particular features of a situation – is taken into account.[5]

At first glance, this understanding of rationalization seems to leave little room for irrationality to enter the picture. But certain actions of the individual are nonetheless conceived of as irrational by therapists and clinical psychologists.

Let me present some examples for how (ir)rationality is usually ascribed by psychotherapists or clinical psychologists. While these examples are *not* the outcomes of a literature review or of qualitative interviews, they are based on my experience as a student of psychology and on discussions with several psychotherapists, psychotherapists in training, and clinical psychologists. They serve as a first source in developing my two notions of (ir)rationality. We will later see that my two notions can, among other things, account very well for how cognitive disputation works, a psychotherapeutic method that relies on understanding patient's beliefs as (ir)rational.

In anxiety disorders, someone's safety behavior – at least if she knows about the relationship between safety behavior and the maintenance of her disorder – is, from the perspective of someone who has *all* of the currently existing knowledge about the disorder, best described as an *irrational* action. From her perspective, it results from assigning too much weight to the short-term goal of getting rid of the anxiety in a particular situation and too little weight to the long-term goal of eliminating the anxiety disorder. In other words: Given the patient's ordering of goals, and the perspective of someone who has the relevant knowledge, it does not make sense for her to engage in safety behaviors, since they run counter to the goal of getting rid of her disorder (e.g. Salkovskis 1991).[6]

[5] I take it that many rationalizations of actions in depression can also be understood as instances of this schema, even though this is a little bit harder to see. But consider the following rationalization of a depressive person's social withdrawal (which seems puzzling, at first) that I have been confronted with several times both as a student of psychology and as an intern in psychiatric hospitals:

(P_1) Anne intends to feel better.

(P_2) Anne believes that she feels bad because she is exhausted.

(P_3) Anne believes that she will not feel better in a context where she feels bad due to exhaustion, unless she (relaxes by) withdrawing socially.

(P_4) Anne is capable of and unprevented from withdrawing socially.

In this case, again, Anne's reaction to her alleged exhaustion by withdrawing socially makes perfect sense. In fact, there is an important distinction to be made between exhaustion and depression (or exhaustion due to a depressive syndrome and exhaustion due to external factors like working too much) which she is simply not aware of.

[6] One may very well worry about the question whether the problem is with the patient's conscious

Given that a patient usually neither knows that her intrusions are actually harmless nor about the relationship between neutralizing actions and the disorder's maintenance before it is pointed out to her, carrying out safety behavior does not qualify as irrational *before* the causal relation between neutralizing actions and the disorder's maintenance is explained to her. This is because the rationality and irrationality of specific beliefs is partially dependent upon the agent's background beliefs: Before she knows that neutralizing actions causally contribute to the maintenance of her disorder, it may be rational for her to engage in these behaviors. After all, what she desires most when confronted with an intrusive thought are two things: (1) to make sure that she and others are safe and (2) to reduce the feeling of anxiety (resulting from her false interpretation of the intrusion as signifying danger).[7]

Neutralizing actions do just that: They provide a sense of safety for the individual who carries them out. On this view, engaging in this action is rational, whether or not it actually prevents the feared event from occurring. Furthermore, the decrease of anxiety in response to neutralizing actions often makes sense: Washing one's hands *is* connected to getting rid of germs that the patient may be afraid of, for example. But even when the neutralizing behavior would not actually prevent the feared event, it still seems reasonable from the perspective of the patient to carry out this particular action, simply because it decreases the intensity of an unpleasant emotion, in this case, fear. Before the agent knows that this action actually helps to *maintain* her disorder, it seems perfectly rational for her to carry it out. This is because the long-term goal of getting rid of her OCD and the short-term goal of reducing her anxiety are not, to her mind, in conflict. Clearly, the kind of rationality at play here is not concerned so much with whether the belief is well-grounded in the available evidence. Rather, it asks whether it makes sense for the patient to assume that this belief helps her to accomplish her explicit, high-level goals.

By contrast, once the patient knows about the relationship between neutralizing actions and her disorder's maintenance and nonetheless carries these ac-

ordering of different goals – that is, whether that ordering is irrational – or with the patient not acting in accordance to her actually *rational* goal-ordering. I will simply assume the latter for the sake of simplicity here.

[7] They can come apart when the individual actually *knows* that carrying out particular behaviors does not really have any effect on the danger that may be present in a situation, but carries out the behavior in question nonetheless, since it reduces the anxiety that she feels. In that case, she does not have to do anything to ensure the safety of everyone, but she does it anyway. A typical example of this is the obsessive need to perform certain calculations.

tions out, she acts irrationally. This is because getting rid of her mental disorder, which presupposes not carrying out neutralizing behaviors, would result in sparing herself of a lot of future anxiety and suffering. By contrast, carrying out neutralizing behaviors will only get rid of her anxiety in that very moment, while effectively causing future anxiety and taking up enormous resources.[8]

Above, I have presented some cases in which someone's action, emotion, or behavior is described as (ir)rational. With these cases in mind, we are well equipped to consider in more detail several *objects* that are regularly identified as (ir)rational in psychotherapy. I would like to start with those objects that are less central for psychotherapy, that is, reasoning processes and emotional reactions, to then work my way towards an analysis of how rationality is ascribed to beliefs. The (ir)rationality of beliefs underlies judgements of actions as (ir)rational.

The two notions of rationality that I introduce in the following do not require the belief in question to be *correct*. By contrast, I understand psychotherapists as more concerned with the – empirical or pragmatic – *justification* of beliefs at particular points in time and the coherence of these beliefs with the agent's more general belief system. Although *many* beliefs that are rational in either of those ways that I will delineate are also incorrect, not *all* of them are.[9] It is an important part of my analysis that, on the psychotherapist's understanding, beliefs can be rational to hold by being well-justified, although they are false.[10]

As I already stated, I want to present a reconstruction of certain parts of therapeutic practice, based on two different notions of (ir)rationality. These parts of therapeutic practice include rationalization as well as normalization, practices that occur in explaining mental disorders to patients, and practices that challenge the rationality of patients' beliefs. Now, rationalization of the patient's underlying beliefs often requires that her reasoning processes and

[8] Of course, this builds on the assumption that the individual has reason not to discount future events very steeply (Frederick et al. 2002, compare, e.g.). We can see that this makes a difference concerning what is considered to be rational or useful in the therapeutic treatment of terminally ill patients, for example.

[9] This matter actually gets more complicated since I deal with evaluative and normative statements as well and would very much like my account to be neutral on the question of the metaethical status of those statements. It should not matter too much, I take it, whether these statements *do*, in fact, have truth-values.

[10] Although a thorough discussion of this is beyond the scope of this investigation, if taken literally, this would seem to commit psychotherapists to *internalism*, that is, the view that a belief's justification is determined by the internal states or reasons of an individual (compare, e.g. Poston 2019).

emotional reactions are *normalized* or *de-pathologized* first. Plausibly, having relatively rational beliefs presupposes having reasoning faculties that work relatively normally. Furthermore, patients are often concerned about their emotional reactions to events. By showing that (and how) these can be made sense of, therapists present their patient's symptoms and psychological processes as understandable, and not as a sign of potential "craziness".

4.3 De-pathologizing Reasoning Processes and Emotions

Let me start with how reasoning processes are normalized in psychotherapy.

Instead of explicitly calling human reasoning processes "rational" or "irrational", clinical psychologists are more apt to describe them as (1) (a)normal or (2) (dys)functional, as became clear in my qualitative interviews.[11] In individuals with mental disorders, particular reasoning processes are systematically and significantly biased. In those models of mental disorders that I have presented, reasoning biases are mentioned only when they *actively contribute* to the maintenance of the mental disorder in question. For example, in OCD, misinterpreting intrusions as implying danger causes certain biases in attention and reasoning, such as focusing one's attention on information that might signal danger. In doing so, these reasoning processes also behave statistically *abnormally*. These kinds of reasoning processes are then singled out by mental health professionals and, often, the correctness of the resulting thought(s) is challenged. But, on their account, the individual's reasoning is not *inherently* biased, but the bias hangs together causally with particular dysfunctional beliefs and dysfunctional behavioral patterns of the individual.[12] According to *classical* CBT, once the individual's beliefs are corrected, their reasoning will

[11] I will refer to this again in the next chapter, but, as one of my interviewees put it: "'Dysfunctionality' has replaced 'irrationality', because people [mental health professionals] wanted to get away from telling people [patients]: 'You are not rational, you are basically crazy if you are thinking like this."'

[12] Several newer psychotherapeutic treatments like *Metacognitive Therapy* (*MCT*) differ from classical CBT here in focusing more strongly on the dysfunctionality of attentional and reasoning processes in these disorders (Wells 2011). Thus, therapists of this orientation often seem to suggest that the patient's thoughts and beliefs are not the source of her problems (and that it may be relatively normal to have thoughts with a particular content about the self), but that it is the patient's way of reasoning and interpreting these thoughts which are problematic. This also leads them to intervene on the latter. Sadly, a comparison of classical CBT and other therapeutic orientations like MCT is beyond the scope of this dissertation. But let me note that I am certain that my concepts of (ir)rationality and dysfunctionality could in principle be adapted to apply to MCT as well – even though this would require a more detailed analysis of reasoning processes than I provide here.

return to normality.

When rationality is conceptualized as relative to the standards of logic and probability theory (compare Tversky & Kahneman 1974), the majority of human beings must be understood as having *some* irrational beliefs, as reasoning in an irrational manner or as acting irrationally. Often, these individuals feel better doing so than they would if they reasoned in a perfectly rational manner.[13] If this is the case, the clinical psychologist will not assign too much weight to this kind of irrationality. Instead, he will point out that the subject's reasoning processes are *normal* in the sense of being skewed in the same way in most other human beings. That is, the therapist will point to the *statistical* normalcy of the individual's reasoning processes. One observation making it clear that clinical psychology is more interested in statistical normalcy and harmfulness of cognitive processes than in whether they fit the actual facts of the matter concerns how Beck's way of talking about the negative bias in depression changed over time. In his earlier work (compare, e.g. Beck 1967), he seems to suggest that the processing of individuals with depression is skewed in the sense of systematically leading to false beliefs. This changed – probably with the observation of phenomena like *depressive realism* (for a meta-analysis of studies on the phenomenon, see Moore & Fresco 2012) –, and, in recent articles, he talks about the patient's negatively skewed reasoning processes differently, that is, without referring to the belief's correctness (compare, e.g. Beck & Bredemeier 2016). This description of the patient's reasoning processes as relatively statistically normal serves the function to de-pathologize.

Thus, when reasoning processes are concerned, psychotherapists are usually more interested in *normalizing* than in rationalizing.[14] Usually, those processes that are of interest to the clinical psychologist and psychotherapists do not differ from normal cognitive processing by being more inaccurate, but by

[13] A rather obvious example of this is the so-called "self-serving bias", a bias that consist in interpreting successes as due to one's own doings, while interpreting failures as due to external factors (compare, e.g. Campbell & Sedikides 1999). This bias tends to lead to more positive emotions in individuals than negative emotions, which is why psychotherapists are not interested in challenging them.

[14] To be able to see this, remember the following quote from one of my qualitative interviews that I already referred to in the last chapter: "We are all in the same boat, what you are having, that's not your mistake, your deficit. You're not crazy, [...] in your case, something has run out of control, you usually have too much of something, and too little of something else. And this dysbalance, it causes you to suffer. This produces tension, pressure, weird experience." The vocabulary which this psychotherapist uses in describing his explanations clearly aims at making the patient feel more normal, but not necessarily more rational.

having harmful effects (compare, for example, the description of the thinking disorder in depression by Beck 1967, p. 255-269). But I take it that, especially in classical CT, such problematic reasoning processes are usually taken to be the result of problematic core beliefs.

Now that we have seen how therapists deal with abnormal reasoning processes in therapy, what about the patient's *emotions*? As I see it, therapists are much less concerned with the (ir)rationality of emotional reactions and more with their *appropriateness* to a given situation. I take it that, when a therapist appears to be concerned with the (ir)rationality of emotions, she is really interested in the (ir)rationality of the *beliefs* that bring these emotions about.

This of course raises the question whether, in reality, practitioners of CBT might actually be concerned with the (ir)rationality of emotions, but take emotions to have cognitive content as well. But even if emotions are understood as having (at least) a cognitive and a feeling component (Scarantino & de Sousa 2018), the point of interest here is that the emotion becomes understandable *only* once the cognitive component – and the reasons that the individual has for taking it to be true – is made explicit by the therapist. In other words, this issue is not really important for the point I want to make, and I would like to remain neutral on questions of what emotions really are.[15]

Even if those emotions seem irrational at first glance, as soon as the individual's background of beliefs is taken into account, they usually make sense, because they are appropriate to the situation *that the person sees herself in*. To give an example: Think of Nick, who reacts to giving a presentation in a seminar with intense feelings of anxiety. A fellow student, sitting in the audience, might notice his anxiety and classify it as irrational, given the fact that nothing really hinges on Nick's performance. On the classmate's reasoning, if Nick gives a bad presentation, the PhD candidate holding the seminar might be annoyed, the class will probably learn nothing new, but this does not really strike him as something worth worrying about. In psychotherapy, Nick's emotional reaction might be reconstructed as caused by his belief that he will not meet the standards of the other participants in the seminar and, as a result, be socially excluded (in keeping with the cognitive-behavioral model of SAD, compare Heimberg et al. 2010). He has adopted this belief in school, where

[15] I might nonetheless sometimes use the term "irrational emotion" or the like without intending to commit myself to the view that emotions are irrational.

he was socially excluded and bullied for saying "stupid things" and acting insecurely, for example, stammering when speaking in class and when giving presentations. Since belonging to a social group is one of his top-level explicit goals, his emotional reaction becomes understandable – even though not appropriate to the actual situation. This is because, to *his* mind, the situation he finds himself in involves actual threat. This holds especially since he has good reasons for his conviction that people will react in the manner in question.

Thus, cognitive-behavioral therapists would reduce the apparent irrationality of Nick's emotional reactions to them being brought about by a judgement about the current situation that derives from at least one belief of his that was, at one point in time, relatively well-justified, although this belief might not be well-justified anymore.

The emotional reactions of individuals with mental disorders are, at least in CBT, usually understood from a perspective that makes them emerge as understandable, given the individual's dispositions, background beliefs and her (perhaps pre-conscious) judgements about the situation at hand.[16] We can also see this when considering the two exemplary models of MDD and OCD, respectively. In the latter, the individual's anxiety emerges as plausible, because she appraises the situation as one where actual danger is present. In the former, the individual's extreme sadness and emptiness emerge as understandable, because the individual has a myriad of negative beliefs about herself, the future and the world. But in another sense, these emotions nonetheless are inappropriate, since the beliefs they rely on are actually *false*.

A moment's reflection shows that this is actually built into the framework of this form of therapy: Since CBT assumes that emotional reactions are largely *dependent* on the cognitive evaluation of the situation (see, e.g. Beck 1995), the appropriateness of an emotion to a situation will always hinge on the *correctness* of those thoughts of the individual that cause them. That is, while we can *understand* Nick's anxiety when he needs to give a presentation in a seminar, this does not make his anxiety *appropriate* to the situation, since

[16] Put quite generally, CBT – at least in the original formulation that was presented by Aaron Beck – appears to assume that for every disposition to interpret some stimulus in a particular manner, there is an underlying belief that may be conscious as well as pre-conscious. This makes sense, as CBT is committed to the idea of *schemas*, that is, particular cognitive structures of the individual that serve to appraise incoming stimuli and information (Beck 1964, p. 562-563, my italics). Importantly, schemas always appear to have content that may be formulated in terms of an individual's deep-seated beliefs. I introduced the concept of schemas already in Chapter one.

– if understood as a danger-detection signal – it misrepresents an actually harmless situation as dangerous. Thereby, suffering from certain emotions that one experiences is reduced to having particular problematic beliefs.

The resulting view of individuals with mental disorders is that, while her mental (that is, cognitive and emotion-processing) faculties are largely in order, her mistake consists in holding on to particular beliefs for too long despite contradictory evidence. Now, cognitive mistakes are something that we as humans are used to dealing with – arguably, we make such mistakes all the time (compare, e.g. Tversky & Kahneman 1974). Cognitive biases that might lead to so-called "belief perseverance" (see, e.g. Guenther & Alicke 2008) are actually quite widespread (we might think, for example, of the confirmation bias, see, e.g. Plous 1993). Thus, making such mistakes in reasoning seems relatively normal. Presenting the patient's emotional and reasoning processes in this way thus reduces what at the outset appears to be a quite substantial problem to a relatively minor one. It simultaneously shows that the problem that is at the core of the mental disorder is one that can be dealt with by the means of intervention, and it shows that the agent cannot reasonably be blamed for either her thoughts or emotions.

To sum up: In therapy, an individual's emotional reactions are usually presented in a way that makes them appear understandable. Emotions *do not* emerge as irrational, but only as inappropriate to a given situation. If they are inappropriate, this is merely in virtue of those beliefs that bring them about. This understanding of the interplay between emotions, reasoning faculties and underlying beliefs is important, because it lends further support to the importance of rationalizing the patient's beliefs in therapy: If the patient's emotions are plausible reactions to these beliefs, then pointing out that the patient is relatively rational in holding these beliefs – even if they may actually be false, like Nick's beliefs – not only de-pathologizes the patient's thoughts, but also her emotions. It is thus time to finally deal with how (ir)rationality is ascribed to beliefs in psychotherapy.

4.4 Rational and Irrational Cognitions

Above, I described the puzzling situation that, when their mental disorders are explained to patients in psychotherapy, they are often represented as simultaneously (1) *having good reasons for forming* particular beliefs or cognitions that may seem "crazy" or irrational at first and (2) being irrational in holding those same beliefs or acting on them. What I will try to do, in this section, is to understand the different senses of "(ir)rationality" that are at play here. My aim will be to offer two understandings of (ir)rationality that solve this puzzle.

Cognitions include both thoughts and beliefs, and I think of both of them as in principle introspectively available to the agent. Thoughts are relatively non-stable cognitive phenomena that have propositional content – or whose content can easily be given propositional form, like intrusions – that might be present in the individual's mind at one point in time, but absent in the next. By contrast, I take beliefs to be more stable mental entities that are held by someone over a longer period of time and that are implicitly held to be true (compare, e.g. Schwitzgebel 2019). In CT and later in CBT, beliefs are understood as mental entities with propositional content that the agent – at least implicitly – takes to be true and that guide the agent's reasoning, thoughts, and actions (compare, e.g. Beck 1995, p. 14-18). Thoughts, on the other hand, may sometimes be held to be true by the agent, but they may also just occur to her without her actually believing their content to be the case. For example, intrusive thoughts belong into this category. As I have pointed out in the first chapter, these thoughts often misrepresent the state of the world, and most individuals also do not take them to be true. One exception are, as we have seen, individuals who suffer from OCD (compare, e.g. Salkovskis 1999)

In the following, I will present those conditions under which *beliefs* seem to be thought of as (ir)rational by psychotherapists. I will be less interested in thoughts, since therapists normally focus on challenging more deep-seated beliefs by – at least implicitly – pointing out that they are irrational.[17] Since I assume that different criteria are at play for descriptive and prescriptive cognitions, I will tackle them one after the other, starting with the case of descriptive beliefs.

[17] Even if they *do* focus on changing particular thoughts, the same conditions of (ir)rationality apply.

4.4.1 Descriptive Cognitions

Under which conditions do therapists ascribe rationality to a patient's descriptive beliefs? In more precise terminology, I am concerned with the rationality of a subject's *believing that* p *at a point in time* t, where p is a proposition. I think that believing at time t that a proposition with descriptive content p is true can be rational in at least two different senses. For one, it can be rational to believe something for someone because the evidence that is available to that person speaks in favor of that particular belief. To put it slightly more precisely:

(TR_{dc}) An agent a is theoretically rational in believing that p at t because p coheres with the relevant background beliefs of a and there is better (empirical) evidence for p available to a than for either member of a set of relevant alternatives $p_1, ..., p_n$.

Furthermore, someone can be rational in holding a particular proposition to be true for very different reasons. That is, she can be rational in having a particular belief because it is reasonable to believe that this belief does her well. To put it more precisely:

(PR_{dc}) An agent a is pragmatically rational in believing that p at t because at t, p coheres with the relevant background beliefs of a and there are good reasons for assuming that acting as though p were true will yield better results relative to the agent's explicit goals than acting as though either member of a set of relevant alternatives $p_1, ..., p_m$ were true.

I take these two characterization of (ir)rationality to be good first passes at those notions that are relevant for therapeutic practice. As I have said, psychotherapists tend to represent their patients' beliefs as simultaneously relatively rational and irrational. Usually, their idea is that the patients' beliefs were rational at the point of adoption, but became dysfunctional – and very often also irrational – later. This feature makes it necessary to introduce a time specifier.

Several conversations with therapists and therapists in training convinced me that these two characterizations are roughly correct. In these conversations, the individuals in question pointed out that, when the empirical evidence on the correctness of a belief was impartial, they often pursued the strategy of asking the patient whether a certain belief – think of "Other people don't like me" – was *helpful* or *useful* to hold. This fits nicely to a cognitive-behavioral therapeutic technique, so-called "hedonistic disputation" (Wittchen & Hoyer 2011, p 555). I will expand on this point later in this chapter.

The first kind of rationality may be called *theoretical rationality*, since it concerns the fit between the belief in a certain proposition and the evidence (that is available to the patient) for the correctness of that proposition at this particular point in time. Note that this notion of rationality allows to make sense of different attributions of rationality depending on whether the person knows that these beliefs will serve to maintain her disorder. By speaking of "relevant" beliefs, it also allows for a certain amount of cognitive *partitioning.* Since most people will hold *some* contradictory beliefs, we should not count *all* beliefs of an agent to be relevant for such judgements of rationality of particular propositions.[18]

The second kind of rationality may be called *pragmatic rationality*, as it concerns the fit between the belief that p and the agent's *explicit* or *considered* goals. In other words, it is pragmatically rational to believe that p whenever believing that p is the case fosters the achievement of the agent's high-level goals.

In his book on rationality and compulsion, Nordenfelt (2007) suggests an analysis of "having good reasons for an intention" that bears some similarity to this notion of pragmatic rationality: According to the author, an agent a has good reasons for intending to do x whenever doing x "fits well into the agent's general life-plan, [...] if it supports or at least does not interfere with the other top-level wants that the man has. I will call this variety of good reasons the *coherence sense.*" (p. 25 Nordenfelt 2007, my italics).

Let us now look at one example of how psychotherapists ascribe (ir)rationality to beliefs in more detail. Consider the following case: Someone, call him Pete, has a belief with descriptive content to the effect that he is not as good at his job to get the position he wants. Let us assume that this belief was theoretically rational to hold at a the time of belief formation, when his boss systematically devalued his job performance, and that he formed it by using

[18] It might even allow to make sense of the distinction between *active* cognitive schemas, as only those schemas that are active will also produce *relevant* beliefs or thoughts.

his reasoning faculties in a statistically normal way. This belief may be theoretically irrational at the current point in time, since the evidence available to the agent shows – and has shown for some time – that he is actually doing good work. he is often praised for what he does, but interprets this as his boss merely feeling sorry for him. Furthermore, let's say that the belief is pragmatically irrational, since it stands in tension with the agent's conscious goal to feel good about himself, and even with his goal to actually get this position.

The agent's reasoning mistake thus merely is that he has failed to update his belief when contradicting evidence became available to him. As I already pointed out, even this process of failing to update his beliefs in the face of contradictory evidence is usually framed by psychotherapists in a particular manner: Often, the persistence of old beliefs that were justified at one point in time, when currently, their opposite is better warranted by the evidence (so-called "belief perseverance", see, e.g. Guenther & Alicke 2008) is represented as statistically *normal* functioning of the agent's reasoning faculties (e.g. Nestler 2010). Representing the agent's mental faculties as normal often makes patients feel better, because it reduces their perceived blameworthiness for erroneously holding onto this belief. His mistake then becomes that he has not overcome the effects of a normal and pre-conscious mechanism by making use of "proper" reasoning – and this can hardly be expected from anyone.

This might look relatively similar to a distinction at the heart of a psychological debate in the area of research on judgement and decision-making that has been called "the rationality wars" (Samuels et al. 2002): After all, the opposition in the case above becomes one between, firstly, a strategy that arguably relies on evolutionarily useful traits of individuals and, secondly, a strategy that would theoretically be optimal.

There are two understandings of rationality that are of relevance for the "rationality wars": According to one tradition, human beings reason rationally if they tend to draw the right conclusions on the basis of the evidence available to them, that is, if their reasoning processes comply with the rules of logic and probability theory (Tversky & Kahneman 1974). According to another tradition, human beings reason rationally if they draw conclusions that make them, by and large, well adapted to their environment (e.g. Gigerenzer & Selten 2002). But as a concept of *rationality*, this does not play a big role in explanations of mental disorders. Instead, the adaptiveness of a particular behavior or belief plays a role in judgements of *functionality*, which I will talk about in the next chapter. Thus, let us set this aside for now.

One interesting fact about those descriptive beliefs that psychotherapists tackle in therapy is that, even though these beliefs have descriptive content, they are often interpreted by the individual as having normative or evaluative *implications*. Just take the well-known “Other people will never like me”. This belief arguably has normative consequences for whether the person perceives herself as a good person. I think that this is actually one of the reasons why they are interesting for psychotherapy in the first place. The other reason is that descriptive beliefs may have harmful behavioral consequences – just think of the descriptive belief that intrusive thoughts are indicative of unconscious wishes. I take it that usually, descriptive beliefs become problematic and interesting for psychotherapy, because they indicate that a particular need of the patient cannot be satisfied. For example, think of someone who takes “I am incompetent.” to be true. This belief is problematic for the individual only if the individual actually has a need for being or feeling competent. This issue will be tackled in more detail in the next chapter.

For now, I would like to leave this account of rationality of *descriptive* propositions as it is and turn to the more complicated issue of *prescriptive* cognitions.

4.4.2 Evaluative Beliefs

How is the rationality of propositions with *prescriptive* content assessed? And what do I mean by “prescriptive cognitions” in the first place?

What I mean here are beliefs that, implicitly or explicitly, refer to norms and that cannot easily be described in naturalized vocabulary. Evaluative beliefs are those beliefs that contain a value-judgement. Many beliefs that are at the core of mental disorders do not have descriptive, but also *normative* content.

Even more, those beliefs that are disputed in psychotherapy often contain *thick* normative-evaluative concepts, that is, concepts that *both* have evaluative and descriptive components (Williams 2006, p. 143-144). To give some examples, I think mostly of beliefs like “I am a loser”, “I am not worthy of love” or the like. The statement “I’m a loser” does two things: for one, it ascribes specific features to the own person. For the other, it contains a value-judgement. In the case at hand, the descriptive component arguably is something like not being able to finish a project at work, while the normative component consists in being evaluated negatively on this basis.

Arguably, these beliefs are both hard to justify and hard to challenge by relying only on descriptive statements about the state of the world.[19]

In my experience, psychotherapists only rarely refer to descriptive facts when challenging evaluative propositions (this also becomes clear when considering how psychotherapists challenge different kinds of beliefs, compare, e.g. Wittchen & Hoyer 2011, p. 555). While beliefs that only have descriptive content are vulnerable to the charge of not being based on sufficient empirical evidence about facts in the outside world, the same cannot be said for beliefs with non-descriptive content. Take someone's belief of not being worthy of love: Even if this person is *actually* loved by someone else, this will not count as evidence against this proposition. Instead, when arguing against these propositions in psychotherapy, seem to rather rely on their own normative beliefs, according to which too strict or too negative normative beliefs should be abandoned. For example, what was mentioned by several of my interviewees were normative beliefs invoking either universal quantification or specific modal operators to the effect that the agent "should", "need", or "must" do something.

Plausibly, what counts as evidence for or against the truth of the proposition will depend crucially upon the agent's other prescriptive beliefs. To assume that the correctness of propositions with prescriptive content could be determined by solely taking into account descriptive propositions would be to commit the naturalistic fallacy (Moore 1988, p. 38).[20] This maps well on certain therapeutic techniques for challenging beliefs, as we will see in the following.

The therapist's goal, I take it, in utilizing these therapeutic techniques will be to show that beliefs like the evaluative "I am a horrible person" or "I always have to please other people" impede the individual in living up to her goals.

How do psychotherapists actually reason about evaluative beliefs? Firstly, the

[19] Even though I realize that this is a contested issue among philosophers who might claim that the naturalistic fallacy (Moore 1988, p. 38) is not really a fallacy at all, I would also like to note that it is not really important to me whether normative or evaluative facts can *in principle* be derived from natural facts. What is important for me is what the actual practice in psychotherapy looks like.

[20] Although there is much discussion about the question whether the naturalistic fallacy is, in fact, a fallacy (see, for example Tanner 2006), I will go with the received view on this matter in assuming that it is, indeed, a fallacy.

individual's credence in a given proposition is weakened, or they show that the proposition is inconsistent with several other beliefs of the agent. This is what so-called "normative disputation" (Wittchen & Hoyer 2011, p. 555) does, a cognitive technique that I will discuss in more detail later. As a second step, the therapist shows that the belief is inconsistent with the patient's goals. This two-fold strategy is intended to make patients question their dysfunctional convictions and instead adopt new, healthier ones.

For propositions p^* with prescriptive content along the lines of "I should always please other people.", the criteria for pragmatic rationality are the same as the criteria for beliefs with descriptive content. In this case as well, it is pragmatically rational for the agent to believe something if there are good reasons for assuming that acting in accordance with that very proposition actually helps the agent in pursuing her explicit goals. That is, we get the following result:

(PR_{pc}) An agent a is pragmatically rational in believing that p at t because at t, p coheres with the relevant background beliefs of a and there are good reasons for assuming that acting as though p were true will yield better results relative to the agent's explicit goals than acting as though either member of a set of relevant alternatives $p_1, ..., p_m$ were true.

A particularly simple case of this is one where the agent has "acting morally" as one of her conscious, top-level goals. Besides not being goal-conducive in the sense of helping the patient achieve his top-level goals, such beliefs may also be irrational in a sense more akin to theoretical rationality for descriptive beliefs: That is, they may be irrational in the sense of not cohering with those normative standards that the agent is implicitly or explicitly committed to.

(TR_{pc}) An agent a is theoretically rational in believing that p at t because p coheres with the relevant background beliefs of a when compared with the members of a set of relevant alternatives $p_1, ..., p_q$

I think that this is the sense of irrationality that psychotherapists employ when engaging in normative disputation: One of the key techniques of this method is to point out to the patient that she devalues certain actions or features of herself that she would hold to be unproblematic when observed in other people. In my understanding, what therapists are getting at when asking how other people would be evaluated by the patient are those norms that the individual actually holds to be generally correct – after all, these are applied to everyone else but the agent herself. By pointing this out, I have already hinted at something that I will deal with now: That is, I believe that my two concepts of (ir)rationality can help us make sense of particular aspects of therapeutic practice.

4.5 Relevance for Therapeutic Practice

How may my two notions of rationality help us make sense of therapeutic practice? Let me start by pointing out the main aims of psychotherapy.

Arguably, the – implicit or explicit – aim of psychotherapeutic interventions is to make the patient think and behave in a less harmful manner than before. That is, psychotherapists intend to assist their patients in changing their actions and thought patterns such that they stop maintaining her disorder. I believe that therapists do so by first convincing the patient that her subjective health and well-being should be among her explicit goals, if it is not one of those already. Once this has been achieved, the therapists intends to make the patient understand that she does not *currently* act in accordance with that particular goal. The therapist does so by drawing both on cognitive techniques – for example, disputation techniques that attack either descriptive or normative beliefs – and on behavioral techniques – for example, trying out alternative behaviors. This is also how I take dysfunctionality and pragmatic rationality to hang together: Once the patient makes her well-being and mental health one of her most important goals, her dysfunctional beliefs *automatically* become pragmatically irrational – they induce significant harm in the patient, after all. Since these beliefs are usually both theoretically *and* pragmatically irrational – at least at the current point in time, it makes sense for the patient to replace them with beliefs that are either pragmatically or theoretically *more* rational. Ideally, the new beliefs will be both.

I take it that it is usually rather easy for psychotherapists to convince their patients to assign their subjective well-being the status of one of their top-level aims, since attending psychotherapy *at all* presupposes that the individual considers her mental health a priority. In fact, the process of convincing a patient to make her mental health a high-level goal may itself be best understood as based on considerations of pragmatic rationality: That is, for almost *any* set of high-level goals that patients have, a certain level of mental health is necessary. In addition to that, patients are usually in therapy because they are unable to attain particular explicit goals of theirs.[21] Thus, adopting the belief "My mental health should be among my top-level goals" appears to be pragmatically rational for almost every patient.

Thus, after normalizing and rationalizing a large part of the patient's beliefs – and building upon this, the patient's emotions and actions as well – the psychotherapists follows either of two strategies:

1. Showing how at least some of the patient's beliefs (namely, the dysfunctional, harm-inducing ones) are not conducive to the patient's considered goals *and* currently not supported by the available empirical evidence on that matter or incompatible with her background beliefs.

2. Showing how at least some of the patient's beliefs (namely, the dysfunctional ones) are currently not supported by the best evidence on that matter *and* not conducive to the top-level aims that the patient *should* have (among other things, being mentally healthy).

How exactly do therapists implement this, and how does it relate to the two notions of (ir)rationality that I have presented above? In particular, how does the distinction between theoretical rationality and pragmatic rationality help us in understanding these practices?

As I have hinted at above, the two different types of (ir)rationality are relevant for understanding the psychotherapeutic method of so-called "cognitive disputation" (Wittchen & Hoyer 2011, p. 555). Disputation techniques focus on challenging those beliefs of an individual that are understood as *harmful*, *maladaptive* or *dysfunctional.* But in challenging these beliefs, therapists *do*, I think, implicitly or explicitly rely on judgements of *rationality.* How so? Let me explain. As a matter of fact, CBT incorporates several forms of disputation. These are *empirical disputation*, *logical disputation*, *normative disputation* and *hedonistic disputation* (the following characterization of the four is based on Wittchen & Hoyer 2011, p. 555).

[21] Whenever this is not the case, the patient still is in therapy because he or she is *suffering* in a particular way, either as a direct or as an indirect consequence of their disorder. That is, their actions show that they *do* have a particular implicit aim that only needs to be made explicit: for the suffering to end.

In *empirical disputation*, a therapist challenges her patient's belief by assessing whether it is empirically accurate, that is, fits with what are plausibly conceived of as the empirical facts about the matter at hand.

In *logical disputation*, the therapist challenges her patient's dysfunctional beliefs, arguing from contradictions or tensions of a certain dysfunctional belief with the rest of the patient's belief-system. As we have seen before, the therapist will search for relatively specific kinds of contradictions. Normally, psychotherapists do not worry too much about the patient having *some* contradictory beliefs, as long as they are not detrimental to the patient's health and well-being. But once there are blatant contradictions between the patient's other beliefs and a particular dysfunctional cognition – or even obvious mistakes in reasoning –, the therapist will use this as evidence to argue against the belief in question.

These two techniques, I take it, rely on a notion of theoretical rationality along the lines presented above. I think that empirical disputation and logical disputation exploit different components of my notion of theoretical rationality. That is, empirical disputation exploits the fit with the empirical facts of the matter. Logical disputation, by contrast, exploits the idea of coherence with the background beliefs in question. Thereby, the latter technique may be used to challenge both descriptive and prescriptive beliefs.

I do not understand the distinction between these two kinds of disputation techniques as a sharp distinction but as one that primarily has *heuristic* value. It should thus be understood as a distinction that singles out which aspect of rationality is *most important* for challenging a particular belief. While empirical disputation *also* needs to take into account some of the patient's background beliefs – such as the patient's beliefs about what constitutes good evidence, but also other beliefs of hers. Logical disputation also sometimes relies on empirical evidence. An interesting instance of logical disputation occurs when a therapist challenges her patient's dysfunctional belief – think of the patient's belief that he is not lovable, because he is not good at his job – by pointing out that he *actually* judges other people to be lovable on the basis of a very different set of criteria, suggesting that he does not, generally, think that people are only lovable to the extent to which they perform well at their jobs (compare Wittchen & Hoyer 2011, p. 555).

What happens here is this: Firstly, on the basis of a set of background beliefs that therapist and patient share, the patient's (supposedly) *actual* moral attitudes and beliefs are inferred from his actions. In a second step, these attitudes are shown to be incompatible with the way he judges himself. Clearly, a crucial background assumption is that one should apply the same moral standards to everyone, including oneself.

Furthermore, there is *normative disputation*, a technique that is intended to challenge certain beliefs of the patient that have normative content. Usually, it is used for changing particularly *rigid* normative beliefs like the belief that the patient needs to be friendly to *everyone*. These beliefs are usually challenged by pointing out to the patient that this particular belief is incompatible with other beliefs that she may hold, or that she does not hold other people to the same standard (Wittchen & Hoyer 2011, p. 555).

Hedonistic or *functional disputation* questions whether the subject's belief is, as the authors frame it, "useful" for the patient. That is, does this belief contribute to him achieving his life-goals or does it hinder his progress towards them? According to these psychologists, only those harmful beliefs of a patient that are also *irrational* are challenged this way. This would suggest that, in therapeutic practice, only those normative beliefs are challenged by hedonistic disputation that have before emerged as theoretically irrational (Wittchen & Hoyer 2011, p. 555-556).

That is, they would first need to be shown to either be incompatible with the best empirical evidence on the issue available to the patient or the therapist, or with the patient's background beliefs. But I think that this is too strong a requirement: Instead, it suffices if the *and* the fit with the patient's background beliefs do not yield a clear result concerning the belief's correctness. Actually, I think that therapists are especially prone to use hedonistic disputation in circumstances where the evidence concerning a belief's correctness is impartial – if only for pragmatic or practical reasons.[22]

I find this particularly interesting, as it seems to assume that a belief's effect

[22] This leaves open the question whether it might sometimes happen that beliefs that are taken to be theoretically rational – that is, consistent with the empirical evidence for descriptive beliefs or consistent with the relevant background beliefs and attitudes of the individual for prescriptive beliefs – may sometimes be disputed hedonistically, despite their theoretical rationality. While I cannot offer any conclusive evidence for this being the case, several psychotherapists in training have pointed out to me, in private conversations, that they sometimes use hedonistic disputation for prescriptive beliefs that they take to be theoretically rational, that is, well integrated into the patient's belief system. If true, this would suggest that sometimes, pragmatic rationality can trump theoretical rationality.

on an individual's emotional state may be part of a *good reason* to stop holding this belief. That is, hedonistic disputation simply consists in appealing to the pragmatic rationality of holding the belief in question.

Usually, therapists will refer to both kinds of rationality in challenging a patient's belief in a particular proposition.

To reiterate, even those beliefs that are challenged when therapists employ disputation techniques do not emerge as *irrational in every possible way*: Adopting them is taken to be the result of a normal, often even rational, reasoning process. This is where the reference to points in time comes into the evaluation of someone's belief(s) as either rational or irrational.

To come back to something that I commented on very briefly before: Why would it be important, in the context of therapy, to be able to say at which point in *time* it was either (1) *theoretically rational* to believe that p or (2) *pragmatically rational* to believe that p^*?

The reason is that it may very well have been *both* theoretically and pragmatically rational to believe that p at a point in time t_0, but that it may have *become* theoretically as well as pragmatically irrational to believe in the truth of p at the current point in time t_c. This seems to result in attribution only a limited amount of responsibility to the agent, since the only thing she can plausibly be accused of is that she did not *update* her belief system according to the evidence available to her – but this, – as I already pointed out – is statistically normal, and thus, hardly something that would make her blameworthy.

I think that the pragmatic (ir)rationality of beliefs is important in therapy partially because it is used to point out to the patient that changing a particular belief is important for her in terms of her goals. This is substantially different from theoretical rationality. In the case of theoretically irrational beliefs, it may happen that, while the patient has *some* beliefs that are not well-grounded in the empirical evidence or that stand in tension to other beliefs of hers, these are not really problematic. This happens if they have no consequences besides resulting in the subject adopting some false beliefs about the world.

Above, I have shown how my concepts of pragmatic and theoretical rationality might elucidate parts of psychotherapeutic practice. Now that we have seen which theoretical considerations speaks in favor of my account, I would like to present some cases that serve to show how my account can handle more complicated cases.

What will this look like in practice? Consider a situation where a psychotherapist wants to claim two things: firstly, that her patient's emotional reaction to a particular situation is, all things considered, understandable. Secondly, that at least one of the patient's underlying dysfunctional beliefs is, in one sense, rational – while being, in another sense, *irrational.* This therapist will not be in a position to claim that, given *the same set of beliefs*, both the patient's emotional reaction is reasonable, while the proposition in question is irrational.

Consider the following example: Think of a woman – let's call her Maria – who has been suffering from SAD for eleven years. In CBT, her psychotherapist points out to her that one of the core beliefs that contributes to maintaining her anxiety is that she will be socially excluded by others if they observe her making mistakes in public. Her resulting emotion, that is, severe performance anxiety when she speaks in public, emerges as reasonable, since it fits the underlying proposition as well as her background beliefs – presuming that her background beliefs includes propositions like "Making a mistake makes other people regard me as a less valuable person". But what about the belief itself? It will be reconstructed as rational at the point of time of belief formation and irrational at the current point in time in at least one of the two senses.

At the time of belief formation, it was either rational for Maria to adopt this belief in the sense of (1) theoretical rationality or (2) pragmatic rationality. That is, this belief was either well supported by the empirical evidence when it was formed, or it was to be expected that acting in accordance to it would probably contribute to her achieving her explicit goals, when compared to the set of relevant alternative beliefs. Let us assume that Maria experienced it more often to make a mistake in public *without* being explicitly devalued by her peers, but that there were some instances where she was treated badly by her peers.

In that case, forming her belief was not theoretically rational, since the evidence available to Maria actually supported the opposite conclusion, namely, that other people would, were she to make a mistake in public, not change their behavior in ways that are relevant for her. But forming this belief still counts as *pragmatically rational*: When Maria formed it, she experienced some situations in which she was devalued and treated badly by her peers after making a mistake in a presentation in public. Since being valued by her peers was one of her main goals at the time, it was rational for her to adopt the belief that making mistakes in public would lead to devaluation by others. By acting in accordance with this proposition, she avoided even those rare cases of devaluation.

In fact, psychotherapists would regard Maria's belief as neither *fully* rational or irrational, but as being formed on the basis of *some* empirical evidence in connection with specific general human tendencies and biases. In the case at hand, we may think that Maria's belief is caused by a combination of being devalued when making a mistake in a public presentation several times and the very human tendency to assign more weight to negative experiences than to positive ones (for a review of the literature on this matter, see Baumeister et al. 2001). Especially when anxiety – understood as a danger-detection signal (Eysenck 2013) – is concerned, it makes sense to think that, in the EEA, false positives were less problematic for survival and reproduction than false negatives, thus resulting in a smaller net cost of false positives compared to false negatives (e.g. Haselton et al. 2016).[23] If this is true and human beings, due to their evolutionary history, tend to overestimate threat[24], then the assessment of Maria's belief changes: Adopting this belief is then best described not as theoretically rational, but instead, as relatively *normal* in the sense of being brought about at least partly by mechanisms that have evolved in the EEA. In this case, "normality" does not refer to mere statistical normality, but to whether or not a behavior is brought about by a disposition that carries out its evolved function.[25]

Moreover, the belief in question is taken to be irrational in at least one of two senses: For one, it is theoretically irrational to hold at the current point in time, because it is not well grounded in the available evidence. For the other, it is irrational because it *runs counter* to achieving her goals. That is, given the fact that her dream job would require Maria to speak in public more

[23] This is an application of error management theory, which, strictly speaking, applies only to judgements of threat, not to the emotional response – that is, anxiety – as such. But this is relatively unproblematic, since, in the cognitive framework of Beck, anxiety can be caused by such judgements of currently being in danger.

[24] This is very rough, of course. When I speak of human beings overestimating threat, what I mean is their tendency to overestimate how threatening a particular stimulus is. This can take the form of overestimating the size of an apparently threatening agent, the misinterpretation of particular stimuli as threatening – as is the case in arachnophobia, for example – or the like.

[25] I will investigate those issues in more detail in the next chapter.

often than she had to speak in public at the time of belief formation, it is now pragmatically *irrational* to hold onto this belief.

An interesting case for testing my analysis of rationality are specific kinds of *predictions*. In particular situations, those beliefs that maintain a mental disorder seem to be grounded in good evidence, that is, they are theoretically rational, and may even be pragmatically rational. Think of a woman who is in a toxic relationship with a partner who continually devalues and mistreats her. Holding fixed her specific situation, the belief "My future is hopeless because there will be much more suffering than happiness in it." is warranted by the available evidence – if we take it for granted that having more suffering than happiness in one's future means that it is hopeless. I will simply do so for now. Nonetheless, a psychotherapist might think of this as a harm-inducing belief that should be changed. He would then point out that this belief actually *does* emerge as irrational if the current situation is *not* held fixed, but alternative scenarios are taken into account. If these are reasonably close to reality, they serve as good evidence against the certainty that is implicit in this person's belief. When such alternative scenarios are taken into account, there is more happiness than suffering in her future.

Consider now the case of a terminally ill patient who has this very same belief. In his case, therapists would probably not deny that his future is hopeless in a certain sense, and that there will probably be more suffering than good in it.

There is yet another matter that I should deal with here. This is the fact that many people in psychotherapy report contradictory beliefs. For example, patients often report situations in which they consciously know that, in fact, many people *do* like them. At the same time, they have the deep-seated, enduring belief of not being liked. Our analysis thus seems to have the result that this belief cannot be either pragmatically or theoretically rational, since it stands in tension with the agent's background beliefs. But I take it that psychotherapists do, in fact, want to ascribe relative rationality to such beliefs as well. This seems to stand in tension with the idea that any belief of an individual, in order to be evaluated as rational, needs to be consistent with the background of beliefs of this very individual.[26] I think that the solution of

[26] Interestingly, if one takes the background of beliefs to be inconsistent and belief to be closed under entailment, according to the rules of standard propositional logics (more precisely, the so-called "principle of explosion"), the patient will need to believe *everything* (Priest 1998). I take it that this implies that psychologists take belief not to be closed under entailment, just as Priest

psychotherapists here is to only require that beliefs, in order to be counted as relatively rational to hold, must be consistent with only some *relevant* background beliefs of the individual. Concerning this set of beliefs, the cognition in question should be relatively well integrated, though. Now, when considering the (ir)rationality of those *actions* that the individual performed at the outset, they look at those subset of the agent's beliefs that are action-guiding in that particular situation.

Now, why does it make sense to think that patients suffering from mental disorders are irrational in *any* of the senses sketched above? The fact that they are pragmatically irrational, that is, have beliefs that impede their ability to achieve their explicit goals seems almost trivial: otherwise, these individuals would not be in therapy. Usually, mental disorders have harmful effects on individuals that, at some point or another, also counteract their efforts to achieve their pragmatic goals. Furthermore, the central assumption in talking therapy is that the patient can actually change his beliefs, or, more easily, his actions to intervene on his symptoms.

I will elaborate on what exactly this means in the following chapter on functionality. For the moment, though, the functionality of an action, behavior, or cognition may very roughly be understood as the extent to which it brings about subjective well-being in the individual in question.

4.6 Conclusions

In this chapter, I proposed to distinguish two notions of rationality in order to make sense of how psychotherapists either utilize the term or how they dispute their patient's beliefs. This analysis used exemplar cases where the notions of rationality and irrationality are employed. I distinguished between the notions of pragmatic and of theoretical (ir)rationality. After characterizing these two notions, I showed both how my analysis can make sense of certain aspects of therapeutic practice, and I elaborated on how it can deal with what might appear to be cases that stand in tension to this analysis.

As I have pointed out above, the concepts of (ir)rationality, (dys)functionality and normalcy are used commonly in explanatory practices within psychotherapy in order to de-pathologize, that is, to represent the patient as relatively

suggests in his paper.

rational, while simultaneously singling out particular beliefs of hers as both harm-inducing and irrational. These practices often rely on the notion of the (statistical) *normalcy* of particular reasoning processes or belief updating processes. Sometimes, this judgement of normalcy is accompanied by a clarification that not only is this process *normal*, but even more, it even has a *function*. In the next chapter, I will delve deeper into the interrelated notions of function and (dys)functionality.

What is the added value of analyzing this concept, given that I have just provided an understanding of several concepts of (ir)rationality? What can talk about dysfunctionality[27] *do* that is not provided by either concept of irrationality discussed above? In my understanding, notions of irrationality are mainly used by psychotherapists for cognitive disputation, that is, to convince her patient that a particular thought or belief is not reasonable to hold and thus, *should* be abandoned, given the patient's individual belief system and her considered preferences. By contrast, the concept of dysfunctionality primarily serves the purpose of pointing towards factors that are causally relevant for the patient's symptoms and that can be intervened upon. These need not necessarily be irrational in any sense of the word, but nonetheless, dysfunctional beliefs are a very common target for disputation techniques. I intend to make sense of this apparent tension in the next chapter.

[27] In clinical psychology papers, the authors seem to often use the term "maladaptive" (roughly) interchangeably with "dysfunctional". Even though one may argue that there is a slightly different sense to this term, I will, for the purposes of this chapter, take maladaptiveness to be nothing more than dysfunctionality. The reader is thus kindly asked to understand the two terms as synonyms.

Chapter 5

Concepts of (Dys)functionality in Psychotherapeutic Practice

5.1 Introduction

In this chapter, I will analyze one particular notion of *dysfunctionality* that is highly relevant for psychotherapeutic practice and is also at play in clinical psychological research. The thesis of this chapter is that this notion of dysfunctionality is best understood as pointing towards factors that counteract the agent's psychological, physical, social, or other *needs*. This focus on the individual's needs makes dysfunctionality the more *objective* notion when compared with those notions of rationality that I presented in the last chapter. Together, my analyses of dysfunctionality and irrationality allow taking an instructive perspective on further aspects of psychotherapeutic practice.

This introduction serves the purpose to motivate analyzing the notion of dysfunctionality over and above the notion of irrationality.

Before noting the differences between these two concepts, let me briefly talk about their similarities. There is an intimate historical connection between these two concepts: What is today referred to as "dysfunctional" by proponents of CBT often coincides with what was historically framed as "irrational", especially by proponents of *Rational-Emotive Therapy* (*RET*), one of the early precursors[1] of CT that is based on similar principles (Ellis 1980). Although *explicit* talk about the "irrationality" of certain beliefs, actions or behavioral

[1] Although several psychologists understand RET rather as a *version* of CT than as another form of therapy, I will, since I use "Cognitive Behavioral Therapy" and "Cognitive Therapy" to denote those forms of psychotherapy that have originated with Aaron Beck (Beck 2005), not count RET as a version of CT or CBT.

strategies is not commonplace anymore in clinical psychology and psychotherapy, they still make *implicit* reference to the patient's reasonableness and rationality when explaining mental disorders and when intervening on their beliefs by means of cognitive restructuring. I have called those processes that are involved in the explanation of mental disorders to patients *rationalizing* and (following Bolton 2008) *normalizing.* Often, they include reference to the *normal function* of a behavior or belief, that is, to the fact that it serves the individual to accomplish a specific aim and to the fact that it is statistically normal. Some of the patient's beliefs and behaviors result in suffering, though – in these cases, psychotherapists talk about "dysfunction" and "dysfunctionality".

Consider the following quote of one of my interviewees about the relation of these two terms:

> "'Dysfunctionality' has replaced 'irrationality', because people [mental health professionals] wanted to get away from telling people [patients]: 'You are not rational, you are basically crazy if you are thinking like this.', but instead, to say: 'dysfunctional', somehow 'does not lead me where I want to be'. [...] Thus, 'it does not work for me', and this is how, in my experience, 'dysfunctional' [...] is usually used. Ehm, by therapists. Thus, as something which does not have the effect that I wanted to bring about emotionally. Yes? [...] Thus, your thought of 'dysfunctional means to induce suffering' is not that bad."

Here, my interviewee presents his view on *why* talk of "dysfunctionality" has replaced talk of "irrationality". His account clearly has therapeutic goals and values at heart. When he claims that therapists did not want to implicitly represent their patients as irrational or even "crazy", he seems to emphasize the need to tell patients a story that allows them to understand themselves as *relatively normal* or – as he seems to suggest – *relatively rational.*

This also fits an early characterization of dysfunctional thoughts by Beck:

> "Dysfunctional thoughts [...] may be defined as stable and unrealistic rules, beliefs, or attitudes about the world and oneself, which hamper an adaptive coping with the environmental demands." (Beck 1979, p. 116, cited in Losada et al. (2006))

Beck's characterization resembles the definition of "irrationality" that I presented in the preceding chapter, just as the quote of my interviewee would suggest. Both characterizations have the same two components: Firstly, a thought or belief is dysfunctional (or irrational) only if it *misrepresents* the state of the world somehow. Secondly, it is dysfunctional (or irrational) only if it hampers adaptive coping with the demands set by the environment – that is, it hampers acting in a way that would satisfy one's needs, given the environmental constraints.

It seems that certain beliefs are intervened upon *because* they are dysfunctional, not because they are irrational. Later, Beck and Weishaar state the following:

> "The cognitive therapist eschews the word *irrational* in favor of *dysfunctional* because problematic beliefs are nonadaptive rather than irrational. They contribute to psychological disorders because they interfere with normal cognitive processing, not because they are irrational." (Beck & Weishaar 2011, p. 279)

Here, the authors claim that dysfunctional beliefs are the target of clinical interventions, because – *by virtue of being dysfunctional* – they contribute to the development and maintenance of mental disorders. But there is something noteworthy about this quote, when compared to the earlier definition of the term by Beck. That is, the authors do not describe these beliefs as "unrealistic" or incorrect anymore. This suggests that, today, there might be an interesting distinction to draw between irrationality and dysfunctionality.

In developing further on this idea, I will, in this chapter, offer a reconstruction of clinical psychologists' and psychotherapists' talk about dysfunctionality. In doing so, I take into account also to non-empirical virtues like simplicity and precision, instead of *only* trying to account for how the term is actually used in practice. It is supposed to account for why most therapists seem to think that dysfunctional beliefs are *also* irrational while simultaneously highlighting the subtle differences that exist between those two concepts and the work that these respective concepts do.

The notion of "dysfunctionality" is used in connection with many different objects: Therapists and researchers in clinical psychology alike talk about (ordered by frequency of occurrence) "dysfunctional behavior", "dysfunctional attitudes", "dysfunctional beliefs", "dysfunctional (family) dynamics", "dysfunctional schemas", "dysfunctional (behavioral) strategies", "dysfunctional thought processes", and the like. Sometimes, "rules" are also referred to as dysfunctional. These examples are taken from various papers from clinical psychology and from conversations with several psychotherapists. I used an academic search engine to generate a rough estimate for how common each of these terms is. Clearly, the terms "dysfunctional attitudes" and "dysfunctional beliefs" were used most frequently.[2] I think that these different uses can be reduced to three categories: namely, (1) beliefs (or thoughts), (2) actions (or behaviors) and (3) strategies.

Let me give some reasons to think so. When Beck and Weissman characterize the aim of their "Dysfunctional Attitude Scale", they write that "The development and validation of an instrument to identify the common assumptions underlying the typical idiosyncratic cognitions in depression is described." (Weissman & Beck 1978, p. 3). The term "attitude" thus is used roughly in the same way as "assumption" or "belief". Similar considerations apply to the usage of "schema" here, since beliefs, according to Weissman and Beck, "act as schemas" (Weissman & Beck 1978, p. 3). Furthermore, I think that "dysfunctional thought processes" usually are instances of the class of (in that case internal) behaviors. This is because what psychologists refer to as dysfunctional thought processes are tendencies of the respective individuals to reason in a certain way, for example, to overgeneralize, to conduct selective abstractions, and the like (Beck 1963). Finally, I take it that strategies and rules can simply be grouped together.

My working hypothesis is that the notion of "dysfunctionality" has similar connotations in all of these contexts. That is, I will assume that the same notion is used in all of these cases. This sets the agenda for the following: In developing my account of dysfunctionality, I will use instances from these different categories as examples.

To reiterate: Why should we be interested in analyzing "dysfunctionality" over and above "irrationality"? After all, both terms are mostly used in connection

[2] For those readers interested in the precise numbers: I typed these respective expressions into GoogleScholar in September 2019. At the time, "dysfunctional behavior" produced most hits, that is, 20.400, while "dysfunctional attitudes" generated 16.100 hits, followed closely by "dysfunctional beliefs" with 15.300 hits, and then came "dysfunctional family dynamics" (1.730), "dysfunctional schemas" (1.510), "dysfunctional strategies" (1.350), "dysfunctional thought processes" (776) and "dysfunctional rules" (452).

with the same kind of cognitive entity, namely, *belief*. Additionally, both are used for the same kinds of beliefs, namely those that have problematic consequences for the individual.[3] But they nonetheless denote different features of such entities. Let me illustrate this difference by means of an example.

Consider the case of André, who suffers from anorexia nervosa. One of his dysfunctional beliefs might be "I will be loved by other people only if my body is attractive enough". That is, he takes his own physical attractiveness to be a necessary condition for being loved by others. On the face of it, this belief is probably false. Nonetheless, it might be rational for him to both (1) have formed this belief and (2) to act in accordance with it, relative to his explicit goals. To see this, consider a scenario in which André was overweight as a teenager, was criticized by his parents for his weight, and bombarded by media images of extremely muscular men that were simultaneously represented as happy and amiable. At some point in time, he lost some weight, built some muscle, and, as a result, got positive feedback from significant others about his changed outer appearance. If we now read his belief not as a *strict* conditional but instead as one that admits for exceptions – which is the most plausible way to treat actual beliefs of real people –, we might say that, on the basis of the evidence available to him back then, it was indeed theoretically rational to form the belief that he will only be loved by other people if his body is attractive.

Furthermore, it might *also* have been pragmatically rational for André to behave according to this belief: If one of his high-ranking goals was to be liked by other people, then it seems only pragmatically rational to bring about whatever one takes to be a necessary condition for this being the case. It might even *still* count as pragmatically rational for this individual, if he explicitly values being loved by other people over and above being healthy. At the same time, this belief would qualify as dysfunctional according to clinical psychologists and psychotherapists. In his particular case, believing that other people only love him if he is skinny contributes to his anorexia nervosa, thereby being actively harmful.

Even if most actual cases in psychotherapy are not as neat as this example, it nonetheless suffices to show that, at least *in principle*, the category of dys-

[3] That is, other than we might have thought, it is *not* the case that the notion of irrationality is mainly only in connection with reasons and actions, while the notion of dysfunctionality is used in connection with mechanisms, behaviors, and the like.

functionality does not amount to the same thing as either sense of rationality discussed in the last chapter.[4]

When taking the psychological literature at face value, we will notice that the authors utilize several different notions of functionality and dysfunctionality, often without explicitly keeping the different senses apart. Sometimes, the term is used to denote the fact that a certain mechanism does not carry out its evolved function, sometimes to point out that something is not acting normally, and, finally, to stress that an action or belief is *harmful*. Wakefield (1999) also notices this, making the following statement about the latter sense of dysfunctionality:

> "There is a less precise, colloquial sense of *dysfunctional* that applies to any trait of an individual or family that causes undesirable or ineffective behavior in relating to the current environment. This sense is used to negatively evaluate traits, people, or families as *dysfunctional* and has no necessary relationship to the [harmful dysfunction] analysis's dysfunction component. Indeed, this colloquial sense of the word *dysfunctional* is more closely related to the harm component of the HD analysis than to the dysfunction component." (Wakefield 1999, p. 376)

I am inclined to agree with Wakefield's suggestion that there is one meaning of dysfunctionality that is more closely related to harmfulness – having more to do with whether a belief or strategy allows an agent to satisfy her *needs* than with whether it fulfills its evolved function. But other than Wakefield, I believe that it is epistemically fruitful to take a closer look at this meaning of dysfunctionality. I take it that his understanding of the "colloquial sense" of the term is not as developed as it could be. If we can provide a clear understanding of this term, we may also gain insights about the intricate puzzles

[4] An interesting feature of dysfunctional beliefs is that, more often than not, dysfunctional beliefs are *rigid* in the sense of including universal quantification, and very often, these beliefs will also contain evaluative operators like "should", "must" and the like. As one of my interviewees said:

> "These universal operators! 'Always', 'never', 'no one', and so on. Right? This is a topic I discuss often, when clients say: 'I have never...', 'I always have to...', I say: 'Oh, watch this. They go on the black list."

At least those beliefs that only have descriptive content will, for this reason alone, hardly count as *theoretically* rational. But they may still, in many cases, be pragmatically rational in the context of belief formation, but theoretically irrational in the current context. For those beliefs with prescriptive content, the same holds, at least according to many therapists.

about psychotherapeutic practice that I have presented above. Finally, this "colloquial" meaning of dysfunction is not only at play in therapeutic practice, but has become an essential part of theorizing about mental disorders in psychology through explanatory models. For this reason, I will focus on it in the following.

At this point, I would like to remind the reader of one interim conclusion from the first chapter: There, I showed that the main explanatory strategy of these two models relies on identifying beliefs of an individual that made sense when they were first adopted but have become maladaptive for the patient later, nonetheless remaining stable over time. In the fourth chapter, I described two accounts of when it is rational for someone to adopt a belief at a particular point in time. One relied on whether the evidence for that very proposition is more substantial than the evidence for its negation, while the other relied on the evidence for the *instrumental* value of a proposition in relation to the patient's explicit goals.

In this chapter, I will analyze what psychologists mean when they claim that a specific belief is dysfunctional. As I pointed out in chapter three, I take it that vocabulary surrounding the notion of rationality is used mainly to attribute only limited responsibility for developing the disorder to the patient and to make her believe that she is not fundamentally alien from other human beings because of his disorder. Finally, it gives the therapist opportunities to challenge particular beliefs of the individual. By contrast, the notion of dysfunctionality is used primarily to identify those beliefs that are at the core of a mental disorder and its maintenance over time and that should thus be challenged in order to treat it. Importantly, therapists need not rely on a specific concept of *normal function* in order to characterize their concept of dysfunctionality.

Clearly, it is not only important here to talk about dysfunctionality as a marker of those beliefs and strategies that are at the heart of mental disorders, but it will also be necessary to talk about uses of the term "function" and what these might refer to – as I have already pointed out, evolutionary functions are explicitly referred to in the explanatory model of depression presented by Beck & Bredemeier (2016). Furthermore, there is frequent reference to the *function* of certain kinds of behavior or thought and patterns of behavior *for* a particular individual. This functional terminology will also be discussed here.

In developing and arguing for my account of dysfunctionality, I will make use of several *criteria of adequacy.* These are based on what we might call a "meta-criterion" of *descriptive adequacy.* I assume that my analysis should, at least to a certain extent[5], fit the usage of these terms by psychologists and psychotherapists. In the following, I will further specify the content of this criterion, through my analysis of both therapeutic practices and research papers. In distinguishing between important and negligible factors that are mentioned by mental health professionals, I will rely on considerations of simplicity and parsimony. A first consequence for my analysis of this notion is the following:

(DA_1) The notion of "dysfunctionality" should in principle be applicable to beliefs, behaviors, and strategies of human beings.

In arguing for my analysis of dysfunctionality, I will proceed as follows: In a first section, I will discuss several different concepts of function, dysfunction and dysfunctionality that occur in clinical psychological research and theorizing about psychotherapy. Secondly, I will narrow the discussion down to that particular notion of dysfunctionality that is central to both explanatory models of mental disorders and most important for psychotherapeutic practice. The second section then deals with how this concept is used in therapy. In a third section, I will sketch selected philosophical accounts of (dys)function to provide some theoretical and conceptual resources for my ensuing analysis of this concept. Finally, in the fourth section, I will suggest an analysis of dysfunctionality, drawing on potential examples and counterexamples. I will argue that my account can accommodate these cases well enough to be satisfying as an analysis of dysfunctionality.

5.2 Functional Concepts in Clinical Psychological Research

I already pointed out in the first chapter that there is an abundance of functional concepts used in clinical psychology and illustrated this presenting several uses of functional concepts in the literature. Here, I would like to provide a more in-depth treatment of these notions. This serves the purpose to provide some background on the later focus on only *one* of those concepts.

I include clinical psychological *research* – over and above psychotherapeutic practice – into the discussion of my concept of dysfunctionality, since I am interested here in the interrelations of research and therapeutic work. The

[5] Compare my methodological discussion in the last chapter.

concept of dysfunctionality is one object that is of relevance for both domains, being present both in explanatory models of mental disorders and at the center of psychotherapeutic interventions. Thus, to develop an account of dysfunctionality that fits both domains, it is important to also take research papers and monographs on the respective mental disorders into account. Furthermore, many therapeutic practices and interventions also depend on those very models of mental disorders.

There are three main cases that I will take a closer look at in the following. I discuss them because they are paradigmatic occurrences of functional vocabulary in the clinical psychological literature. For one, there is the APA's characterization of mental disorders that both refers to "mental functioning" and "dysfunction". Secondly, in Beck & Bredemeier (2016), the authors speak – among other things – of the *evolutionary function* of particular behavioral and emotional patterns. In the same paper, they also introduce a concept of dysfunctionality that is closer to the one employed in explanatory models of mental disorders and in psychotherapeutic practice. Thirdly, I will go back to the model of Salkovskis et al. (1998). On the basis of these cases, different notions of function, functionality and dysfunctionality can be distinguished. Once I have provided a first analysis of the meaning of these terms, I will narrow my focus to only one of them. I do so because, of the three, it is the one most relevant for psychotherapy and those models of mental disorders that I have discussed in the first chapter.

To begin, which understanding of "dysfunction" is encoded in the DSM-5's characterization of mental disorder? As a reminder: According to the APA, "[a] mental disorder [...] reflects a dysfunction in the psychological, biological, or developmental processes underlying mental functioning" (American Psychiatric Association 2013, p. 20). Here, two instances of functional vocabulary occur: for one, they speak of "mental functioning", for the other, about a "dysfunction" of underlying processes.

The first usage concerning so-called "mental functioning" is best understood as making reference to relatively *normal* and *good* workings of the human psyche. This becomes clear when considering that it is used to draw out a contrast between those individuals who suffer from mental disorders and healthy individuals, whose normally-working psychological (and other) processes account for their mental functioning. Thus, I think that it is best understood as a concept referring to either statistical normalcy in the workings of mental faculties or to a particular, positively evaluated and relatively common state of human agents.[6] That is, someone functions well mentally if her psychological processes allow her to live her life – given statistically relatively normal background conditions[7] – without these processes causing her to feel significantly more harm or significantly less positive emotions than other individuals experience in their daily lives because these processes operate worse than is statistically normal.[8]

A similar meaning is also employed by the APA when they list so-called "functional consequences" of mental disorders (e.g. American Psychiatric Association 2013, p. 158). For *Disruptive Mood Dysregulation Disorder*, a mental disorder classified among the depressive disorders, they list the following consequences:

> "[...] marked disruption in a child's family and peer relationships, as well as in school performance. [...] difficulty succeeding in school; they are often unable to participate in the activities typically enjoyed by healthy children; their family life is severely disrupted [...] and they have trouble initiating and sustaining friendships [...] Levels of dysfunction in children with bipolar disorder and disruptive mood dysregulation disorder are generally comparable. Both conditions cause severe disruption in the lives of the affected individual and their families." (American Psychiatric Association 2013, p. 158, my italics)

Thus, "functional consequences" of a mental disorder are negative conse-

[6] Other than one might think, this cannot be well captured by Cummins' functional analysis. This is because, in his account, functions are always identified relative to an explanandum disposition. That is, the interesting functions will instantiate statistically normal workings of the human psyche only if the explanandum in question is itself a capacity that most human beings have. As we have seen before, one might also understand my exemplary explanatory models of mental disorders as functional analyses of complex dispositions, where the complex dispositions in question are mental disorders. In this case, *neutralizing actions* or *safety behaviors* may emerge as functions, since they allow the analysis of the complex explanandum disposition into simpler sub-dispositions. Clearly, these kinds of behaviors are emphatically *not* statistically normal, even though on might make sense of why they occur in psychological explanations.

[7] What these are will clearly depend on the individual's cultural and social setting.

[8] One might have doubts about this, since the APA has set up their diagnostic criteria in a way such that the lifetime prevalence of mental disorder is very high, with estimates ranging from 12 to 47% (Kessler et al. 2007, p. 168), which makes it likely that their actual understanding of mental functioning also incorporates an evaluative component, not only a statistical one. But I think that what underlies this characterization that we have seen above is actually the idea of statistical normalcy.

quences of the disorder for the so-called "functional level" of the individual. I take it that the functional level is understood relative to socially accepted markers of the good life: for example, the APA lists being successful at school, engaging in enjoyable activities, and friendships. That is, both usages of functional vocabulary have a norm at heart that determines when something (e.g., someone's mental faculties) or someone *functions* well enough. This normativity of functional terminology is actually widespread in clinical psychology and psychotherapy. I will refer to it again when I deal with my second case.

The second usage of functional vocabulary by the APA, i.e., talk about the disorder "reflect[ing] a dysfunction", seems quite different: I think that one can make the case that it is about evolved functions not being carried out anymore *or* about the statistical abnormality of processes. In the latter case, those processes would not contribute to those outcomes that they normally contribute to. For now, I will remain neutral on which of the two is more plausible, since this discussion is beyond the scope of my dissertation.[9]

Secondly, there are several uses of functional vocabulary in Beck & Bredemeier (2016). I focus here on selected instances of functional vocabulary in their paper, both omitting those that I have either already covered and those that are not of interest for me in the following.[10] For example, the authors refer to so-called "dopaminergic dysfunction" (Beck & Bredemeier 2016, p. 605) and "immune functioning" (Beck & Bredemeier 2016, p. 607). Although both are interesting instances of functional vocabulary, they pertain to physiological processes and are thus less interesting for my analysis of function concepts in clinical psychology and psychotherapy.

At one point, the authors make explicit reference to Wakefield in employing the notion of "dysfunction":

> "We also address the potential functions and adaptive value of milder (i.e., 'subclinical') symptoms, as these are key to an understanding of when and how a given level of depression is dysfunctional or maladaptive (see Wakefield (1999))." (Beck & Bredemeier 2016,

[9] There are, indeed, some reasons to think that Wakefield's analysis cannot capture all the dysfunctions that psychologists are thinking of here. I think of examples that have been presented as counterexamples to Wakefield's account that supposedly *are* mental disorders, but do not have evolutionary dysfunction of the kind thought of by Wakefield at their core. One example of this is, arguably, *dyscalculia* (Lilienfeld & Marino 1995).

[10] It should be noted that the authors use functional vocabulary 20 times in their paper (Beck & Bredemeier 2016).

slightly adapted)

This is interesting in part because the main claim of the authors in this paper is that the symptoms of depression are the result of a behavioral program that has an evolved function. This would suggest that the authors actually misinterpret Wakefield here, who takes a dysfunction to be present *only* if the respective mechanisms does *not* carry out its evolved function. But Beck and Bredemeier appear to argue for precisely the opposite: according to them, in depression, a process *does* carry out its evolved function, but is dysfunctional for the individual because the environmental conditions have changed. On the face of it, this would seem to imply that MDD is not a mental disorder *at all*.

Furthermore, the authors repeatedly refer to the *function* of the depression program, clearly using the term with an evolutionary meaning. This becomes clear when the authors compare the behavior of depressed individuals to that of other animals in order to make an argument for this behavioral program having an evolutionary function, in this case, attracting the attention of significant others (Beck & Bredemeier 2016, p. 604).

When speaking about such behavioral patterns, psychologists seem to use a more clearly evolutionary notion of dysfunctionality. On this understanding, a condition is *not* dysfunctional if it is brought about by a mechanism that carries out its evolved function. I think that the reverse is also true: That is, a pattern of behavior is dysfunctional just in case it is the result of a mechanism not carrying out its evolved function any longer. This is the closest clinical psychologists' and psychotherapists' usage of functional terminology comes to Wakefield's understanding of the term (see, e.g. Wakefield 1992). As we have seen, at least in my exemplary model of depression, this notion of evolutionary function and dysfunction is used differently than we would expect according to Wakefield.

Finally, there is a third notion of dysfunctionality, which seems to be a *primarily* evaluative concept. As stated in the first chapter, Beck developed his "Dysfunctional Attitude Scale" (Weissman & Beck 1978) as an instrument for measuring long-held attitudes with negative content that occur frequently in depressed people. This raises the question whether what makes those attitudes dysfunctional is the fact that they have negative content. I think that this is not quite right: After all, dysfunctional vocabulary is also frequently employed in characterizing relevant features in other mental disorders that are not characterized by their negative content, for example, in OCD. There, what is dysfunctional is the patient's belief that his intrusions are indicators of *danger* (Salkovskis et al. 1998). For these reasons, I think that dysfunctionality is not solely about the *negativity* of underlying beliefs, but more about their *harmful effects*.[11]

This understanding of dysfunctionality is also exploited in the model of Salkovskis et al. (1998). In this model, another instance of functional terminology is present. Just consider a patient's disposition to show neutralizing actions. When viewed from the perspective of the individual, engaging in neutralizing actions serves a *function*: namely, that of *neutralizing* the thought(s) that one is responsible for preventing harm. But this intention is not fulfilled by the behavior in question. This is because neutralizing actions, in the long run, *increase* instead of decrease the number of responsibility beliefs the individual experiences. At the same time, their function is fulfilled in the short term: For a certain period of time, the responsibility beliefs of the individual decrease.

Here, clinical psychologists employ a *teleological*[12] notion of function, that is, a notion that implies goal-directedness[13]: Neutralizing actions have the *function* to reduce responsibility beliefs because they are *aimed at* reducing the person's responsibility beliefs. Thus, goal-directedness is inherent in this conceptual-

[11] The confusion here might partially stem from the fact that those kinds of beliefs that have bad or negative effects seem to be thought of as having *negative* content because of how we evaluate the effects they bring about. But this is not correct, as the example of someone with massively positively biased self-concept shows: Just like beliefs that are extremely *negatively* skewed, those beliefs with extremely positively skewed content may be harmful, too. Just think of individuals who suffer from manic episodes (For more on this, please consider the diagnostic criteria of manic episodes in appendix A.1).

[12] Teleological notions of function were heavily debated in the philosophy of biology, in particular. Some of the reasons why such notions are controversial in biology and the philosophy of biology are the fact that they seem to require backwards causation, they are often thought to be incompatible with mechanistic explanation, and they are mentalistic, apparently relying on actions of mind where there is no mind (Allen & Neal 2019). In the context of my investigation here, such teleological notions would arguably not be problematic, since I investigate dysfunctionality in agents who have goals and intentions.

[13] Although one might think that *every* notion of function needs, by definition, to be teleological, this is actually not true: Cummins' notion of function, when taken at face value, is not teleological, since it analyzes functions in terms of causal roles. The inability of Cummins' notion of function to distinguish between effects of something that are due to the fact that it carries out its function and its mere side-effects – to use a classic example, think of the heart pumping blood and the heart making sounds, the first of which appears to be part of its actual *function*, and the second being a mere side effect – is one of the major criticisms that is regularly raised against this analysis (e.g. Couch 2019).

ization of function. Since this teleology is based in the person's *intentions*, this does not mean that we have to think of goals as being objectively *out there* in nature.[14]

Furthermore, as we have seen above, there is another component of their model that is frequently dubbed "dysfunctional". This is the problematic *core* belief of an individual that leads to misinterpretations of intrusive thoughts as indicative of danger.[15] The core of what these clinical psychologists appear to mean when they say that this thought is *dysfunctional* is that it is *detrimental to the well-being* of an individual that engages in it.[16]

To summarize, let me offer a quick classification of those functional notions that I have discussed above.

Firstly, I have presented an account of the APA's understanding of (mental) *functioning* as a value-laden concept that primarily exploits societal norms to establish certain features of an individual as indicators of good or sufficient functioning. This concept is vague and highly context-dependent. Furthermore, I reconstructed their notion of *dysfunction* as sub-standard workings of mental processes.

Secondly, I described three notions of *dysfunctionality* as occurring in the clinical psychology literature. Roughly, the notions that I identified can be classified as either (1) primarily statistical, (2) primarily evolutionary or (3) primarily evaluative.

In what follows, I will develop a more precise account of the last understanding of dysfunctionality. There are several reasons for this. The primary reason is that this last notion of dysfunctionality is central not only for research in clini-

[14] It is fascinating to note that Beck actually explicitly states at some points in his 1967 book that he intends to eschew both functionalist and teleological concepts in his theory. As he states: "As the history of science demonstrates, theories that ascribe some design or purpose to natural phenomena have generally been superseded as basic knowledge increased." (Beck 1967, p. 253). He uses this statement as a reason not to think of symptoms as having particular functions for the individual in the situation in question. It is interesting to note that, in my understanding of his theory, he indeed *does* need to speak of teleology and function, but in an, as I argue, unproblematic manner.

[15] For the purpose of this chapter, I will assume that, since the relevant beliefs underlying the disposition to misinterpret one's thoughts is usually said to be dysfunctional, the disposition itself counts as dysfunctional as well.

[16] Let us, for the moment, ignore the fact that psychologists also often speak about the *dysfunctionality* of parenting styles (e.g. Morawska & Sanders 2007, p. 760) and other features of groups such as group dynamics (e.g. Norvell & Forsyth 1984, p. 297), where either the suffering party is not identical with the one that engages in the behavior or either of the two is not an individual. I take it that the general idea of a systematic infliction of harm upon someone is what remains constant in these cases.

cal psychology and explanatory models of mental disorders, but also employed in psychotherapeutic practice, whereas the other notions of dysfunctionality that I have mentioned are less relevant for psychotherapeutic practice. To reiterate, the core of this notion is that a belief is dysfunctional if it is detrimental to the patient's well-being.

From this, we can derive two more features that my account of dysfunctionality should have. What seems to be implicit in how "dysfunctionality" is used is that dysfunctional processes are one important marker to distinguish individuals who are mentally healthy from individuals who are mentally ill. I will thus add this criterion to my criteria of descriptive adequacy:

(DA_2) The beliefs, behaviors and strategies in question typically differ between mentally healthy individuals from mentally ill individuals.

Furthermore, what we may also derive from this – in particular, in conjunction with the discussion of the DSM-5 criteria for mental disorder and Wakefield's remarks on the "colloquial sense" of dysfunctionality – is that what the notion of "dysfunctionality" that I am interested in is centered around the belief's, behavior's or strategy's *harmfulness*:

(DA_3) The beliefs, behaviors and strategies in question lead to significant (subjective) *harm* in the individual holding the belief or carrying out the action/strategy.

In this section, we have seen how researchers in clinical psychology utilize functional concepts. Furthermore, I have identified and provided some reasons to be interested in the particular notion of dysfunctionality that I will analyze in the following. Now, I would like to turn my attention to psychotherapists' usage of functional vocabulary in practice.

5.3 Functional Concepts in Psychotherapy

In the following, I will put particular emphasis on *one* notion of dysfunctionality that is of particular importance for therapeutic practice and that has also influenced clinical psychological models of mental disorders.

One sense of dysfunctionality that therapists seem to employ is the sense of someone's action not contributing to a goal that she intends to achieve. This became clear when one of my interviewees said the following about conditions under which someone's (in this case, the therapist's) actions are dysfunctional:

> "Yes, such a therapist could act in a *dysfunctional* manner, for example, by not activating enough emotion."

Here, the therapist's action is described as dysfunctional, because it does not contribute to her goal of making the patient feel better. Clearly, this use of "dysfunctional" runs the risk of collapsing into the notion of pragmatic rationality that I have discussed in the last chapter.

Note here that there is a small, but important, difference between the notion of pragmatic rationality and this notion of dysfunctionality: While someone may act in an *objectively* dysfunctional way, this might not be pragmatically irrational, because, given his background beliefs, it may seem to be the best course of action. For example, he might think that his patient would profit from a detached, non-emotional perspective – and thus act pragmatically rational. But this may nonetheless be dysfunctional, since it leads to more suffering in the patient.

Furthermore, several psychotherapists I interviewed claimed explicitly that actions, beliefs or strategies of patients are dysfunctional if they *do not contribute* to their explicit goals. While this may be *one* meaning of the term that psychotherapists employ in therapy, other statements suggest that therapists also refer to those thoughts or beliefs as "dysfunctional" that *actively counteract* the satisfaction of the agent's *goal* and, thereby, lead to significant harm.

I take this to be due to the fact that most therapists simply do not distinguish between a behavior that does not contributing to the satisfaction of a goal and a behavior that actively runs counter to someone's needs being satisfied.

One statement that is suggestive of this is the following:

> "Dysfunctional beliefs hinder the person to carry out what she would like to, they are not target-aimed. So, if she has a particular goal, right? To go to work regularly, and then the thought 'I am not good enough.' interferes with that, then this is dysfunctional. [...] So, causing harm, really not getting anything done, right? So both in the external, that *things really do not work*, that one can speak of a restricted functional level, but also that it *causes suffering.* That the thought in itself already... that it impairs the person's well-being so much that no psychic flexibility is there anymore."

There are several things to say about this. It is a characterization of dysfunctional beliefs as *hindering* the person to act in accordance with her goals and values, and as not being *target-aimed*. These clearly are two distinct – although compatible – ways of understanding dysfunctionality. From the last sentence, though, it can be inferred that what really is at the heart of their understanding of a belief's dysfunctionality is its interference with the patient's goals.

Since I have been told this or something very similar by several psychotherapists in my qualitative interviews, I take it to be important for an analysis of dysfunctionality to incorporate this feature. Thus, let me further specify my criterion of descriptive adequacy:

(DA_4) The beliefs, behaviors and strategies in question are either non-goal-conducive or actively counteract the satisfaction of the agent's goals.

There is something more to say concerning this quote that is of interest for us here, though. That is, the therapist in question distinguishes two kinds of ways in which someone can be hindered in his ability to achieve her goals, that is (1) a seemingly objective level, and (2) a more subjective level. More specifically, he seems to be saying that there is a relatively objective *functional level* of someone. There are several possibilities for what this might refer to. For one, functionality may be understood as something that is relative to the individual, or as something relative to the socially constructed expectations that we have about how well someone ought to function. First of all, it might be described as the extent to which a patient can carry out particular acts that she values in themselves or that she values as a means to achieve a state she values. But secondly, this talk about the "functional level" of an individual is dependent upon the norms of a social group. If this is true, someone's functional level describes the extent to which she is able to carry out particular acts that we, in a particular society, consider constitutive of *normal* or *good human functioning*. These two understandings of someone's functional level will, often, coincide. When someone intends to act according to social expectations, as many of us do. But they can become quite distinct in certain individuals who do not care about going to work, for example.[17] I think

[17] For example, both the DSM-IV and the DSM-5 include scales on which to measure the functioning of a patient. In the DSM-IV, this scale was the "Global Assessment of Functioning Scale", whereas the DSM-5 includes the "World Health Organization Disability Assessment Schedule 2" as an assessment of someone's functioning and impairment (Gold 2014).

that this second reading of a person's functional level is actually correct: think again of the APA's characterization of so-called "functional consequences" of a particular mental disorder that I quoted above. According to it, a behavior is dysfunctional when it makes it harder for the individual to engage in activities that we, as a society, consider constitutive of the good life.[18] Interestingly, this is distinguished by the therapist from the subjective suffering that is induced in the individual.

From this, we can derive a further criterion to flesh out *descriptive adequacy*:

(DA_5) The beliefs, behaviors and strategies in question lead to a reduced (objective) *functional level*, that is, a reduced ability of the agent to take part in activities that we, as a society, deem important or worthy of taking part in (work, friendship or the like).

I think that, with this, we have almost everything we need for an analysis of dysfunctionality. But I would like to draw the reader's attention back to something that I said early on, concerning the *intention to intervene* on dysfunctional beliefs.

A helpful quote concerning this issue from my qualitative interviews is the following, in which the psychotherapist talks about those words and phrases that he takes to be indicative of someone talking about his dysfunctional features:

> "And then I will hear, what does someone understand, or what could be translated in such a way, where does someone say: 'And this is how I realize that this is dysfunctional for me: ...', '*this does not do me well*', '*this does not work for me*', 'I do not feel well with it', 'I'm at an impasse', these are images that are used a lot. 'Impasse', 'treading water', 'going around in circles', 'hamster wheel'. [...] Okay. What could better fit the criterion of 'it works', 'it feels good', 'it does me well'?"

In this quote, the therapist mentions different ways that clients might talk about their situation that are *signals* of dysfunctional thoughts or behaviors. The images he mentions seem, by and large, be indicative of an individual

[18] I take it that functional consequences refer, implicitly, to societal norms of what constitutes a good life, not simply to social norms, since their characterization does not only make reference to the individual's ability to work, for example, but also to someone's ability to form friendships or the like.

acting in a particular way without getting the result that he desires and suffering as a consequence of this. Importantly, this does not *only* mean that the dysfunctional beliefs or action strategies are not goal-conducive. Instead, dysfunctional beliefs in particular are better understood as beliefs that *ultimately bring about* this non-goal-conduciveness. Even though, in this quote, dysfunctionality is characterized as relative to the patient rather than to social expectations, I think that we can make sense of this apparent tension by noting that meeting social expectations is often important for individual well-being.

What is implicit in this statement is *also* that usually, dysfunctional beliefs or strategies can be intervened upon. When the therapist asks which thoughts or strategies "could better fit the criterion of 'it works', 'it feels good"' and the like, he clearly thinks about alternative, more adaptive beliefs that someone might acquire.

Let me present an example to further clarify this matter and provide some evidence for it. Consider the case of Andrea, a patient with OCD.

She experiences intrusive thoughts about wanting to kill her colleagues and intrusive images of standing over the corpse of her dead colleagues with a butcher's knife. Remember that, according to our exemplary model of OCD, these intrusive thoughts and images are *not* what brings about her disorder that these thoughts and images indicate which intentions she has. Instead, it is Andrea's more stable, dysfunctional belief, that – given the background conditions of her case – effectively results in her not wanting to go to work, as she concludes on their basis that her intrusive thoughts indicate that she actually desires to and will attempt to kill her colleagues. But why do clinical psychologists and psychotherapists alike describe *this* as the dysfunctional belief, instead of singling out the intrusive thought as dysfunctional, which seems to have the problematic content? After all, *both* beliefs seem to contribute causally to her symptoms, as a very basic counterfactual test shows: If Andrea did not have these intrusions, she *would* be able to go to work daily, wouldn't she? Although this is true, two things are worth highlighting here: For one, having harmful consequences is not sufficient for dysfunctionality. For the other, even though intrusions *are* causally relevant for the maintenance of such a disorder in the sense of being necessary conditions for it, they are not what psychotherapists, in their search for interventions, are interested in. This makes sense, because what differs between mentally healthy and mentally ill individuals is *not* the occurrence of intrusive thoughts (Rachman & de Silva 1978, p. 233).

This consideration shows that dysfunctional thoughts have to be both (1) part of how patients with the mental difficulty in question differ from healthy individuals[19] and (2) in principle suitable for interventions by means of talking therapy.

That is, intrusive thoughts are not relevant for the psychotherapist who is interested in intervening on the mental disorder, in part because they are harder to intervene upon and to change than dysfunctional beliefs. That is, these models have considerations of psychological possibility and of *changeability* at their core: I take it that dysfunctional beliefs are something that can be changed by intervening with the means of psychotherapy.

Thus, another criterion of adequacy is the following:

(DA_6) The beliefs, behaviors and strategies in question can be intervened upon in psychotherapeutic treatment, and they are introspectively available to the agent.

Now that I have pointed out which features this primarily normative notion of dysfunctionality should have, I would like to discuss selected philosophical accounts of function and (dys)functionality with the aim to select components of these philosophical accounts that serve as tools for developing an analysis of the term.

5.4 Philosophical Accounts of Function and Dysfunction

In this section, I will discuss philosophical analyses of function and dysfunction. In doing so, I will point to features of these accounts that we should keep in mind for our analysis of the concept of dysfunctionality. Remember that I aim to provide an account of this term that has at its core that dysfunctional features of an agent are detrimental to that agent's well-being. More precisely, I will use this section to assemble a conceptual toolkit that I will use in the subsequent section to provide an account the concept of dysfunctionality that I am interested in. To do so, I will present the most important accounts of

[19] This is of course very much open to interpretation – especially if we want to allow for the possibility of dysfunctional beliefs and dysfunctional behavioral strategies in individuals without mental disorders. But I think that a similar case can be made when contrasting particular difficulties of the people with the belief with the lack of such difficulties in individuals without this belief.

function and dysfunction from philosophy, focusing particularly on the plausibility of these accounts on their own and on their respective relevance for my analysis.

Let me begin with some words about those accounts of *function*, *dysfunction* and *dysfunctionality* that have been presented in the philosophical literature – in the philosophy of psychiatry, philosophy of medicine and philosophy of biology, in particular. In the following, I will briefly sketch those positions that are important for the following.

For my purposes, the most important philosophical accounts of function (drawing on Boorse 2002, p. 64-68) are three: Firstly, the causal-role analysis by Robert Cummins, according to which a feature's function – roughly – consists in its disposition to causally contribute to the exercise of an explanandum disposition (i.e., Roth & Cummins 2014). Secondly, there are etiological analyses of function, according to which an item's function consists in those of its effects that causally explain its existence. Usually, this is made explicit by claiming that a feature has a particular function F just in case carrying out F helped the organism that exhibits this feature survive and reproduce (compare, e.g. Wright 1973). Finally, there are so-called goal-contribution analyses. Boorse, for example, favors the latter kind of analysis. Very roughly, they make good on the intuition that a function of something is what it does in contributing to a specific goal of an organism, where "goal" is analyzed in naturalistic terms, as we will see below.

Since everyone of these views has its problems, many philosophers have adopted a pluralist stance about functions, assuming that function statements may refer to either causal roles or selected effects. Additionally, there are views that understand every function statement as implicitly and fundamentally *normative*, that is, as referring to an "effect useful or good for some beneficiary" (Boorse 2002, p. 67).

For my purposes, I will engage in more detail with the causal-role analysis and the goal-contribution analysis. I will show that Cummins' account cannot be used to understand dysfunctionality, even though it is helpful for analyzing the explanatory power of these models: The goal-contribution analysis, by contrast, provides conceptual resources that I will exploit in the following. I will also be concerned with clarifying how these accounts of function relate to a widespread analysis of what mental disorder is. That is, I will also be concerned with Wakefield's harmful dysfunction analysis of mental disorder. I mainly deal with his account for two reasons: Firstly, it is important to show how much my account has to offer over and above this notion of dysfunction. Secondly, my account both exploits one of the components of Wakefield's account and will also be *structurally* similar to it.

To begin, there is Robert Cummins' analysis of functions as *causal roles*. Functional analysis asks how a particular system *works* (Cummins 2000). That is, given that a system exhibits a certain kind of behavior, it asks how that behavior is brought about. Cummins' approach is well-suited to the psychological context mainly because it does not rely on identifying components of a system in order to explain one of its properties (Cummins 1983).

Cummins' account is pragmatic in the sense of understanding the ascription of functions as relative to the epistemic activity of explanation. As I have already pointed out before, Cummins takes it that the function of an entity is identified relative to a systemic property that we try to explain. Given this explanandum property, the *function* of a less complex property is identified in terms of input-output pairings (Roth & Cummins 2014, p. 7-8), where pairs whose output contributes to the exercise of the phenomenon to be explained *are* their functions in that context. Again, the contribution to the exercise of a complex disposition is thus everything there is to a particular function.

To provide an example: Let's assume that you are interested in the capacity of a human being to store and manipulate information over a short period of time. To adequately explain this capacity of human beings – also called "working memory" –, you refer to the model of Baddeley (2012), according to which it can be accounted for by the orchestrated workings of four sub-systems, that is, the "central executive" and three storage units, the "phonological loop", the "visuospatial sketchpad" and the "episodic buffer" (see figure C.5).

Take a closer look at what the central executive *does* to contribute to working memory: the central executive controls the flow of information to and from the three storage units. This is the *causal role* of the central executive. If this is true, then the *function* of the central executive in a human being is just that: taking in information, distributing that information among its so-called "slave systems" and controlling where the information flows *from* these slave systems. Importantly, the kind of information differs, depending on the subsystem. For example, the information fed to the visuospatial sketchpad is

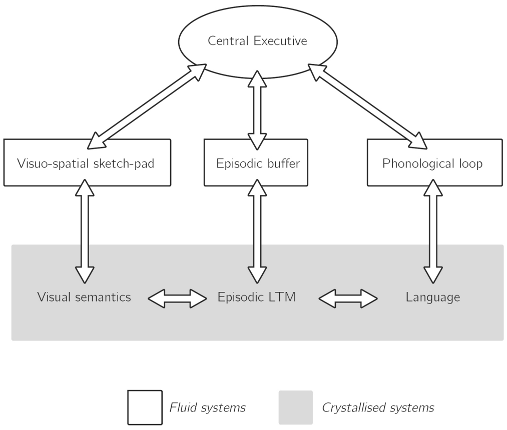

Figure 5.1: Model of the working memory as presented by Baddeley (2012).

spatial information, whereas the episodic buffer is fed information about events happening in the life of the person (Baddeley 2012).

The example of working memory highlight some of the positive features of Cummins' account: We can make sense of the relevant content of some explanations without needing to refer to the material manifestations of the functions in question and without being committed to a rich account of *function.* Instead, the explanatory content only consists in functional components of a disposition which are organized in a specific way.

One of the major criticisms levelled against Cummins' analysis is that it appears to be too liberal (compare, e.g. Couch 2019), since it does not allow to distinguish between those things that a particular system part[20] just *does* – mere side-effects, so to speak – and those things that are part of its function.

On Cummins' understanding, the function of something is given by its *causal role.* Is his concept of *function* also helpful in understanding what psychologists and psychotherapists mean when they refer to the *dysfunctionality* of certain beliefs or other features? I argue that this is not the case, even though Cummins' account helps to understand the explanatory power of models of mental disorders like the one by Beck & Bredemeier (2016) or Salkovskis et al. (1998) – that is, they emerge as *functional analyses* of complex dispositions.[21].

[20] Please think of "parts" here as not necessarily spatially individuated.

[21] Relatedly, Vosgerau & Soom (2018) have argued that mental disorders can be understood as dispositional properties.

On Cummins' understanding, if we intend to explain OCD, the *function* of system parts like *neutralizing actions* would simply be their causal role – that is, being brought about by misinterpretations of significance, resulting in further intrusive thoughts and enhancing the tendency for misinterpretation. I think that this stands in stark contrast to how most psychologists refer to neutralizing actions: That is, as a *dysfunctional* behavioral strategy. Thus, on Cummins' view, the dysfunctionality of this strategy cannot have anything to do with its causal role in the system of interest. But it *has*, since this strategy is referred to as dysfunctional partially *because* it is harmful. For this reason, dysfunctionality cannot be analyzed with Cummins' account.

According to the goal-contribution analyses of the concept, functions are causal contributions to *goals* (Boorse 2002, p. 68). Since this analysis is supposed to hold for living beings and artifacts in general, not only for human beings, the understanding of a system's "goal" is not given in cognitive or intentional terms. Instead, Boorse, making reference to the so-called "Sommerhoff–Nagel view" (Boorse 2002, p. 69), describes the following view of a system's *goal-directedness*:

> "[A] system S is 'directively organized', or 'goal-directed', toward a result G when, through some range of environmental variation, the system is disposed to vary its behavior in whatever way is required to maintain G as a result." (Boorse 2002, p. 69)

I take this to be a very intuitive understanding of goal-directedness: When a system is disposed to vary its behavior to get to a particular state no matter its environment, it seems plausible to say that this state is one of its goals.[22]

Now, how do we get from goal-directedness to functions? According to Boorse, this is how:

> "X performs the function Z in the G-ing of S at t if and only if at t, the Z-ing of X is a causal contribution to G." (Boorse 2002, p. 70)

In other words: Something performs a certain function, when, by doing what it does, it contributes to the goal in question.

[22] Note that this notion of "goal" is importantly different from the notion that I employed above, when referring to consciously held and explicit goals of human beings. These require many more conditions being met.

This is similar, and yet, substantially different from Cummins' account of functional analysis insofar as the goal-contribution analysis relies on actual goals of an organism to identify functions that are performed in an organism. By contrast, Cummins' analysis relies on identifying causal contributions of system components to some disposition that we intend to explain. As is pointed out quite regularly in the literature, this supposedly solves crucial problems of Cummins' view like the fact that it ascribes functions to too many things. These latter accounts either – in the case of selected-effects views – take the function of an entity to be that which it was selected for, or – in the case of goal-contribution analyses – that which contributes to the organism's goals (Boorse 2002).

The goal-contribution analysis of functions is interesting for me since it relates well to the claim of several of my interviewees that dysfunctionality is primarily about whether a certain thought, behavior or strategy of someone helps that person to attain her *goals*.[23]

Keep in mind the central features of the goal-contribution analysis, since I will use them later in developing my analysis of dysfunctionality. Now, I will switch to an account that will be of relevance in the following: that is, Wakefield's understanding of the term.

This understanding of *dysfunction* is at the heart of Wakefield's account of mental disorder. It has an etiological analysis of function at its heart – and thus, stands in tension to both the causal-role analysis and the goal-contribution analysis of function, since Boorse does not define "goal" evolutionarily. Wakefield, as I stated already, takes mental disorders to be *harmful dysfunctions* (i.e. Wakefield 1999). A dysfunction is present – or, in the clinical psychologist's vocabulary, a mechanism or process is dysfunctional – whenever a mental *mechanism* does not instantiate or carry out its *evolved* function.

The view of function that Wakefield exploits here is the etiological view of function. It has been characterized by Hardcastle as follows:

> "Roughly speaking, according to this view, a trait T has the function of producing effect E in some organism O if T contributed to the

[23] Let me just ignore for the moment that the statements of my interviewees seem to suggest that "goal" needs to be something stronger than what Boorse is pointing to in this quote. I take this to be feasible, because at the end of this chapter, I will argue that their notion of dysfunctionality really does *not* need to rely on a notion of *explicit* goals of someone, but that it is sufficient to think of the individual's *needs*.

> fitness of O's ancestors in virtue of doing E and T is heritable." (Hardcastle 2002, p. 145)

In the case of psychological function, we may take it that interpreting many actually harmless stimuli as signals of danger has the function of making human beings cautious, since being cautious contributes to the survival of human beings in the EEA, and (plausibly) interpreting harmless stimuli as harmful is heritable.

According to Wakefield, a *dysfunction* is present once the evolved function in question fails to be carried out. For example, think of the evolved function of the heart to pump blood. Once the heart fails to pump blood, a *dysfunction* of the heart is present. Wakefield takes this account and applies it to mental processes or mechanisms.

As we will see, Wakefield's concept of dysfunctionality does not align well with psychotherapist's understanding of the term. For example, he points out himself that there is nothing dysfunctional about SAD[24], on his account (Wakefield et al. 2005). According to him, the symptoms of SAD are the effect of a psychological mechanism that evolved as a response to the need for social connections as a resource for survival in the EEA. In this environment, individuals were best served by not risking their status in a group too easily, and thus, shying away from behaviors that would involve such a risk, like speaking in front of large groups. As long as this anxiety still allows the individual to engage in basic social interaction with familiar individuals, it does not count as dysfunctional on Wakefield's view. Thus, according to him, SAD is not a mental disorder as long as it only occurs in performance situations like speaking in front of audiences constituted of unfamiliar individuals, since these situations are "biologically plausible triggers" (Wakefield et al. 2005, p. 317).

This does not map neatly onto the clinical psychologist's and psychotherapist's way of thinking and speaking about dysfunctionality. The relevant literature on SAD shows that clinical psychologists both understand SAD as a disorder and they frequently refer to dysfunctional cognitions and dysfunctional behavioral strategies when speaking about this disorder (e.g. Hofmann 2007). While

[24] In his paper, Wakefield actually speaks of "Social Phobia", but I will stick with the more recent term of "Social Anxiety Disorder" from the DSM-5, as the diagnostic criteria for both disorders are virtually identical, and thus, Wakefield's argument also applies to SAD as characterized more recently (American Psychiatric Association 2013, p. 202-203).

this does not directly contradict Wakefield's account of mental disorder – at least if we don't take it to be a requirement for an account of mental disorder that it should cut along the same lines as our current classification systems of mental disorders[25], it *does* show that his and the psychologist's use of the concept do not map onto one another neatly.[26]

Now, I will draw the different strands together and finally offer the promised analysis of dysfunctionality, using the six criteria of adequacy I identified in my analysis of notions of dysfunctionality in psychotherapeutic practice. Furthermore, I will make use of selected parts of these philosophical accounts of function. More precisely, I will use Boorse's notion of *goals* and Wakefield's *harmful dysfunction* view.

5.5 Analyzing the Concept of Dysfunctionality

In this section, I will provide an analysis of dysfunctionality that agrees with the observation that, when clinical psychologists are asked what they mean by "dysfunctionality", they present descriptions indicating that they think of something being *detrimental to the well-being* of an individual or group. To begin, let me go back to those *criteria of adequacy* for a good analysis of dysfunctionality that I have developed over the course of my above discussion of how the concept of dysfunctionality is used in clinical psychology research and psychotherapeutic practice:

[25] Here, one might be inclined to point out that not being dysfunctional is not identical to not being a disorder. But importantly, on Wakefield's view, mental disorders *just are* harmful dysfunctions. That is, if what is classified as a mental disorder does not involve a harmful dysfunction, then the thing so classified is actually not a mental disorder, or so her argues (Wakefield et al. 2005).

[26] I do think that a philosophical account of mental disorder should roughly have the same results when it comes to the conditions under which we diagnose mental disorder – in part because licensing treatment is one of the most important functions of the ascription of mental disorder, which Wakefield seems to see differently (Wakefield et al. 2005). It is also important to note that SAD is, by many psychologists, held to be a rather *paradigmatic* case of mental disorders, as opposed to conditions like, for example, schizophrenia. But this is a topic too broad to be discussed here.

(DA_1) The notion of “dysfunctionality” should in principle be applicable to beliefs, behaviors, and strategies of human beings,

(DA_2) that typically differ between mentally healthy individuals and mentally ill individuals,

(DA_3) that lead to significant (subjective) *harm* in the individual holding the belief or carrying out the action/strategy,[27]

(DA_4) that are supposedly either non-goal-conducive or actively counteract the satisfaction of the agent’s goals,

(DA_5) that lead to a reduced (objective) *functional level*, that is, a reduced ability of the agent to take part in activities that we, as a society, deem important for someone to take part in (work, friendship or the like), and

(DA_6) that can be intervened upon in psychotherapeutic treatment, and they are introspectively available to the agent.

In the following, I will develop an analysis of dysfunctionality, based on observations presented before in this chapter and on specific examples that show how the notion of dysfunctionality is employed in practice. Finally, I will show how my analysis adheres to my six criteria of descriptive adequacy. If it does not adhere to a specific criterion, I will make explicit why I think that it does not have to accord with it. I will start with a first pass at an analysis and further refine it on the basis of individual counterexamples for these earlier passes at a definition. This method allows me to show, in each instance, what *exactly* is wrong with a particular notion, and it gives us a better grasp on how this vocabulary is *actually* used. Thus, it makes it possible develop a sufficiently precise account of dysfunctionality.

Let me start with a first pass at a characterization of what dysfunctionality might be and then see in the following why this does not yet lead us where we want to go. This first pass is rooted to the same extent in Wakefield’s comments, according to which the “colloquial sense” of dysfunctional is nothing more than harmfulness (Wakefield 1999, p. 376), in my above review of the literature as well as explanatory practices in psychotherapy and in my qualitative interviews:

[27] In his book on mental disorders, Bolton (2008) points out that the *harm* that is distinctive of mental disorder need not necessarily be harm that the individual herself suffers from. While I take this to be correct, I nonetheless believe that the notion of dysfunctionality is rather about harm that the agent (or agents) holding the belief suffer from.

> (D_1) Some x (that is, a belief, an action, a strategy) is dysfunctional for an agent a just in case x is causally relevant for significant harm in a.

This already captures an important feature that I wanted my account to have: I argued that my notion of dysfunctionality should be centered around harmfulness. What might not be immediately clear is why I added the proviso of "significant" harm. This is intended to allow for cases where someone's acts lead to a negligible reduction in well-being without having to count this individual as acting in a dysfunctional way. Think, for example, of someone drinking alcohol. Although this action is harmful, it is arguably not harmful enough to count as dysfunctional.

Note that this fits to the idea that something's dysfunctionality is not determined by the *correctness* of those mental states that are dysfunctional (or bring about the dysfunctional action/strategy), but by the content's *harmfulness* for an individual in her specific context. This is important, because building correctness into the notion would run into problems when we try to square it with the issue of *depressive realism*: the phenomenon that, in individuals with mild depression, the representation of oneself and one's future is systematically more accurate than in healthy individuals (Moore & Fresco 2012). Thus, it might actually be more functional for many individuals to have skewed beliefs.

Let me show that this first pass at an analysis of dysfunctionality is not sufficient by way of giving an example. Let's assume that Michaela suffers from SAD. In her case, this means that she experiences intense fear when speaking in front of groups, especially when her performance is evaluated. While she would like to pursue a PhD after finishing her studies in history, she does not consider herself capable to speak up in seminars and can hardly give presentations because she fears that people will spot her nervousness and hold her to be too stupid to pursue an academic career. Certain dysfunctional beliefs underlie her symptoms, most prominently, the belief that if she were to show signs of nervousness in public, people will exclude her socially and she will have to bury her career plans. This leads to significant harm, since it results in high levels of anxiety and might even lead her to decide against a career that she would actually like to pursue, thereby leading to further unhappiness.

Her strategy in dealing with this anxiety is to wear only long, black clothes to hide her physical reactions to the stress, that is, sweating and blushing, and to prepare her presentations extremely well, to have her manuscript finished weeks in advance and to then repeat it several times in front of the mirror, until she memorizes the text by heart. This strategy calms her down sufficiently, such that she can present in seminars without her nervousness being spotted by others.

At this point, you might ask: Why on earth would Michaela's strategy be dysfunctional? According to my narrative, she seems to be doing just fine, and her strategy *does* allow her to accomplish her explicit goal of giving presentations in seminars. Who could possibly have an issue with that? Well, psychologists *do* have an issue with this. Although Michaela's strategy allows her to achieve her short-term goal, it effectively makes it impossible to challenge her underlying, harmful belief(s). This reduces her well-being in the long run, since these beliefs continue to cause her harm. That is, both her strategy and the underlying belief are dysfunctional. The belief that showing signs of nervousness will make other people exclude her socially is dysfunctional, because it brings about anxiety, and, thereby, is harmful to her. Her strategy is dysfunctional, because it "immunizes" her harmful belief against being tested empirically. As we have seen, her belief has the structure of a conditional, and thus, it can only be empirically tested by *exposing* one's alleged imperfections to others. But Michaela's strategy is in place precisely *to avoid* exposing her flaws and mistakes. Thus, it effectively *maintains* the underlying, dysfunctional assumption that she will be made fun of or heavily criticized if she exposes herself.[28] Behavioral strategies like these that decrease the individual's anxiety in the short run, but effectively maintain the mental disorder by keeping the underlying problematic beliefs from being tested – so-called "safety behaviors" – are very common in individuals with SAD (Clark et al. 2005, p. 196-198).

Since she makes sure that she is never perceived as nervous when present-

[28] Of course, one might argue that there are certain contexts in which this belief is actually well-grounded, that is, where people *are*, in fact, heavily criticized when exposing themselves. Think of, for example, bullying of schoolchildren or of increasingly rare cases of aggressive academic discussions. Luckily, these contexts are quite rare. Furthermore, in contexts where flaws and imperfections *do*, in fact, lead to heavy criticism or the like, the individual might still reject the conclusion that this means that she is stupid and instead adopt the position that it is the other individuals in the situation who are acting in a morally problematic way. This latter position is more easily adopted if the context in question is one that can be left or if significant others are available who hold a different position.

ing in seminars, it is impossible for her to find out that, if she *were* to show signs of nervousness, she would not be socially excluded and treated as stupid. Thus, her strategy is dysfunctional in virtue of immunizing a dysfunctional belief against disconfirmation. In other words: Michaela's behavior is pragmatically rational relative to her background beliefs, but her strategy still is dysfunctional, because it has a substantive, negative effect on her well-being.

This suggests that further conditions must be added to my analysis so that it can deal with cases like the one I identified above.

I take it that the main problem with the case described above derives from the fact that my account does not yet tell us anything about time: that is, in its current form, my account would be consistent with claiming that a particular strategy is functional in the short term, but dysfunctional in the long run. In my understanding, this is not how psychotherapists reason, however: usually, they will simply refer to a strategy as dysfunctional if it has long-term negative consequences for the individual.

When does a belief, a strategy or a behavior have long-term, harmful effects, though? Slightly altering what my interviewees have said in the interviews, I take it that this is the case if the satisfaction of particular, relatively stable *needs* of the patient is causally counteracted by the belief, strategy or behavior in question. Thinking in this way about dysfunctional beliefs allows for a better understanding of where the *repeated* harmfulness originates: That is, if certain stable needs of someone are kept from being satisfied systematically, this explains why the individual is continually harmed. It also helps to understand what therapists search for in alternative beliefs or behavioral strategies: namely, the feature of allowing for such needs to be satisfied.

Let me present the next step towards my analysis of dysfunctionality:

> (D_2) Some x (that is, a belief, an action, a strategy) is dysfunctional for an agent a just in case x *causally counteracts at least one need n of a*, and produces significant harm for a as a result.

Why am I talking about *needs* instead of goals here, as my reliance on Boorse would suggest?

I think that this makes good on the intuition that isolated instances of significant harm do not count for judgements of dysfunctionality. Instead, what counts will only be the long-term, net effect of the action. This is important because, as we have seen, safety behaviors are pragmatically *rational* when regarded in the short term, and they would also arguably count as dysfunctional, if dysfunctionality did not have a certain time-criterion built in. I take it that understanding dysfunctionality as relative to particular, relatively stable needs of a person allows us to circumvent this problem. Plausibly, one of her underlying *needs* is to feel relatively safe in most everyday situations in the long run[29], and this is where the strategy turns out as dysfunctional: That is, while it decreases the amount of anxiety felt in the short term – in keeping with her explicit goals –, it runs counter to a need of Michaela's by increasing how much anxiety she feels in the long run, and thus, exposing her to a lot of harm.

Nonetheless, societal norms are still involved in determining whether a certain amount of harm makes the entity bringing it about *dysfunctional.* Thus, the context-dependency involved in fixing the threshold of "significant" harm in the concept of dysfunctionality remains, but I take it that referring to needs makes this slightly less problematic, since we often have a relatively good grasp on whether certain needs of individuals are satisfied in a particular situation.

The characterization I give here is inspired by the goal-contribution analysis of function, presented by Boorse (2002). In particular, it takes Boorse's notion of *goal.* Nonetheless, I take it that what the author refers to with "goal", if applied to human beings and when specific background conditions are added, rather serves to identify people's *needs.* After all, according to Boorse, systems are goal-directed just in case they are disposed to vary their behavior such that the result is achieved under variations of environmental conditions. In the case of human beings – by contrast to other organisms –, we can distinguish the individual's explicitly held goals from her potentially subconscious needs, and I take it that Boorse's account rather helps us to identify the latter.

To go beyond what Boorse suggests, this notion of a *need* should have particular additional conditions built in. For one, psychologists are arguably interested only in a specific level of description: For example, I am currently disposed to vary my behavior in such a way that I add a particular number of words to my thesis on any given day. But this does not indicate that, in doing

[29] It is important to note that, of course, this need is only one of a myriad of different needs that people have, but it is the need where this strategy turns out to be dysfunctional. While the agent probably *also* has the need to actually *be* safe, this is not the need that is at issue here, because I simply assume, for the sake of simplicity, that the latter need is actually satisfied.

so, I act in accordance to a need to write that particular number of words. Instead, I take it that we get at my actual needs by taking into account the behavior that I would show under different kinds of variations of my environment, even relatively far-fetched ones. Another relevant consideration is how I would react to not being able to reach the end state in question. If not being able to reach this state significantly reduces my well-being, then I am arguably acting in accordance to a need (compare, e.g. Brock & Miller 2019). Thus, I take it that the notion of need should roughly be understood like this:

> (N) An agent a has being in state s as a need n just in case a would vary her behavior under different (possibly counterfactual) environmental conditions in such a way that a achieves s, and if a does not achieve s, a experiences significant harm.

Of course, one might object that understanding harmfulness as being brought about by someone's needs not being met and then characterizing someone's needs by referring to harmfulness again is actually circular. But I think that levelling this criticism against my analysis would be to misunderstand the importance of being able to point out where the harm in question *originates* in the individual and being thereby put in a position to actually search for alternative strategies that allow for these needs being met.

To come back to the case in question, it would quickly be observable that my underlying needs are to earn money and to engage in a kind of work that gives me a sense of purpose. For these variations in environmental conditions to not become too extreme, psychologists usually take a set of historical, social, and economic as well as individual conditions as given.[30]

This allows us to distinguish between what individual patients explicitly and consciously aim at, and what they implicitly and often unconsciously *require* in order to remain or become mentally healthy, a distinction that is important for psychotherapeutic practice.

To present an example: If human beings are indeed disposed to vary their behavior in a way that allowed them to be members of a social group, then

[30] We may observe that my analysis bears a certain similarity to analyzing the meaning of specific model operators by means of possible-world semantics. Understanding the notion of "need" as, in a sense, relative to particular background conditions, for example, a certain society, a particular historical period or the like maps on the idea of different kinds of necessity such as physical necessity, logical necessity and the like (compare, e.g. Menzel 2017).

being a member of a social group plausibly counts as one of those things a human being's behavior is generally[31] directively organized towards, and if human beings do not achieve this state, it will cause them significant harm. That is, being a member of a social group seems to be an (evolved) *need* of human beings. This is implied by the claim of *obligatory interdependence* (Caporael & Brewer 1995) of human beings in the EEA. If someone acts in a manner that continuously undermines being a member of a social group, his behavior may – prima facie – qualify as dysfunctional. Insufficiently satisfying those needs that human beings have *as a species* makes several strategies or beliefs dysfunctional – just think of social withdrawal, a maintenance factor of depression: It makes sense to think that what makes it dysfunctional is that it impedes social contact, and social contact is something that human beings, as a species, are directively organized towards. Importantly, though, not all needs of individual human beings are generic in this sense. Instead, we may use my analysis of needs also to identify specific needs that only few individuals have or that are socially constructed. If someone is – under many different environmental conditions – disposed to vary her behavior in a way that she ends up in a monogamous, committed relationship, then we may assume that this is a need of hers.[32]

One may wonder whether it might not make sense to give an evolutionarily based meaning to dysfunctionality. This view would be very similar to my own, but assume that an agent's needs must be analyzed evolutionarily.

But I take it that an understanding of dysfunctionality along evolutionary lines severely underestimates the influence of societal factors on someone's needs: There are some individual needs that are not the outcome of evolutionary processes but that can, nonetheless, contribute to mental disorders.

Just consider a need that many people have, namely, to be in a committed, long-term, monogamous romantic relationship. It is not clear whether there is an evolutionary basis for this need – instead, there is some evidence according to which human beings did not evolve as a monogamous species and that

[31] This should be understood as a generic generalization, that is, a statement that, other than universally quantified statements, remain true even if not *every* member of the category has this as a goal (Leslie & Lerner 2016).

[32] Clearly, identifying needs is a tricky business, as identifying something as an instance of one and the same need may only be possible by describing the behavior in question in a very particular way. Usually, this will mean taking the intentional stance (Dennett 1971) and then going through different possible descriptions of a behavior until one finds one where a pattern can be observed.

mating and human relationships worked differently in other historical periods (Henrich et al. 2012). That is, the need for this particular *kind* of relationship – other than the more general need for deep interpersonal, committed relationships – seems to depend on raised in a society with a particular set of values. Nonetheless, for many people, those beliefs that actively run counter to their ability to actually engage in a long-term, committed, monogamous romantic relationship, are harm-inducing, and thus, *dysfunctional* beliefs.

Talking about *needs* instead of explicit goals gives dysfunctionality the relative objectivity that psychologists seem to assume it has. By "objectivity", I mean a feature's relative independence of potentially false beliefs of the individual. I take it that psychologists think dysfunctionality as in this sense more objective than the notion of irrationality because of statements like the one by Beck that I referred to above, who identifies the feature's *dysfunctionality* as problematic for the individual's mental health, not the feature's irrationality. It does so by allowing for such needs to be relevant for judgements of dysfunctionality that might not be introspectively available to the patient or that might only become so in the therapeutic process.

There is a further advantage to this, though. As I pointed out in the beginning of this chapter, it should in principle be possible for particular beliefs of an individual to be simultaneously dysfunctional and *not* pragmatically irrational. I argued that there should at least be some *conceptual* elbow room to allow for cases in which an individual values some other goal over the satisfaction of her needs, without necessarily having to count this person as pragmatically irrational. This elbow-room is provided by talking of needs instead of conscious goals here.

Consider cases in which someone values a particular ideal over her own well-being. One instance of this is valuing honor more highly than one's own well-being. This was common in some historical periods in Europe and North America, for example (LaVaque-Manty 2006). One of the consequences of this understanding of honor was the need to challenge somebody to a duel who had damaged it. As these duels were likely to end in the death of at least one of the participants, it seems plausible to say that someone who values honor as much as was common there did indeed value it higher than his own well-being. Still, while it does seem to be dysfunctional for the individual to believe honor to be that important – primarily because most other needs of the individual are secondary to her need for survival, this does not make all of the individual's actions by definition pragmatically irrational. But given his beliefs, it is, indeed, quite rational to behave in this way. To give a less extreme example, think of a woman who values the well-being of her romantic partner over her own. If this ordering counteracts a need of hers – like being safe –, it counts as dysfunctional. This is the case even if avoiding harm is not one of her explicit goals. Nonetheless, she need not be pragmatically irrational in caring more for him than for herself, given she has particular background beliefs and preferences.[33]

It is important to point out that this implicit relativity of a belief's or action's functionality to the agent's needs does, in reality, only hold for particular kinds of needs, namely those that can be satisfied without violating the needs of other individuals and very basic social norms. For example, most psychotherapists would probably not consider it functional behavior if a pedophile consumed child's pornography, even if – let's assume this to be the case – this individual's sexual preferences were such that his sexual needs could not be satisfied by anyone who was above a certain age.

This suggests that our analysis should be adapted as follows:

> (D_3) Some x (that is, a belief, an action, a strategy) is dysfunctional for an agent a just in case x causally counteracts at least one *socially and morally sanctioned* need n of a, and produces significant harm for a as a result.

There is something more to consider, namely, the fact that dysfunctionality is a *dispositional* notion: According to the theory underlying CBT, someone can have dysfunctional beliefs without these beliefs actually having *any* harmful consequences. This becomes clear when considering the fact that CBT assumes that dysfunctional beliefs are adopted at some point during the individual's childhood, remain within the individual's belief-system and only become *activated*, and thus, thought- and action-guiding later. The belief in question not having harmful effects despite being dysfunctional may either be due to the fact that the individual is not in the type of situation that would serve as a

[33] Clearly, this depends also on a certain ordering of different needs. Needs seem, in a certain sense, more important than others if the individual would have them in more, and stronger, variations of her environment. It seems like the need for survival that I referred to above is biologically inbuilt and a need that many other needs are geared towards.

stimulus for the disposition to be manifested, it may be due to the individual having further, more positive, core beliefs (Beck 1995, p. 21), or it is due to the fact that the individual carries out coping strategies that keep the harmful effects from occurring.

To understand how this looks like in practice, consider the following example: Someone, let's say Andy, has the dysfunctional belief that he always has to please other people. In one situation, he is continually lauded for his academic performance. In this situation, Andy feels good about himself – his dysfunctional belief, in my understanding, does not manifest in bringing about harmful effects. A little later, his life situation changes, and he finds himself in a much more competitive context. Suddenly, he gets only very little positive feedback, and professors start criticizing his work continually. Andy feels that he does not please them. As a result, his self-esteem diminishes. This is the kind of situation in which the dysfunctionality of his belief *is* actualized, thus leading to harmful effects he would not experience if he did not have this conviction as part of his set of beliefs.

This is roughly how clinical psychologists and psychotherapists think about dysfunctionality: According to CBT, a dysfunctional belief may also lie dormant in someone, that is, have no harmful effects, until it – in their terminology – is *activated* due to a stressful situation (i.e., Beck 1995, p. 15). In these cases, I would rather speak of the dysfunctionality not being *masked* anymore and thus *realizing* and producing harm (Choi & Fara 2016). It fits my way of thinking quite well that therapists and clinical psychologists have observed that individuals with dysfunctional beliefs may not show symptoms in case they possess sufficiently many and good resources, alternative beliefs or coping strategies that allow them to counteract the otherwise negative consequences of these beliefs (Beck 1995). The account we arrive at thus is the following:

> (D_4) Some x (that is, a belief, an action, a strategy) is dysfunctional for an agent a just in case x *would, given triggering conditions* c[34], causally counteract at least one socially and morally sanctioned need n of a, and produce significant harm for a as a result.

There are further important issues to consider here that have to do with the

[34] In the case of strategies and behaviors, the "triggering condition" simply is being carried out.

kinds of situations that are taken to be the default by mental health professionals: There might be beliefs or behavioral strategies that are not *actually* dysfunctional, but that still are involved in causing harm under very particular kinds of circumstances. This has to do with the simple fact that there are environments so harmful that almost anyone would get mentally ill in them – and those individuals that don't develop a mental disorder as a result count as the *exceptions* rather than as the rule. For example, repeated and excessive traumatization seems to have this effect, as was shown in a study on women who survived being held captive by the so-called "Islamic State", where roughly 60% of women who were raped more than 20 times developed *Posttraumatic Stress Disorder* (*PTSD*) as a result (Kizilhan 2018).[35] For the diagnostic criteria of PTSD, see appendix A.6.

What is noteworthy about this example is that, according to (D_4), those beliefs that individuals have and that brought them into this very situation – think of, for example, specific positive beliefs about the Islamic State that may be wrong, but not dysfunctional – and are, as a result, *causally relevant* for the counteracting of particular socially and morally sanctioned needs of the individual – would emerge as dysfunctional. But this seems highly implausible, since, for a belief to be dysfunctional, it arguably should contribute sufficiently much to the harmfulness in question. This is not the case for the beliefs in question here, though: instead, they become irrelevant for the harmful effects in question once the environment's effects are taken into account.

Even though one might think that the most promising way out of here is to require certain features of an environment for it to be relevant for judgements of a belief's dysfunctionality, I think that what psychologists are relying on here is actually something different. In fact, the problem here is akin to another one that I noticed when applying my account to Salkovskis' model of OCD.

If the conceptualization given so far were true, then it would seem that everything that intrusive thoughts depend upon causally in the system of interest should be classified as *dysfunctional*.[36]

[35] I think this shows that, even though psychotherapy assumes a certain amount of adaptation to circumstances that differ from the EEA, there are environments that simply differ too much from that which the human psyche would be evolutionarily set to be able to endure.

[36] For the time being, I will take "causal relevance" to be quite broad, allowing for several different entities to be causally relevant: On this understanding, both event (types) might be causally relevant for something, in case they, by actualizing a disposition, causally bring about the effect in question (intrusive thought), but also the dispositions themselves. This makes sense, because intervening on the dispositions of interest does lead to systematic changes in intrusive thoughts; thus, they satisfy

Here, an account of dysfunctionality as mere causal relevance for significantly harmful states would have the effect that *not only* the beliefs in question are dysfunctional, but the other system components *as well*, including *mood changes* and *attention & reasoning biases*. These system components are also causally relevant for certain states that are harmful to the individual. Think of mood changes, which may include severe anxiety. Although there are some clinical psychologists who would apply the adjective "dysfunctional" to these system components, it is not very common to speak of "dysfunctional biases" or "dysfunctional moods".[37] Note that most psychotherapists would not want to call those biases dysfunctional solely because they are on the causal path from the actual cause of harm to the harmful effect. Thus, in both of these cases, my analysis seems to give us the wrong result.

These two cases are alike, because both of them are a result of my account being overly inclusive. That is, in the second case, it wrongly assigns dysfunctionality to objects that simply lie on a causal path from the actual cause to the harmful effect. In the first case, it also includes too much, this is just not clear at first glance, since the actual cause of the harmful effect – the extremely hostile environment – is excluded by fiat by my analysis. Furthermore, both of these cases can be dealt with by making one single change to my account, namely, by including a criterion of *screening-off*. This is loosely based on the definition of screening-off that was originally invented to characterize particular kinds of probabilistic relationships. A particular event C is said to *screen off* an alternative event A from the alleged effect E just in case

$$P(E|A \wedge C) = P(E|C).$$

That is, if the probability of some event E, given two other events A and C is identical to the probability for E given C, then C screens off A from E (compare Hitchcock 2018). Or, put more simply, but also more crudely: Once we hold the effect of C on E fixed, the effect of A on E vanishes. Here, I will use this notion of screening off to characterize the harmful effect that one particular event has on another.

I will thus include a new clause into my characterization of dysfunctionality:

interventionistic accounts of causation along the lines of Woodward (2003).

[37] This is a rough estimate, based on GoogleScholar searches for "Obsessive Compulsive Disorder" and "dysfunctional x", where x was substituted by either "beliefs", "behaviors", "strategies", "biases" or "moods".

> (D_5) Some x (that is, a belief, an action, a strategy) is dysfunctional for an agent a just in case x *would*, given triggering conditions c, causally counteract at least one socially and morally sanctioned need n of a, and produce significant harm for a as a result. *The harmfulness of x for a may not be screened off by either the environmental conditions or by another belief, action, or strategy.*[38]

I think that with this, we have reached a sufficiently precise understanding of what psychotherapists see as a belief's, action's or strategy's *dysfunctionality.* But how does it relate to those features that I have listed above as requirements for such an account? Let me conclude by pointing out which ones it satisfies and which ones it fails to satisfy, and for what reason.

Starting at the beginning, my account *does* satisfy the condition of applying to all three objects of claims of dysfunctionality, that is, to beliefs, actions, and strategies. It does so by fiat: That is, I have built this requirement into it explicitly.

Secondly, are *only* those beliefs dysfunctional that *also* distinguish the mentally healthy from the mentally ill? I do not think so. This is because those dysfunctional beliefs, actions and strategies that are of relevance in psychotherapy will usually be those that distinguish between mentally ill and mentally healthy individuals. But there may still be individuals without mental disorders that have dysfunctional beliefs. As I pointed out above, psychologists assume that these individuals will have other beliefs or strategies that counterbalance their dysfunctional assumptions. For example, someone who has the dysfunctional belief that she is only worthy of love if she performs extremely well at her job may not develop any symptoms, if she exhibits a strategy of systematically working more than she has to, thus actually performing very well and getting a lot of positive feedback as a result.

Thirdly, I have built the requirement of harm directly into my account. It emerges as the outcome of someone's needs failing to be realized.

The fourth condition, that dysfunctional entities should not be goal-conducive, is also accommodated, although in a slightly changed form. While I do not talk about goals in the sense of explicit goals that I referred to when analyzing

[38] Although this analysis of dysfunctionality is clearly one that only deals with the dysfunctionality of an individual agent's belief, I take it that it can in principle be extended to cover cases in which multiple agents act dysfunctionally. For reasons of space, I do not cover this issue here.

rationality, I did indeed talk about goals in Boorse's sense, deriving a notion of needs from this. To distinguish from explicit goals of the individual and in keeping with how psychotherapists talk, I have called these objects *needs*.

Fifth, my account of dysfunctionality also satisfies the requirement of being consistent both with the idea of a reduced objective, even though socially constructed, functional level of someone, but this reduction in functional level is not necessarily introspectively available to the patient. Furthermore, the fact that most human beings agree in their needs to a certain extent – in their evolutionary needs, and in some of their socially constructed needs, if they have a similar cultural background – also makes sense of this idea of relative objectivity.

To reiterate, according to the fifth requirements, the beliefs, behaviors and strategies in question lead to a reduced *functional level*, that is, a reduced ability of the agent to engage in activities that we, as a society at a specific point in history, deem important or worthy for someone to take part in. A prototypical activity of this kind is work. This requirement is not directly part of my analysis, but I think that one can make the case that a reduced functional level of an individual that is specified via what a society takes to be indicators of a good life is very closely connected to the idea that a certain feature of an individual causally counteracts at least one socially and morally sanctioned need of hers. One might even argue that this reduction of an objective functional level is merely a heuristic that is used to as an estimation for when someone's needs are not met.

According to the sixth condition, dysfunctional entities should also be suitable targets for psychotherapeutic intervention. I take it that this is actually built into my definition already by virtue of restricting dysfunctionality to beliefs, actions, and strategies. In the last chapter, I have already presented some reasons of why therapists believe that they can intervene upon a patient's beliefs, and with actions or behaviors, this is even easier. Strategies are also good potential targets for intervention, since they are by definition something that the agent has chosen to do, and can thus also choose to change.

One interesting effect of my analysis for understanding psychotherapeutic practice is that, for most people who enter therapy, most of their dysfunctional beliefs will emerge *also* as pragmatically irrational. This explains why several talking therapeutic techniques aim at showing dysfunctional beliefs to be irrational, I take it. As I have hinted at in the last chapter, one of the central goals of psychotherapeutic treatment is to make the patient's health one of her top-level *explicit* goals. Once this is the case and the individual knows how her mental disorder is caused and maintained, she must, according to my analysis, be understood as both theoretically and pragmatically irrational in holding and acting in accordance to these beliefs.

There is a further consequence of it that is helpful in reconstructing certain aspects of psychotherapeutic practice. That is, actions count as dysfunctional just in case they *actively counteract* the satisfaction of the agent's needs, and as functional just in case they promote the satisfaction of them. Usually, dysfunctional actions are brought about in part by dysfunctional beliefs. Conversely, given certain background conditions such as the absence of intervening factors like coping strategies or certain positive beliefs, dysfunctional beliefs will bring about dysfunctional actions.[39] This implies that it is also possible for the functionality and dysfunctionality of someone's beliefs and actions to come apart. Just consider the example of someone who has a dysfunctional belief about needing to avoid making mistakes in public. Simultaneously, that individual carries within him the strong wish to take part in a theatre production as an actor. According to her dysfunctional belief, it would be advisable to not act according to her wishes. If she takes up acting in a theatre play anyway, she acts in a functional manner, despite the presence of her dysfunctional beliefs. By behaving in this way, she puts herself in a position to produce counter-evidence to the correctness of her deep-seated, dysfunctional belief. Generally, this holds for many different mental disorders. This fact is exploited relatively often in psychotherapeutic practice, for example, in more recent forms of psychotherapy like MCT (e.g. Fisher & Wells 2009).[40]

[39] One might wonder, though, about whether dysfunctional beliefs or dysfunctional actions are actually *primary*. That is, we might think that it does not make sense to think that a belief is dysfunctional without having harmful consequences *in the form of actions or open behaviors*. But I think that this is not quite right. Just consider the possible case of someone who has a particular belief about being inadequate that leads to negative automatic thoughts, and thereby, to sadness, loss of motivation, and the like. In principle, it would be possible for this individual to act in the same way as a healthy individual. Nonetheless, the belief in question clearly counts as dysfunctional, because it has an effect on the individual's cognitive processing that leads to significant harm by being in conflict with a socially sanctioned need of the individual – that is, the need to feel competent. In this case, a dysfunctional belief would have harmful consequences without influencing the open behavior and intentional actions of the individual.

[40] It seems that there is a significant difference between more "traditional" forms of CBT and newer forms of therapy such as MCT concerning how intimately related they take thoughts, emotions and behavior to be. While the more traditional forms of CBT appear to assume that someone's emotion is basically determined by his thoughts about a situation, MCT seems to suggest that one can have

Among other things, I think that this analysis, and, in particular, the distinction between pragmatic irrationality of beliefs and the a belief's dysfunctionality allows to differentiate between different processes in therapy. Think, for example, of someone's motivation for psychotherapy. Someone who does not have his mental health as an explicit, top-level aim will hardly listen to arguments according to which he should not carry out actions that are detrimental to his mental health. In doing so, the patient would not necessarily act pragmatically irrationally, even though he does actually damage his health. What *will* emerge as pragmatically irrational, though, is not making one's mental health a top-level, explicit goal. This is because the patient's mental health will either be a precondition for or it will substantially facilitate achieving many (if not most) of the other aims that the patient may have. This also accounts for the fact that, often, the individual will experience the need to change her priorities and top-level goals in such a way as to include her mental health. Thereby, she achieves a better mapping of her explicit goals onto her underlying, more objective needs.

5.6 Conclusions

In the last section, I have developed an account of *dysfunctionality* as used in clinical psychology and psychotherapeutic practice. To do so, I combined several general criteria that I derived from my analysis of both clinical psychological literature and psychotherapeutic practices with certain individual examples. I then pointed to several positive consequences of understanding dysfunctionality in this way. My notion of dysfunctionality differs in important ways from the two notions of (ir)rationality from the preceding chapter. By pulling these different notions apart, we can offer an enlightening reconstruction of important parts of therapeutic practice.

Furthermore, I take it that all of this can also tell us something about the concept of functionality that is at the heart of this. In my understanding, someone acts and/or thinks in a functional manner, if, according to her needs, given the restrictions of society and morals, the person enough of her needs

all sorts of thoughts about a particular kind of situation *without* necessarily having to have a certain emotional reaction to it. But I think that my account is, in fact, compatible with both views: If we assume that the sequence is not, as CBT would have it, situation → thought → emotion, but situation → thought → acceptance/rejection of thought → emotion, then all of this might be neatly fitted into one picture.

to reach relatively stable equilibrium state that she feels sufficiently content about. This then allows her to engage in other activities, such as to actively pursue her explicit life goals. This is actually very similar to something that Bolton points out when he identifies one notion of "normality" in psychology to be "behavior that is on balance beneficial to the agent [...], consistent with their needs and intentions" (Bolton 2008, p. 268). What I offer in addition to his understanding is, though, that I differentiate between behavior that is rational (behavior that fits the agent's *intentions*) and behavior that is functional (behavior that fits the agent's *needs*).

How does the irrationality of particular thoughts or behaviors and their dysfunctionality hang together, on my account? I take it that one of the primary goals of psychotherapy is to make the patient's subjective well-being one of their primary goals, if it does not yet have this status. Once this is the case, there is an alignment of personal goals and well-being, and thus, virtually *all* dysfunctional cognitions, behaviors and behavioral strategies emerge as pragmatically irrational. Once they are thus understood, they may be challenged by one of the different kinds of disputation techniques to show its irrationality.

When considering causal models like the two exemplary models I presented before, it becomes clear that it is part of the nature of many mental disorders that it is particularly hard for the patient to undertake those actions that would break the circle of feedback loops between different symptoms to reduce her suffering. This is the case for *both* of the cases that I have focused on in this dissertation, namely, depression and OCD. In the case of depression, one of the actions most likely to at least reduce the disorder's symptoms is so-called *behavioral activation* (compare, e.g. Cuijpers et al. 2007): that is, pursuing activities that the individual, under normal circumstances, likes to engage in. But those very activities are those that are particularly hard for the patient to pursue, given her state. Similarly, in the case of OCD, not carrying out neutralizing behaviors as a response to intrusive thoughts is what would be most likely to break the circle. But due to the association of intrusive thoughts and neutralizing actions that then lead to a short-term reduction in anxious emotions, it can seem almost impossible to the patient not to carry out the action in question.

Partially for this reason, we might say that, although individuals with mental disorders who possess a full understanding of their problems are, to a certain extent, *responsible* for preventing or lessening future symptoms, they nonetheless do not count as *fully* blameworthy if they do not manage to act in accordance to their better judgement as to what would be reasonable to do, given their considered preferences. This is because blameworthiness does not only depend on one's actions in a particular sort of situation, but also on the individual's dispositions and capabilities.

Chapter 6

Conclusions and Open Questions

6.1 Overview

In this chapter, I will revisit those questions that I posed in the beginning, intending to give a clear and exhaustive overview of what we have learned over the course of this investigation.

I started out with the very generic question of how clinical psychologists and psychotherapists explain mental disorders. In particular, I was interested in the interplay between practical applications of explanatory models and their theoretical formulation. I developed the hypothesis that explanatory practices in psychotherapy influence the form and content of models of mental disorders, which I argued for in chapters two and three. There, I did two things: Firstly, I investigated the processes of model construction and further development. Secondly, I reconstructed how practitioners explain mental disorders in psychotherapy, which I think of as the *application* of these models.

On the basis of qualitative interviews with psychotherapists, I developed an account of how cognitive-behavioral therapists explain their patients' mental disorders in practice and which aims guide these explanatory practices. To gain a better understanding of the models at issue and to provide an enlightening account of particular aspects of therapeutic practice, I analyzed the concepts of *rationality* and *dysfunctionality* that are at play both in these more theoretical models of mental disorders and in therapy. Let me now recapitulate the questions that were at the heart of this dissertation.

In the beginning of this dissertation, after providing an explication of the notion of a mental disorder that is presented in the DSM-5, I offered an analysis of the content of two explanatory models of mental disorders, one model of depression and one model of OCD. There, I identified several noteworthy features of these models. That is, they are intended as *causal* models, they mention mainly *normal* psychological processes, as Bolton (2008) points out, they contain folk-psychological vocabulary and they make use of vocabulary surrounding the notion of *function*.

As I pointed out, these models are not *only* the outputs of clinical psychological research, but also used by psychotherapists to explain mental disorders in practice. As I hypothesized on the basis of informal conversations with psychotherapists in training, these explanatory practices are governed by very specific, practical aims. This leads to explanations that are substantially different from what we would expect of mere epistemic practices. Thus, I found it necessary to distinguish between

1. explanatory aims of the models and
2. aims of using the models as part of a particular explanatory practice

While (1) refers to the target phenomenon that stands in need of explanation, (2) refers to aims that the speaker wants to achieve in providing an explanation. One may also talk of *epistemic* aims in contrast to *pragmatic* or *practical* aims. It is both theoretically interesting and important to distinguish between these two kinds of aims, since they might pull into different directions. As I have argued extensively over the course of this investigation, explanatory models of mental disorders can only be used for the purpose of achieving particular pragmatic aims if they represent the mental disorder in question in a specific vocabulary and as having particular features. While this by no means contradicts that such models may provide proper explanations of these phenomena, it arguably has other effects, such as favoring models that employ folk-psychological vocabulary.

In contrast to much existing work in philosophy of psychiatry (compare, e.g. Murphy 2010, 2006, Cooper 2007), I decided to focus my attention on the more *pragmatic* and *practical* side of this distinction. For one, I believe that aims of explanatory practices actually have relevant effects on the form and content of explanatory models. Thus, I take it that the interconnectedness of psychotherapy and clinical psychology research has so far received too little attention from philosophers. For the other, I consider it philosophically worthwhile to analyze psychotherapy in its own right, as an essentially *discursive* practice.

Let me now give an overview of how I carried out this analysis over the course of the preceding five chapters.

In chapter one, based on my description and subsequent reconstruction of two models of mental disorders, I put forward the hypothesis that specific noteworthy features of these models might also be due to do with the work that they do in practice. That is, I pointed out that simultaneously describing these disorders in causal, folk-psychological, and functional vocabulary might have something to do with the importance of the context of application for these models.

In the second chapter, to provide some evidence for this thesis, I investigated how the two exemplary models of interest were first constructed in clinical psychology. I argued that these models are created on the basis of evidence that derives primarily from the context of application. I backed this up with personal reports from Aaron Beck and Paul Salkovskis and with the fact that both of them combined research and psychotherapy. The kinds of changes that these models underwent in practice make it very likely that their development over time was influenced by pressures from therapeutic practice. In doing so, I did not provide an exhaustive analysis of how research and application processes interact in the construction and development of these models. Nonetheless, given the statements of the two scholars that I have referred to, I take it to be extremely plausible to regard these two processes as strongly interwoven.

The third chapter took this primacy of the context of application as a starting point, focusing on how these models are actually *used* in psychotherapy to explain mental disorders to patients. This investigation was based on the results of six qualitative interviews with psychotherapists. I drew on the notion of aims of explanatory practices that I had introduced in the very beginning, identifying three main aims of explanatory practices:

1. limited attribution of responsibility, that is, presenting the patient as not at fault for developing and still suffering from her condition(s)
2. understanding the patient as still having *agency*, and

3. putting forward possibilities for intervention that may be realized in the psychotherapeutic process

I pointed out that there are several strategies by which these aims are achieved, namely,

1. pointing to stable (dispositional) features of the patient that she could hardly control and that contributed to her falling ill,
2. representing mainly *normal*, *functional* and *reasonable* psychological processes (what I called *normalization* and *rationalization*) as bringing about these symptoms,
3. pointing to epistemically accessible features that causally maintain the symptoms in question and that are – at least in principle – under the agent's control.

I argued that a good explanation of this fit between model and aims of explanatory practices is that the structure and content of these models is influenced by those three aims of explanatory practices. Particularly in conjunction with the finding from the second chapter, namely, that the context of application is essential for the construction and further development of these models, this appears plausible. Finally, I drew the reader's attention to another effect of these model's being used in the psychotherapeutic context, that is, the occurrence of a phenomenon akin to *looping effects*, arising from the fact that these explanations of mental disorders are presented to individuals who suffer from them, to make these patients adopt a certain view of themselves. Changing how patients view themselves arguably changes the way they behave.

This change in behavior, caused by a *deliberate* intervention on the patient's self-perception, might seem to provide new evidence for the correctness of the model. While these feedback effects usually have positive effects on patients by, for example, giving them more hope, they might also have negative epistemic or practical consequences by resulting in a merely *apparent* fit of the patient and the intervention at issue. It may happen that the correct explanation of someone's mental disorder and the therapist's explanation come apart without the patient or the therapist noticing this – in part because of such feedback effects.

In the fourth chapter, I started out by providing a more detailed analysis of the processes of rationalization and normalization in psychotherapy that I first hinted at in chapter one. I provided an account of those normative concepts that are relevant for understanding how (and why) these strategies are carried out in therapy. These are concepts of statistical normalcy and, more importantly, of rationality. I argued that, to understand how psychotherapists rationalize, we have to distinguish between two notions of rationality that I called

1. theoretical rationality and
2. pragmatic rationality.

While the theoretical rationality of beliefs depends on whether they are well-grounded in the relevant empirical evidence and the extent to which they are compatible with the agent's background beliefs, "pragmatic rationality" describes the degree to which those beliefs can be expected – from the agent's perspective, that is, given her background beliefs – to help her achieve her considered goals. By relativizing these two kinds of rationality to particular points in time, the psychotherapist can represent certain beliefs of hers as both (relatively) rational to adopt and irrational to act upon at the current point in time. I argued that pulling these two understandings of rationality apart also allows for a more precise understanding of other parts of psychotherapy over and above explanatory practices.

In the fifth chapter, I provided an analysis of functional concepts that are relevant in psychotherapeutic practice and that are, furthermore, at the heart of those models of mental disorders that I discussed so far. I focused on the notion of *dysfunctionality* that is used very frequently by psychotherapists in practice. I showed how it differs from the two notions of rationality that I discussed in the preceding chapter, most importantly, from the notion of *pragmatic* rationality that it is very similar to. I argued that, while the pragmatic rationality of someone's beliefs or – derivatively – her actions is dependent on the agent's *explicit* goals and her set of background beliefs, the dysfunctionality of her beliefs depends on her personal *needs*. As soon as her beliefs or actions *actively counteract* the satisfaction of her needs, they should be understood as dysfunctional. In this way, dysfunctionality becomes the more *objective* notion in comparison to rationality, which depends on the patient's background beliefs and her explicitly held goals.

This allows for an enlightening reconstruction of psychotherapeutic practice: According to it, CBT is based upon making the agent's *needs* explicit to her. If the agent's mental health or psychological well-being is assigned the status of one of her top-level explicit goals – which can often be taken as a given, provided that she entered psychotherapy at her own free will[1] and usually knows that, without a certain level of well-being, she will not be able to pursue most of her other goals[2] –, what is dysfunctional also becomes pragmatically irrational for the patient. Understanding the process of therapy in this way also makes good on the observation that many patients feel the need to change their priorities substantively when undergoing psychotherapy. At the same time, it allows to understand patients as relatively – theoretically *and* pragmatically – rational agents also *before* entering therapy. At that point in time, their top-level explicit goals were simply different, thus leading to distinct, and sometimes false, judgements of what would be reasonable to do. By making dysfunctionality the more *objective* of the two notions – since it relies on facts about the individual's *needs* –, it is nonetheless possible to describe these patient's beliefs and actions as dysfunctional, just like psychotherapists often do.

I think that the results of my work may not only be interesting to philosophers of science, but also to psychotherapists and clinical psychologists. Although one might say that what I do in chapters three to five is primarily to make implicit, procedural knowledge of practitioners explicit, this is far from trivial. Especially, it may be considered quite problematic from the perspective of psychotherapy that psychology students usually begin their therapeutic work with the naïve picture of individualization in mind. This implies that every single psychologist has to learn anew how to individualize these models in order for explanations to be as useful as possible to patients in therapy. From private conversations with several therapists in training, I know that they found it difficult to adapt their explanations to the mental abilities of the patients, and that they also did not find it easy to present the disorder in question in such a way that the patient feels simultaneously taken seriously in

[1] This makes the psychotherapeutic treatment of individuals that are forced to be in a mental health institution a topic for another time.

[2] And, if it is not, I have presented some reasons to think that psychotherapists very often intend to convince their patients to make it so.

his suffering and gets the impression that he can nonetheless change something about his symptoms.

Being aware of the difficulties involved in individualizing and knowing which factors are taken into account by experienced therapists might help a little bit with this issue. Even more, it might be useful for practitioners to consider my account of how their disputation techniques depend on particular judgements of rationality.

6.2 Tying up Loose Ends

Now that I have presented an overview of what I have done in this investigation, I will tie up some loose ends by reconsidering issues that I have brought up and did not resolve so far.

To start, I would like to remind the reader of one of the central questions of the preceding chapters: How do psychotherapists simultaneously represent the patient as not to blame for developing her disorder and relatively rational, while still being responsible for intervening on her symptoms? In my understanding, what psychotherapists do during therapy is to make their patients understand themselves in this particular manner by pointing out that, while it might be the case that the patient did not reason perfectly in the past and acquired several *harmful* beliefs, there is nothing about this which would make her blameworthy. The patient is not blameworthy because, given her past experiences, she did as well as she could. In reconstructing the agent's process of belief formation as relatively rational, psychotherapists show that she simply *did not know any better* or did not have enough control over her behavior to actually do better. This fits two conditions of moral responsibility that most philosophers agree on, namely, (1) the *control condition*, which asks whether the agent was acting freely, and (2) the *knowledge condition*, which asks whether the agent knew what she was doing (Rudy-Hiller 2018). Let me try to make these last points a bit more explicit.

I take it that what is going on when therapists explain their patient's mental disorders is that two perspectives about the patient's suffering are taken simultaneously, resulting in two distinct, but interwoven, narratives: one narrative that centers around concepts of *causality* and *dysfunctionality*, and another one that focuses on *rationality*, *functionality* and *normality.* Then, given the patient's individual background (including her personality structure, her rough genetic make-up and the like) and beliefs,

1. therapists point to dysfunctional beliefs and behavioral strategies that, the relevant facts about the patient's situation being equal, would have made it hard for most individuals to think and act in a more health-preserving manner (that is, she could hardly have done better), while
2. therapists tell a coherent, *meaningful* story about why this person adopted these dysfunctional beliefs and strategies, and in this story, the patient emerges as, by and large, relatively *rational* and *normal*.[3]

I take it that both perspectives are required in order to represent the patient in the particular way that is intended by psychotherapists in practice: That is, the patient did not know that adopting the beliefs or strategies in question would have these harmful effects, and she adopted them for good reasons. According to the therapist's narrative, once she adopted those beliefs, they "developed a life of their own", contributing substantively to both the development and the maintenance of her symptoms.

Now, given this analysis of explanatory practices in psychotherapy, how should we evaluate them? Are these actually *good explanatory practices*?

Most importantly, we need an account for when explanatory practices count as *good* that we may use here for the purposes of evaluation. Firstly, we may think that explanatory practices in psychotherapy are good explanatory practices just in case they satisfy a certain philosophical theory of explanation. Secondly, we may understand "goodness" here as nothing more than *utility* for the purpose at issue. The question then becomes whether these practices actually help to achieve the more general aim of psychotherapy – that is, contributing to the patient's well-being. Thirdly, we may also consider them to be good practices in the sense of practitioners' being *morally* justified to engage in them. In this case, therapeutic explanatory practices count as good if the therapists in question *ought to* carry them out, given a particular moral framework. Finally, one might think that all of these dimensions should be

[3] Again, some of this is similar to what Bolton & Hill (1996) describe. Nonetheless, I think that I have something to add to their (very comprehensive) analysis of causal explanation in psychiatry and psychology by showing how these two perspectives are, in fact, used in therapeutic practice for achieving particular pragmatic aims.

involved in judging whether the explanatory practices in question are actually good.

Each of these questions, to be answered in a satisfactory manner, would require a much more detailed treatment than the one I can provide here. Thus, I will provide only a few remarks on them here.

Do these explanatory practices count as good explanatory practices according to philosophical accounts of explanation? Since explanation is often understood as having relatively strict success conditions – like being in a position to derive the explanandum from the explanans logically, as according to the deductive-nomological theory of Hempel & Oppenheim (1948) –, *being an explanation* already implies that the entity in question exhibits specific epistemic virtues. Thus, it suffices to ask whether these practices may be regarded as *proper explanations.* The answer to this question depends crucially on the philosophical theory of explanation one favors.

Nonetheless, it seems clear that, according to accounts with very strict and formal success conditions like the DN-account of explanation or more recent, mechanistic understandings, these explanatory practices will not count as proper explanations. According to the first, that is, because they do not refer to universal generalizations, but to particular events that occurred in the history of that very individual. According to the second, they cannot be so understood, because, most importantly, due to the lack of mention of spatial components, they do not describe genuine mechanisms.[4]

But even on an account like the one presented by Cummins (1983), these practices are arguably insufficient. That is, because they are too sketchy and too *partial* to fully explain someone's mental disorder by decomposing it into underlying sub-dispositions. Only in extremely rare cases will a psychotherapist's explanation of her patient's mental disorder actually amount to decomposing the condition at issue into sub-dispositions whose outputs amount to the realization of the explanandum disposition.[5] Thus, these practices would only count as explanations on comparatively *deflationary* accounts of explanation.

[4] Of course, they could view those as explanation sketches or schemas. But this would imply that these explanations gain in quality the closer they get to actually describing mechanisms, that is, the more spatial components are added to them. Since clinical psychologists seem to have relatively little interest in describing actual, spatial components of mechanisms, I do not think that explanatory practices in clinical psychology is best understood as mechanistic explanation.

[5] Of course, this presupposes reconstructing the mental health professional's talk of individual events and strategies as dispositions, but I do not take this to be particularly problematic.

But explanatory practices in psychotherapy do not actually *aim* at giving the patient a maximally precise or maximally correct understanding of the phenomenon in question. Even though there is, as I have discussed before, a certain amount of realistic representation not only in models of mental disorders, but *also* in these explanations that are tailored to the individual, the practices in question are not evaluated as *better* by therapists when they represent more realistically. Instead, the value of explanatory practices is, to a large part, determined by how *useful* they are for the patient and how much they contribute to the psychotherapeutic process.[6] This means that, even though these explanatory practices do not qualify as proper *explanations*, they may nonetheless be good explanatory practices. Understanding them in this way requires, I take it, to make the individual patient's improvements in his ability to deal with and potentially reduce his symptoms the main success criterion for explanatory practices.

This is closely connected to the aim of *utility*. Are these explanatory practices actually useful for the patient in this sense, and thus, *good*? The answer to this question depends heavily on what we mean by these practices "being useful for the patient". My first understanding would be that these practices are useful for the patient just in case they contribute to his well-being. As I have pointed out, it is one of the primary goals of psychotherapy to serve this function, and at least according to the therapists that I interviewed, their explanatory practices actually contributed to the therapeutic success by changing the patient's self-concept. Thus, if these therapists are correct in their understanding of their patients' therapeutic progress and their gains in well-being, then these explanatory practices actually emerge as *good* from the point of view of their *utility*. Since therapists usually get feedback directly from their patients, it seems plausible to suggest that they would be in a good position to judge this issue – but then, this should rather be decided by future, *empirical* research.

At first, one might think that studies on the relative effectiveness of different kinds of psychotherapy would settle this issue. But problematically, effectiveness studies usually only compare complete psychotherapeutic interventions with one another, which allows to draw conclusions about the effectiveness of psychotherapeutic explanatory practices only to a limited degree.[7]

[6] This has been pointed out repeatedly by my interviewees.

[7] A possible exception are so-called "dismantling studies" that identify the precise contribution of different parts of a specific therapeutic intervention to the therapeutic effect (see Papa & Follette

What about the *moral sense* of goodness? As I have hinted at in previous chapters, I take it to be a fascinating and important question whether explanatory practices in therapy are morally permissible. This is particularly so for what seem to be questionable therapeutic practices. As I noted before, several psychotherapists pointed out to me that they sometimes presented their patient's condition in a more positive light than they held it to be justified by the available evidence. This raises the question whether dishonesty can sometimes be permissible in therapeutic practice, and if so, under which conditions. I will come back to this issue later.

To conclude, I take it that, given what we have learned over the course of this dissertation, it is hard to tell whether the practices in question are to be evaluated in a positive way. But this is relatively unproblematic, since psychotherapists do not actually intend to provide perfectly precise explanations. We have seen that, from an epistemic standpoint, these practices are rather questionable. If we are interested in these explanatory practices' *utility*, they emerge – on the basis of the available information – as good explanatory practices. This is particularly true when we adapt our criteria of epistemic goodness in a way that takes into account that the aim of psychotherapists usually is to give patients the possibility to intervene. To do so, the account of the patient's disorder has to be tailored to the patient's mental capacities. The verdict on the third question, whether these explanatory practices are *good* in a moral sense, is still out.

Now that I have covered explanatory *practices* in psychotherapy, what about the explanatory *models* psychotherapists utilize in providing an understanding of their disorder for patients? Note that these models may still be explanatory, even if therapeutic explanatory practices are not good. This is particularly so, since explanatory practices in psychotherapy do not adhere to the model of simple individualization. Other than psychotherapists, who often favor pragmatic considerations over epistemic ones, researchers in clinical psychology are not bound to the same degree by practical considerations.[8] For these models, the question of epistemic quality is much more important than for these practices. Should we think that these models provide genuine *explanations*?

2015).

[8] To make matters more precise, an individual may, when working as a clinical psychology researcher, take epistemic considerations more seriously, while giving center stage to practical considerations when working as a psychotherapist.

In the following, I will show that both exemplary models of mental disorders that I have discussed over the course of this dissertation may be understood as explanations according to the theory of psychological explanation provided by Cummins (1983). That is, they may be understood as providing *functional analyses* of the target systems.

If mental disorders are complex dispositions or clusters of symptoms that are explained constitutively, then an explanation of a system's *property* might consist in a *functional analysis* of that system (Cummins 1983).

According to Cummins, explaining a disposition by functional analysis involves decomposing it into a number of *more basic* or *less problematic* dispositions which are individuated by their *function* in the system of interest (Roth & Cummins 2014).[9] Cummins' account of explanation is *constitutive* as opposed to *subsumptive* (Cummins 1983). By "subsumptive explanation", the author means explanations of events via general laws and background conditions such as the deductive-nomological model (Hempel & Oppenheim 1948) or the interventionistic account of explanation (e.g. Woodward 2003).

By contrast, Cummins tries to account for explanations of properties – mainly, that is, of dispositions. According to him, the latter are explained by showing *how* the disposition *works* (Cummins 2000). To do so, a property analysis focuses on the capacity's components and their organization. The simpler dispositions referred are of such a kind that their programmed manifestation accounts for the manifestation of the explanandum disposition (Cummins 1975, p. 759). Presenting these sub-dispositions reveals the *functional structure* of the system (Cummins 1975, p. 758). By "functions", Cummins means *causal roles.* He claims:

> "[...] *functional analysis* [...] operates at a level of abstraction that identifies analyzing properties in terms of what they *do* or *contribute* [to the

[9] In his papers, Cummins usually speaks about "capacities" instead of "dispositions". This seems to rest on the view that "capacity" is the more general term than "disposition". I pursue a different strategy here, for the following reasons: For one, I think that "disposition" is the more general term as opposed to "capacity": While capacities are usually understood as *abilities* of someone or something, *dispositions* seem to simply be *tendencies* of people or objects to show a specific behavior under some circumstances. For the other, in the case of cognitive clinical psychology, it seems more reasonable to speak about dispositions than about capacities, for the simple reason that many phenomena of clinical psychology might very well be characterized as a dispositions, but not as a capacities. One example is the tendency of depressed individuals to react to different stimuli in a negative manner: this might be conceptualized as a disposition, but intuitively, it is not a *capacity* of the individual living with it.

> exercise of a systemic property], rather than their intrinsic constitutions." (Roth & Cummins 2014, p. 779, my italics)

Thus, he claims that the *function* of an entity is given by input-output pairings or *regularities* (Roth & Cummins 2014, p. 779), where those pairs whose output contribute to the exercise of the explanandum disposition are the relevant functions. That is, *both* the explananda and the explanantia, on his account, are properties.[10] This means that *all there is* to a specific disposition in the context of a functional analysis is this contribution to the exercise of the complex disposition that we are trying to explain.

Thus, Cummins-functions are those causal roles which the system components play in accounting for the (complex) disposition of interest (e.g. McLaughlin 2001, de Jong 2003).

A positive feature of Cummins' account is that we can make sense of the fact that psychologists refer to certain models of mental phenomena as *explanations*, even though the models do not represent spatial components.

His analysis thus seems to be applicable to the models that we are interested in, which I would like to illustrate with the example of Salkovskis' model of OCD. Plausibly, this explanatory model accounts for OCD by identifying simpler sub-dispositions of it whose actualizations amount to a manifestation of the explanandum disposition. Thereby, it clarifies the abstract causal *design* of the condition.

The complex disposition to experience symptoms of OCD can be partitioned into the sub-disposition to react to intrusive thoughts with misinterpretations of significance, and these misinterpretations of significance (that is, a perception of danger and responsibility) serve as a stimulus for further sub-dispositions: attention and reasoning biases, counterproductive safety strategies, mood changes, and neutralizing actions. Their respective manifestations in reaction to the perception of danger and responsibility causally lead to

[10] Sometimes, the way Cummins' and others talk about causal role functions suggests that really, functions are specified by a subclass of the disposition's *effects*. I take this to be incorrect, since Cummins (1975, p. 758) claims the following two things:

1. Functional statements imply statements about corresponding dispositions.
2. Attributing a disposition to an object is asserting that the behavior of the object is subject to a "certain lawlike regularity".

This means that functions imply generalizations about a certain effect or output that occurs in response to a particular kind of stimulus. Thus, our notion of function has to incorporate more information than merely that the bearer of the function has specific effects.

further intrusive thoughts in the individual. These sub-dispositions are plausibly individuated via their causal role in the system of interest, namely, their causal contribution to what the complex disposition in question *does*: reacting to intrusive thoughts with obsessions and compulsions.

In other words, it makes sense to think of this model of Salkovskis et al. (1998) as one that, to explain a complex disposition, presents simpler sub-dispositions, their operations and their organization. Together, the operations of the simpler sub-dispositions result in the same effects as the realization of the complex disposition – even if it does not *only* do this.[11] The explanandum arguably becomes more understandable through this partitioning into simpler phenomena and their interactions.

Thus, if we take Cummins' view on psychological explanation to be one that provides plausible conditions of adequacy for explanations, it would seem that this model – and arguably, the model of MDD presented by Beck & Bredemeier (2016) as well – emerges as a (reasonably) *good* explanation. I take this to be plausible: After all, functional analysis seems to capture what is explanatory about several other models that we can find in psychology and that are considered to be good explanations of their phenomena by researchers in the discipline – think of, for example, Baddeley's model of working memory (e.g. Baddeley 1992). Furthermore, Cummins' understanding of how psychological explanation works is more plausible for those models at issue here than certain rival accounts, when compared to mechanistic accounts of explanation (e.g. Bechtel & Abrahamsen 2005, Glennan 2002, Machamer et al. 2000).[12] In particular, Cummins' analysis of explanation in psychology shows that the explanatory power of these models lies in showing *how something works*, and, in the process of doing so, reducing the complicated activity of a whole into more simple activities that occur in an organized fashion. Although mechanistic accounts of explanation are extremely powerful in capturing explanation in other disciplines, most of them appear to require mechanisms to have spatial components. As should be clear from the previous chapters, this is not how clinical psychological models of mental disorders are usually set up – and it

[11] Instead, it *also* puts forward what might be understood as a causal explanation for the occurrence of the phenomenon, by referring to preceding, early experiences and activating events. But this point is of rather little importance here.

[12] Strictly speaking, this holds only for those accounts that take mechanisms to necessarily have (spatial) parts or components. I allow myself to be a bit imprecise here since most recent accounts of mechanisms in science have this condition built into them.

also is not what most psychologists seem to aim at with their models. Relatedly, refining psychological models does not appear to progress from functional analyses – or "mechanism sketches", as they are called by proponents of mechanistic accounts of explanation – to complete representations of mechanisms, as one would expect if, for example, Piccinini & Craver (2011) were right. I agree more closely with the arguments of Stinson (2016) to the effect that psychological explanations might not actually map directly onto more detailed neurological models. Like Stinson, I take it that these models have explanatory value nonetheless.[13]

Even though these models satisfy Cummins' account of explanation, we may think that there are independent epistemic problems with how these models are formulated that have more to do with their *justification*. I have repeatedly pointed out over the course of this dissertation that, since these models are constructed on the basis of evidence from the context of application and used in that very context, certain *feedback effects* might occur. That is, since explanatory practices in therapy are geared towards generating a particular self-conceptualization in the patient, they may change the patient's behavior even independently from explicit intervention on behaviors in therapy. This was pointed out by my interviewees when comparing the effects of medical models of mental disorders and psychological models on the patient. These changes in the patient's behavior may, when observed by researchers in the discipline, be incorporated, and thus, change the model in question.

When the patient improves merely as a result of being presented with a particular explanation of his difficulty, this might be taken as evidence for the correctness of the model, when, in fact, the patient merely changed as a result

[13] Of course, Piccinini and Craver might argue that explanations in psychology *should* actually progress from functional analyses to complete representations of mechanisms and that these clinical psychological models of mental disorders are at best how-possibly explanations. With this, we have reached a point where I would disagree with them on a quite fundamental level: Piccinini and Craver seem to think that a philosophical account of explanation like the mechanistic account that is developed with specific scientific disciplines in mind and that seems very well-established there, can also be used for very different disciplines to evaluate their explanations. On their account, it does not matter too much how researchers actually work in practice. I disagree with this idea (similarly with how I disagree with Wakefield's idea that a philosophical account of mental disorder can rule out what one has previously taken to be a prime example of a mental disorder). I take it that a philosophical account should, first and foremost, try to take seriously how a certain scientific discipline operates and try very hard to be descriptively accurate. Of course, this does not mean that a philosophical account of a particular phenomenon like explanation cannot also rule out certain cases that researchers take to be explanations as non-explanatory. But the fact that large parts of psychology do not comply with how they think about explanation should be reason enough to seriously doubt whether their account should actually be extended to this discipline.

of the explanation.

Not all of these changes are brought about by the respective therapist or psychiatrist[14] *intentionally*, and these changes of behavior do not necessarily rely on presenting the patient with a correct model of their disorder, either. To see this, consider the fact that (according to my interviewees), providing clients with medical models of their mental disorders often results in patients refraining from taking over responsibility for changing their behavior. In the case of the serotonin model of depression, this change is brought about by a model that is, to our current knowledge, actually incorrect (Lacasse & Leo 2005).[15] Similarly, psychiatrists do not *intend* to bring about such inertia in their patients – their explanations just often seem to have this effect, because they seem to suggest that the patient cannot intervene on the disorder herself.

That is, the success of a therapeutic treatment does not necessarily constitute good evidence for the correctness of the underlying models. But how problematic is this, really, from an epistemic point of view?

To reiterate, the focus of application may result in an inadequate focus on particular entities that make it easier to explain the disorder to the patient in a *useful* way. Nonetheless, such feedback effects may be counteracted by evidence from independent research projects.

How many of these do actually exist? For one, a substantive amount of evidence that is not directly derived from the therapeutic context has accumulated both for Beck's model of depression (see, e.g. Disner et al. 2011, Beck & Bredemeier 2016) and for Salkovskis' model of OCD (see, e.g. Salkovskis 1999). This may counteract the potentially problematic fact that these models are constructed in the context of application. Taking the example of depression, there are also investigations of depressed individuals that study their information processing faculties, biases and potential neurological correlates of the phenomenon, for example. Such studies arguably are not vulnerable to the same kind of feedback effect that I have described with regards to explanatory practices. For example, the criticism that Coyne (1985) levels against various

[14] Although I am aware that distinguishing between psychotherapists and psychiatrists in this way is a false dichotomy, I use it here to highlight a contrast that has been pointed out by several of my interviewees between what they called "medical" approaches to mental disorder and "psychological" approaches. Of course, many psychiatrists are, in fact, psychotherapists as well – the most obvious example of this being Aaron Beck.

[15] This fact does not seem to keep practitioners from using it, as I have experienced when working as an intern in psychiatric hospitals.

models of depression can be understood as the outcome of such independent research practices.

But we may think that the potentially counteracting effect of these independent studies is somewhat limited. This is because, currently, research in clinical psychology appears to focus more on the effectiveness of treatments, not so much on investigating the accuracy of the underlying, theoretical models. While it is relatively hard to back this up with quantitative measures, it is an impression that I got when studying clinical psychology that was shared by one of my interviewees, himself a professor of clinical psychology:

> "On the one hand, I believe, that we have yet understood far too little about the mechanisms of mental disorders, on the other hand, one can say: This problem is so complex that we cannot wait with the development of therapeutic interventions until we can *really* treat those phenomena that we have to treat in our everyday practice. This means, on the one hand, this has to go hand in hand, we have to try to understand those mechanisms better, and in parallel, we have to find solutions in the here-and-now for our patients [...] I share your impression that currently, [...] in the academic setting, that psychotherapy has gotten a very high weight, therapy research, psychotherapy. This is why I believe that currently, the pendulum has deflected a bit into that direction, that one thinks a lot about treatment, without, or [that it] has fallen a little bit from view: What is it that one treats? The question of mechanisms and models, which we have just discussed, it has fallen a little bit from view."[16]

Of course, this only serves as a first indicator that there might be too little research done on those explanatory models when compared to treatments. In particular, even if this is the state of the discipline *currently*, we have already seen in chapter two that quite some evidence has accumulated for earlier versions of the two models of Beck & Bredemeier (2016) as well as Salkovskis et al. (1998). Investigating the import of these studies for those models would be a worthwhile topic for future research.

Connected with these considerations is the question whether these models of mental disorders *should* actually rely as heavily on folk-psychological theorizing

[16] Please do not focus too much on the fact that this professor refers to "mechanisms" here. When I asked him about this term, he pointed out that he used it to refer to causes or causal processes.

as they do. This question that has occurred in debates in clinical psychology as well, with many researchers arguing that models of mental disorders should not utilize this kind of vocabulary, partially because it would then be more in line with talk in other sub-disciplines of psychology. Among other things, using folk-psychological vocabulary was regarded as a sign for the theoretical inferiority of these models or even the discipline as a whole when compared to the rest of psychology (compare, e.g. Teasdale & Barnard 1993).

But is this a plausible way to think about these models? I take it that this question can be approached by taking one of two very different perspectives on clinical psychology. Firstly, we may think that clinical psychology is a *basic science* for psychotherapy, that is, that clinical psychology provides the theory behind therapeutic endeavours without being heavily invested in them. To a certain extent, we can understand medicine like this. But secondly, we may think that clinical psychology is an *applied* discipline, more akin to certain branches of engineering.[17]

If the former is the case, we may think that these explanatory models should actually not utilize folk-psychological theorizing about mental disorder as heavily, because it would be most important to possess an epistemically good model of the phenomenon in question before intervening. If, on the other hand, clinical psychology is to be understood more along the lines of certain branches of engineering, we may instead think that what matters for judgements about whether these models are actually *good* models is rather whether they are *useful* in the context of application. Similarly to disciplines like engineering, clinical psychologists would surely demand a certain amount of theoretical soundness and, especially, fit to the empirical reality from their models. But they might also disregard a certain lack of theoretical precision, if these models are particularly helpful in treating patients. As I already stated in the beginning, I believe that clinical psychology should indeed by understood as an applied discipline akin to engineering. This actually fits relatively well to the following statement of Teasdale & Barnard (1993, p. 7):

> “Beck [...] and his colleagues [...] have outlined a theoretical account of the origins and role of negative thinking in the aetiology of depression. [...] it is, avowedly, a *clinical* rather than a *scientific theory.* [...] the

[17] Of course, this is an extremely simplified take on these two disciplines. While I am aware of this problem, this rough distinction is good enough for the point that I want to make here.

> main purpose of the theory is to *guide the clinician in understanding and treating patients* rather than to provide a detailed exposition articulated in precise theoretical terms. [...] presentations of the model have tended to be relatively imprecise, to vary from one statement to another and to have shifted in their emphasis over time.'

My only correction would be, to put it simply and with a very common saying: "It's not a bug, it's a feature!". Indeed, I take it that being a clinical theory necessarily contradicts also being a scientific theory.

One question that appeared time and again when writing this dissertation was whether, and if so, how, my analysis of models of mental disorders and explanatory practices in psychotherapy relates to different accounts of what mental disorders *are*. This became particularly clear when I dealt with Wakefield's harmful dysfunction account in the preceding chapter. Let me thus quickly comment on this. Throughout this dissertation, I emphasized that my analysis of the concepts of rationality and dysfunctionality in psychotherapeutic practice is, neutral concerning the question of what mental disorders are. As I pointed out, the explanatory and therapeutic strategies that I describe may simply be the currently best way of intervening on mental disorders without representing what *actually* goes on in mental disorder. In that case, it would still be interesting as an analysis of a practice that has exerted a strong influence on research in clinical psychology. Nonetheless, my analysis squares better with some accounts of what mental disorders are than with others. That is, if psychotherapeutic practices were good evidence for which account of mental disorder is to favor, then the results of this investigation should be understood as supporting accounts that stress the normativity and social construction inherent in concept of mental disorder and health, or as supporting pluralist understandings of the term (compare, e.g. Bolton 2008). My analysis squares less well with the widespread harmful dysfunction view (Wakefield 1992) or purely naturalized accounts of mental disorder (compare, e.g. Boorse 1975)

Now that I have shown how some open questions may be answered, let me shift the focus slightly, discussing open questions that might be worthwhile to investigate in the future.

6.3 Open Questions for Future Work

I will begin this section by presenting some open empirical questions first before laying out the more substantial philosophical issues that were raised, but that I did not investigate in depth in this dissertation.

My analysis of the work that explanatory models do in psychotherapy raises several intriguing empirical questions. For one, it seems important to empirically investigate whether those self-conceptualizations that therapists want to evoke in their patients are in fact evoked by their practices. Furthermore, it would be interesting to compare the effects of conceptualizing mental disorders of patients according to so-called "medical" models of mental disorders and of explaining them according to the clinical psychologist's models.

Furthermore, a topic that I have been able to allude to only in passing is the question of how the need to intervene on the basis of explanatory models interacts with those pressures that these models are subject to from pragmatic aims arising in therapeutic practice. While it seems plausible that the need to intervene pulls in the direction of a realistic depiction of actual causal factors that are operative in mental disorder, there are several questions to ask about the extent to which this is the case: For one, the knowledge that clinical psychology has generated so far about the superiority of cognitive-behavioral therapeutic interventions over other psychotherapeutic interventions – think of psychoanalytic or psychodynamic interventions (compare, e.g. David et al. 2018) – is partial, and often only singles out the *complete* therapeutic intervention as superior to another in alleviating the patient's symptoms. The superiority of CBT may in fact be due to factors that are external to its theoretical models, like, for example, the kind of attitude practitioners of this therapeutic school are urged to take towards patients. That something like this may actually be the case may be suggested by the fact that, even across several therapeutic schools, so-called "therapeutic common factors" (compare, e.g. Tracey 2003) account for a much larger part of the therapeutic effect than factors that are specific to certain therapeutic schools (see, e.g. Lambert & Barley 2001). Add to this the observation that there are so-called "supershrinks" (see, e.g. Okiishi et al. 2003) in every form of therapy, and one may become very skeptical about how much of the difference in outcome between distinct therapeutic interventions is really due to theoretical differences instead of being due to more generic differences in the respective practitioner's attitude. Thus, I take it to be important for future research to investigate the extent to which pragmatic aims of explanation and theoretical aims are in conflict. While the fact that there are therapeutic common factors does not explain why CBT is superior to other forms of psychotherapeutic treatment, it guides our attention to the fact that other factors over and above the allegedly better explanatory models of mental disorders might account for this superiority.

Relatedly, there are differences between the evidential standards that the construction of models of mental disorders has to adhere to in comparison to the evidential standards that novel interventions are held to. It seems that the latter standards are much higher, at least if we take classical *hierarchies of evidence* to be good standards for the quality of evidence (see, e.g., Oxford Centre for Evidence-based Medicine 2009).

Additionally, there are conceptual issues that are in need of more attention than I have been able to spare here. For example, I have only hinted very briefly at the issues surrounding concepts of normality that are at play in models of mental disorders and in therapeutic practices. I think that it would be a worthwhile topic for future research to use existing philosophical work on notions of normality in psychiatry (for example, by Bolton 2008) and analyze more precisely when and for which kinds of (therapeutic) purposes these different notions are employed.

Relatedly, one might ask to which extent judgements of normalcy that are important in psychotherapeutic explanation are reducible to judgements of the trait being evolutionarily adaptive. This is suggested by the fact that many psychotherapists use the fact that particular features of a patient *also* have an evolutionary function as a *means* to normalize that patient's experience and behavior. I think that this only makes sense in one direction, but not the other: While it does make sense to suggest that everything with an evolved function is also normal, it does *not* make sense to think that everything that is normal also has an evolved function. Think of the example of calculating ability that is discussed by Lilienfeld & Marino (1995) in criticizing Wakefield's account of mental disorder as harmful dysfunction of an evolved mechanism. The authors use the example of dyscalculia, a mental disorder that impacts an individual's ability to manipulate numbers in particular, to argue against the harmful dysfunction analysis. Dyscalculia would not qualify as a mental disorder according to Wakefield, since there is not *one* underlying, evolved function that is damaged. Instead, the ability to calculate is, according to the authors, an instance of *adaptively neutral exaptations*, that is, "features not originally shaped by natural selection, but that are by-products of adaptations [...]" (Lilienfeld & Marino 1995, p. 412). Nonetheless, it is statistically *normal* to be in principle able to handle numbers and to perform calculations. Similarly, there are other features of individuals that are statistically normal in today's society, and that are classified as (parts of) mental disorders if they are malfunctioning, but that are not the outcome of evolution. To achieve a full understanding of them, a much more thorough investigation of different notions of normality would be required.

One issue that I have hinted at above and that is in need of more attention is the issue of *dishonesty in psychotherapy.* As I have alluded to several times, both my interviewees and several other therapists mentioned to me in private conversations that they sometimes omitted details about their patient's disorder and presented the expectable therapeutic progress or the disorder itself in a more positive light than they actually thought to be appropriate. Although this seemed to be an important topic for several of the therapists I interviewed, I have found very few detailed treatments of this and similar topics in the philosophical literature. It would be fascinating to evaluate how those practices should be judged from the perspective of moral philosophy.

Such questions interact in interesting ways with issues surrounding what I called "feedback effects": Consider a case in which a therapist presents the expectable therapeutic progress in a more positive light than might be warranted by the evidence. This is something that some of my interviewees admitted to do, and it stands in tension to the idea that some of them expressed, namely, that psychotherapeutic practice is less epistemically unjust than other medical professions in the sense of Fricker (2007), which also brings up issues of paternalism.

Importantly, the description of this progress is only positively skewed *when compared to a world where that positive account has not been given.* That is, the psychotherapist's idea seems to be precisely that, by presenting this account of the situation, they will actually bring this progress about. This means that this kind of deception can be regarded as in certain regards similar to the administration of placebos: That is, it is to be expected that, through administering a certain drug or representation, a particular effect is brought about, while the mechanisms that bring this change about are *not* those that are explicitly described or implicitly suggested by the medical or psychological authority in question, but rather by psychophysiological mechanisms in the patient that are triggered by this description (Lichtenberg et al. 2004). Coming back to Kraines' criteria of explanations in therapy, we may say that describing the expectable therapeutic progress in a more positive light than would be correct, brings about hope in the patient, which very often leads to more progress than there would have been in the absence of this positively skewed description. One might even suggest that the therapist's description – if it is a prediction, after all, and must not be understood more as an *announcement of a decision* (compare Hampshire & Hart 1958) – is not deception at all.

More precisely, it might be understood as the therapist announcing his decision to do whatever he can to make the patient recover, or as declaring to work towards the patient's recovery as a team. Representing the current condition differently than one actually understands it seems to be more problematic morally.

There was another case that I discussed, which cannot even in principle be framed in this way. What is represented in a different light here is not so much the future, expectable therapeutic progress, but the facts of the matter *as they are currently.*

The case of interest is one where the patient's condition is represented differently than the therapist sees it. Usually, this means to present the patient's condition either as less severe than it actually is, or to present the patient as less blameworthy for the condition than he actually is. One example of this is the case of NPD. The rationale that was given by my therapists to justify not presenting narcissistic patients with the full picture of their mental disorder was that most therapists took it to be likely that the patient in question would either discontinue psychotherapy altogether, or that they would severely damage the therapeutic relationship, which they took to be an important factor of therapeutic change. Usually, they described the difference between the full picture of the patient's disorder and their presentation of it as one that concerned the patient's *culpability* for her difficulties. According to them, it is partially built into the phenomenon of NPD that the patient himself is somewhat responsible for the negative reactions he receives from other individuals.

Since this is substantially at odds with how the patient regards himself, it will lead to profound cognitive dissonance in the patient to tell him this. The patient will, understandably, try to avoid such negative evaluation of himself, and thus, probably discontinue therapy. I think that it is an important question to answer whether this actually makes the therapist ethically justified in proceeding this way.

One matter that will be decisive for this judgement is whether this should be understood as the therapist intentionally telling her patient a falsehood, or merely *omitting* certain facts of the matter (compare Mahon 2016). This is particularly important, since those therapists that brought up this issue in my interviewees were quick to add that they would try to correct this representation of the patient's condition over the course of therapy – mainly because real therapeutic process often depends on a correct account of the patient's difficulties, which is in keeping with the fact that those explanations are also intended as providing the patient with the means to intervene on their disorders. I take it that NPD is only one particularly striking example of this, having to do with the fact that the kinds of behavior characteristic of the disorder are negatively evaluated by most people. Importantly, questions of deception and dishonesty in psychotherapy are not specific to this case and may also arise for other mental disorders. One kind of case in which they will almost always arise is when providing the patient with a full understanding of her disorder stands in tension to the improvement of her mental health. In these cases, psychotherapists often found it justifiable to not tell their patients the full truth. Future research should focus on when these cases occur, whether they are morally justified, and if not, what to do about this.

Lastly, there is the issue of *feedback effects* that I described by relying on Hacking's notion of looping effects. This is worthy of further investigation. In particular, it should be analyzed in more detail whether the models that are affected by those feedback loops that I described actually *stabilize*. That is, it is important to investigate whether the models actually *only* provide further evidence for their own correctness through these loops, but do not *move around*, as Hacking (1995) described. Clearly, both results would be interesting. If these models stabilize, what remains is the problem of those models apparently producing evidence for their own correctness. Interestingly, one may think that this is more of a problem for medical models of mental disorders, since psychological models of mental disorders appear to allow for more possibilities for intervention than medical models. If interventions based on these models actually work, then, it would seem, these models have gotten something right. Think of how these models work: They function by telling the patient that she really *can* intervene on her mental disorder by cognitive means, that she still has some *agency*. If this actually puts the patient in a position to change something about her mental disorder, then she actually *has* as much agency as the psychotherapist claims, or at least, she *acquires* this amount of agency. If true, it would seem that psychological models of mental disorders only show feedback effects that are relatively unproblematic.

In this chapter, I have given an overview of what I have done over the last five chapters. In this investigation, I argued for the importance of the context of application for the construction of models of mental disorders. On the basis of five interviews with practitioners, I argued that explanatory practices in psychotherapy should be understood as serving particular pragmatic aims. These aims can only be reached because the models in question have particular noteworthy features that I described in the first chapter. Furthermore, psychotherapeutic practice implicitly relies on distinct concepts of (ir)rationality and (dys)functionality that, as I have argued, should be kept distinct. Keeping them distinct allows, in turn, for an enlightening reconstruction of psychotherapy. Above, I pointed out that a number of fascinating question arise from this investigation, some of which might best be tackled within psychology, while others lend themselves best to a philosophical approach.

Bibliography

Allan, L. G., Siegel, S. & Hannah, S. (2007), 'The sad truth about depressive realism', *The Quarterly Journal of Experimental Psychology* **60**(3), 482–495.

Allen, C. & Neal, J. (2019), Teleological notions in biology, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', spring 2019 edn, Metaphysics Research Lab, Stanford University.

Alvarez, M. (2017), Reasons for action: Justification, motivation, explanation, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', winter 2017 edn, Metaphysics Research Lab, Stanford University.

American Psychiatric Association (2013), *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*, American Psychiatric Publication, Washington, DC.

American Psychological Association, Division 12 (1996), 'About clinical psychology', *American Psychological Association* . **URL**: *https://web.archive.org/web/20150401061600/http://www.apa.org/divisions/div12/aboutcp.html.* Last accessed on 09/30/2019.

Andreasen, N. C. (1985), *The Broken Brain: The Biological Revolution in Psychiatry*, Harper & Row, New York, NY.

Baddeley, A. (1992), 'Working memory', *Science* **255**(5044), 556–559.

Baddeley, A. (2012), 'Working memory: Theories, models, and controversies', *Annual Review of Psychology* **63**, 1–29.

Barrett, L. F. & Bliss-Moreau, E. (2009), 'Affect as a psychological primitive', *Advances in Experimental Social Psychology* **41**, 167–218.

Barth, J., Munder, T., Gerger, H., Nüesch, E., Trelle, S., Znoj, H., Jüni, P. & Cuijpers, P. (2016), 'Comparative efficacy of seven psychotherapeutic interventions for patients with depression: A network meta-analysis', *Focus* **14**(2), 229–243.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C. & Vohs, K. D. (2001), 'Bad is stronger than good', *Review of General Psychology* **5**(4), 323–370.

Bechtel, W. & Abrahamsen, A. (2005), 'Explanation: A mechanist alternative', *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* **36**(2), 421–441.

Beck, A. T. (1963), 'Thinking and depression: I. Idiosyncratic content and cognitive distortions', *Archives of General Psychiatry* **9**(4), 324–333.

Beck, A. T. (1964), 'Thinking and depression: II. Theory and therapy', *Archives of General Psychiatry* **10**(6), 561–571.

Beck, A. T. (1967), *Depression: Clinical, Experimental, and Theoretical Aspects*, Harper & Row, New York, NY and Evanston, IL and London, England.

Beck, A. T. (1979), *Cognitive Therapy of Depression*, Guilford Press, New York, NY.

Beck, A. T. (2002), Cognitive models of depression, *in* E. T. D. Robert L. Leahy, ed., 'Clinical Advances in Cognitive Psychotherapy: Theory and Application', Springer, New York, NY, pp. 29–61.

Beck, A. T. (2005), 'The current state of Cognitive Therapy: A 40-year retrospective', *Archives of General Psychiatry* **62**(9), 953–959.

Beck, A. T. (2008), 'The evolution of the cognitive model of depression and its neurobiological correlates', *American Journal of Psychiatry* **165**(8), 969–977.

Beck, A. T. (2019), 'A 60-year evolution of cognitive theory and therapy', *Perspectives on Psychological Science* **14**(1), 16–20.

Beck, A. T. & Bredemeier, K. (2016), 'A unified model of depression: Integrating clinical, cognitive, biological, and evolutionary perspectives', *Clinical Psychological Science* **4**(4), 596–619.

Beck, A. T. & Haigh, E. A. (2014), 'Advances in cognitive theory and therapy: The generic cognitive model', *Annual Review of Clinical Psychology* **10**, 1–24.

Beck, A. T. & Hurvich, M. S. (1959), ‘Psychological correlates of depression: 1. Frequency of “masochistic” dream content in a private practice sample’, *Psychosomatic Medicine* **21**(1), 50–55.

Beck, A. T. & Valin, S. (1953), ‘Psychotic depressive reactions in soldiers who accidentally killed their buddies’, *American Journal of Psychiatry* **110**(5), 347–353.

Beck, A. T. & Weishaar, M. (2011), Cognitive Therapy, *in* D. Wedding & R. J. Corsini, eds, ‘Current Psychotherapies’, 9 edn, Cengage Learning, pp. 276–309.

Beck, J. S. (1995), *Cognitive Therapy: Basics and Beyond*, Guilford Press, New York, NY and London, England.

Beiser, H. R. (1984), ‘An example of self-analysis’, *Journal of the American Psychoanalytic Association* **32**(1), 3–12.

Blaney, P. H. (1977), ‘Contemporary theories of depression: Critique and comparison’, *Journal of Abnormal Psychology* **86**(3), 203–223.

Bogen, J. & Woodward, J. F. (1988), ‘Saving the phenomena’, *The Philosophical Review* **97**(3), 303–352.

Bolton, D. (2008), *What Is Mental Disorder?: An Essay in Philosophy, Science, and Values*, Oxford University Press, New York, NY.

Bolton, D. & Hill, J. (1996), *Mind, Meaning, and Mental Disorder: The Nature of Causal Explanation in Psychology and Psychiatry*, International Perspectives in Philosophy and Psychiatry, Oxford University Press, Oxford, England.

Boorse, C. (1975), ‘On the distinction between disease and illness’, *Philosophy and Public Affairs* **5**(1), 49–68.

Boorse, C. (2002), A rebuttal on functions, *in* A. Ariew, R. C. Cummins & M. Perlman, eds, ‘Functions: New Essays in the Philosophy of Psychology and Biology’, Oxford University Press, Oxford, England, pp. 63–112.

Borsboom, D. (2008), ‘Latent variable theory’, *Measurement* **6**.

Borsboom, D. (2017), ‘A network theory of mental disorders’, *World Psychiatry* **16**(1), 5–13.

Borsboom, D. & Cramer, A. O. (2013), ‘Network analysis: An integrative approach to the structure of psychopathology’, *Annual Review of Clinical Psychology* **9**, 91–121.

Braun, J. D., Strunk, D. R., Sasso, K. E. & Cooper, A. A. (2015), ‘Therapist use of socratic questioning predicts session-to-session symptom change in Cognitive Therapy for depression’, *Behaviour Research and Therapy* **70**, 32–37.

Bringmann, L. F. & Eronen, M. I. (2018), ‘Don’t blame the model: Reconsidering the network approach to psychopathology’, *Psychological Review* **125**(4), 606–615.

Brock, G. & Miller, D. (2019), Needs in moral and political philosophy, *in* E. N. Zalta, ed., ‘The Stanford Encyclopedia of Philosophy’, summer 2019 edn, Metaphysics Research Lab, Stanford University.

Bundespsychotherapeutenkammer (2019), ‘2015 bis 2019’, *BPtK Spezial* pp. 1–16. **URL**: *https://www.bptk.de/wp-content/uploads/2019/03/bptk_spezial_2015_2019.pdf.* Last accessed on 07/09/2019.

Burckhardt, L. (2015), *Why Is Landscape Beautiful?: The Science of Strollology*, Birkhäuser, Basel, Switzerland.

Busch, F. N., Rudden, M. & Shapiro, T. (2016), *Psychodynamic Treatment of Depression*, 2 edn, American Psychiatric Association.

Butler, A. C., Chapman, J. E., Forman, E. M. & Beck, A. T. (2006), ‘The empirical status of Cognitive-Behavioral Therapy: A review of meta-analyses’, *Clinical Psychology Review* **26**(1), 17–31.

Campbell, W. K. & Sedikides, C. (1999), ‘Self-threat magnifies the self-serving bias: A meta-analytic integration’, *Review of General Psychology* **3**(1), 23–43.

Caporael, L. R. & Brewer, M. B. (1995), ‘Hierarchical evolutionary theory: There is an alternative, and it’s not creationism’, *Psychological Inquiry* **6**(1), 31–34.

Carel, H. & Kidd, I. J. (2014), ‘Epistemic injustice in healthcare: A philosophical analysis’, *Medicine, Health Care and Philosophy* **17**(4), 529–540.

Carnap, R. (1959), *Induktive Logik und Wahrscheinlichkeit*, Springer, Vienna, Austria.

Carrier, M. (2011), Knowledge, politics, and commerce: Science under the pressure of practice, *in* M. Carrier & A. Nordmann, eds, ‘Science in the Context of Application: Boston Studies in the Philosophy of Science, Vol. 274’, Springer, Dordrecht, Netherlands, pp. 11–30.

Carrier, M. & Nordmann, A. (2011), Science in the context of application: Methodological change, conceptual transformation, cultural reorientation, *in* M. Carrier & A. Nordmann, eds, ‘Science in the Context of Application: Boston Studies in the Philosophy of Science, Vol. 274’, Springer, Dordrecht, Netherlands, pp. 1–7.

Choi, S. & Fara, M. (2016), Dispositions, *in* E. N. Zalta, ed., ‘The Stanford Encyclopedia of Philosophy’, spring 2016 edn, Metaphysics Research Lab, Stanford University.

Churchland, P. M. (1981), ‘Eliminative materialism and the propositional attitudes’, *The Journal of Philosophy* **78**(2), 67–90.

Clark, D. A. & Beck, A. T. (1999), *Scientific Foundations of Cognitive Theory and Therapy of Depression*, John Wiley & Sons, New York, NY.

Clark, D. A., Cook, A. & Snow, D. (1998), ‘Depressive symptom differences in hospitalized, medically ill, depressed psychiatric inpatients and nonmedical controls’, *Journal of Abnormal Psychology* **107**(1), 38–48.

Clark, D. M. (1986), ‘A cognitive approach to panic’, *Behaviour Research and Therapy* **24**(4), 461–470.

Clark, D. M., Crozier, W. & Alden, L. (2005), ‘A cognitive perspective on Social Phobia’, *The Essential Handbook of Social Anxiety for Clinicians* pp. 193–218.

Cooper, R. (2007), *Psychiatry and Philosophy of Science*, Routledge, London, England.

Couch, M. B. (2019), ‘Causal role theories of functional explanation’, *The Internet Encyclopedia of Philosophy* . **URL**: *https://www.iep.utm.edu/.* Last accessed on 08/21/2019.

Coyne, J. C. (1985), *Essential Papers on Depression*, New York University Press, New York, NY and London, England.

Cramer, A. O., Waldorp, L. J., van der Maas, H. L. & Borsboom, D. (2010), ‘Comorbidity: A network perspective’, *Behavioral and Brain Sciences* **33**, 137–150.

Craver, C. F. & Darden, L. (2013), *In Search of Mechanisms: Discoveries Across the Life Sciences*, University of Chicago Press, Chicago, IL and London, England.

Cristea, I. A., Huibers, M. J., David, D., Hollon, S. D., Andersson, G. & Cuijpers, P. (2015), ‘The effects of Cognitive Behavior Therapy for adult depression on dysfunctional thinking: A meta-analysis’, *Clinical Psychology Review* **42**, 62–71.

Cuijpers, P., Van Straten, A. & Warmerdam, L. (2007), ‘Behavioral activation treatments of depression: A meta-analysis’, *Clinical Psychology Review* **27**(3), 318–326.

Cummins, R. C. (1975), ‘Functional analysis’, *The Journal of Philosophy* **72**, 741–765.

Cummins, R. C. (1983), *The Nature of Psychological Explanation*, MIT Press, Cambridge, MA and London, England.

Cummins, R. C. (2000), “How does it work?” versus “What are the laws?”: Two conceptions of psychological explanation, *in* F. Keil & R. A. Wilson, eds, ‘Explanation and Cognition’, MIT Press, pp. 117–145.

David, D., Cristea, I. & Hofmann, S. G. (2018), ‘Why Cognitive Behavioral Therapy is the current gold standard of psychotherapy’, *Frontiers in Psychiatry* **9**(4).

de Jong, H. L. (2003), ‘Causal and functional explanations’, *Theory & Psychology* **13**(3), 291–317.

de Sousa, R. (2017), Emotion, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', winter 2017 edn, Metaphysics Research Lab, Stanford University.

Dennett, D. C. (1971), 'Intentional systems', *The Journal of Philosophy* **68**(4), 87–106.

Dennett, D. C. (1996), *Content and Consciousness*, Routledge, London, England and New York, NY.

Diffily, A. (1991), 'Father and child: Tim Beck and his uncommon common sense', *Penn Medicine* **4**, 20–27.

Disner, S. G., Beevers, C. G., Haigh, E. A. & Beck, A. T. (2011), 'Neural mechanisms of the cognitive model of depression', *Nature Reviews Neuroscience* **12**, 467–477.

Ellis, A. (1980), 'Rational-Emotive Therapy and Cognitive Behavior Therapy: Similarities and differences', *Cognitive Therapy and Research* **4**(4), 325–340.

Ellis, A. (1995), 'Changing Rational-Emotive Therapy (RET) to Rational Emotive Behavior Therapy (REBT)', *Journal of Rational-Emotive & Cognitive-Behavior Therapy* **13**(2), 85–89.

Ellis, A., David, D. & Lynn, S. J. (2010), Rational and irrational beliefs: A historical and conceptual perspective, *in* D. David, S. J. Lynn & A. Ellis, eds, 'Rational and Irrational Beliefs: Research, Theory, and Clinical Practice', Oxford University Press, chapter 1, pp. 3–22.

Emmett, S. D. & Francis, H. W. (2014), 'The socioeconomic impact of hearing loss in US adults', *Otology & Neurotology* **36**(3), 545–550.

Eysenck, M. W. (2013), *Anxiety: The Cognitive Perspective*, Psychology Press, London, England.

Fisher, P. & Wells, A. (2009), *Metacognitive Therapy: Distinctive Features*, The CBT Distinctive Features Series, Routledge, London, England and New York, NY.

Flynn, H. A. & Warren, R. (2014), 'Using CBT effectively for treating depression and anxiety', *Current Psychiatry* **13**(6), 45–53.

Foley, R. (1995), 'The adaptive legacy of human evolution: A search for the environment of evolutionary adaptedness', *Evolutionary Anthropology* **4**(6), 194–203.

Fonagy, P. (2003), 'Psychoanalysis today', *World Psychiatry* **2**(2), 73–80.

Frederick, S., Loewenstein, G. & O'Donoghue, T. (2002), 'Time discounting and time preference: A critical review', *Journal of Economic Literature* **40**(2), 351–401.

Freud, S. (1957), Mourning and melancholia, *in* 'The Standard Edition of the Complete Psychological Works of Sigmund Freud, Volume XIV (1914-1916): On the History of the Psycho-Analytic Movement, Papers on Metapsychology and Other Works', The Hogarth Press, pp. 237–258.

Freud, S. (1982), *Die Traumdeutung, Studienausgabe, Band II*, Fischer Wissenschaft, Frankfurt, Germany.

Fricker, M. (2007), *Epistemic Injustice: Power and the Ethics of Knowing*, Oxford University Press, New York, NY.

Frigg, R. & Hartmann, S. (2017), Models in science, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', spring 2017 edn, Metaphysics Research Lab, Stanford University.

Froggatt, W. (2005), 'A brief introduction to Rational Emotive Behaviour Therapy', *New Zealand Centre for Rational Emotive Behaviour Therapy* . **URL**: *https://www.rational.org.nz/prof-docs/Intro-REBT.pdf.* Last accessed on 07/22/2019.

Fus, T. (2006), 'Criminalizing marital rape: A comparison of judicial and legislative approaches', *Vanderbilt Journal of Transnational Law* **39**, 481–517.

Galletta, A. (2013), *Mastering the Semi-Structured Interview and Beyond: From Research Design to Analysis and Publication*, New York University Press, New York, NY and London, England.

Garberoglio, C. L., Cawthon, S. & Sales, A. (2017), 'Deaf people and educational attainment in the United States: 2017', *National Deaf Center on Postsecondary Outcomes* .

Gifford, S. (2008), The psychoanalytic movement in the United States, 1906–1991, *in* J. G. Edwin R Wallace, ed., 'History of Psychiatry and Medical Psychology', Springer, pp. 629–656.

Gigerenzer, G. & Selten, R., eds (2002), *Bounded Rationality: The Adaptive Toolbox*, MIT Press, Cambridge, MA and London, England.

Gill, M. M. (1954), 'Psychoanalysis and exploratory psychotherapy', *Journal of the American Psychoanalytic Association* **2**(4), 771–797.

Glennan, S. (2002), 'Rethinking mechanistic explanation', *Philosophy of Science* **69**(3), S342–S353.

Gold, L. H. (2014), 'DSM-5 and the assessment of functioning: The World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0)', *Journal of the American Academy of Psychiatry and the Law Online* **42**(2), 173–181.

Goldberg, D. (2011), 'The heterogeneity of "Major Depression"', *World Psychiatry* **10**(3), 226–228.

Groves, P. M. & Thompson, R. F. (1970), 'Habituation: A dual-process theory', *Psychological Review* **77**(5), 419–450.

Grünbaum, A. (1984), *The Foundations of Psychoanalysis: A Philosophical Critique*, University of California Press, Berkeley, LA and London, England.

Guenther, C. L. & Alicke, M. D. (2008), 'Self-enhancement and belief perseverance', *Journal of Experimental Social Psychology* **44**(3), 706–712.

Haaga, D. A. & Beck, A. T. (1995), 'Perspectives on depressive realism: Implications for cognitive theory of depression', *Behaviour Research and Therapy* **33**(1), 41–48.

Hacking, I. (1995), The looping effects of human kinds, *in* D. Sperber, D. Premack & A. J. Premack, eds, 'Symposia of the Fyssen Foundation. Causal Cognition: A Multidisciplinary Debate', Clarendon Press/Oxford University Press, New York, NY, pp. 351–394.

Hacking, I. (1998), *Rewriting the Soul: Multiple Personality and the Sciences of Memory*, Princeton University Press, Princeton, England. Reprint.

Hammen, C. (2006), ‘Stress generation in depression: Reflections on origins, research, and future directions’, *Journal of Clinical Psychology* **62**(9), 1065–1082.

Hampshire, S. & Hart, H. (1958), ‘Decision, intention and certainty’, *Mind* **67**(265), 1–12.

Hanwella, R., de Silva, V., Yoosuf, A., Karunaratne, S. & de Silva, P. (2012), ‘Religious beliefs, possession states, and spirits: Three case studies from Sri Lanka’, *Case Reports in Psychiatry* .

Hardcastle, V. G. (2002), On the normativity of functions, *in* M. P. André Ariew, Robert C Cummins, ed., ‘Functions: New Essays in the Philosophy of Psychology and Biology’, Oxford University Press.

Harkness, K. L. & Lumley, M. N. (2008), Child abuse and neglect and the development of depression in children and adolescents, *in* J. R. Z. Abela & B. L. Hankin, eds, ‘Handbook of Depression in Children and Adolescents’, Guilford Press, New York, NY, pp. 466–488.

Hart, B. L. (1988), ‘Biological basis of the behavior of sick animals’, *Neuroscience & Biobehavioral Reviews* **12**(2), 123–137.

Haselton, M. G., Nettle, D. & Murray, D. R. (2016), The evolution of cognitive bias, *in* D. M. Buss, ed., ‘The Handbook of Evolutionary Psychology’, 2 edn, John Wiley & Sons, Hoboken, NJ, pp. 968–987.

Heimberg, R. G., Brozovich, F. A. & Rapee, R. M. (2010), A cognitive behavioral model of social anxiety disorder: Update and extension, *in* ‘Social Anxiety’, Elsevier, pp. 395–422.

Hempel, C. G. & Oppenheim, P. (1948), ‘Studies in the logic of explanation’, *Philosophy of Science* **15**(2), 135–175.

Henrich, J., Boyd, R. & Richerson, P. J. (2012), ‘The puzzle of monogamous marriage’, *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**(1589), 657–669.

Hitchcock, C. (2018), Probabilistic causation, *in* E. N. Zalta, ed., ‘The Stanford Encyclopedia of Philosophy’, fall 2018 edn, Metaphysics Research Lab, Stanford University.

Hoff, P. (2015), 'The Kraepelinian tradition', *Dialogues in Clinical Neuroscience* **17**(1), 31–41.

Hofmann, S. G. (2007), 'Cognitive factors that maintain Social Anxiety Disorder: A comprehensive model and its treatment implications', *Cognitive Behaviour Therapy* **36**(4), 193–209.

Hofmann, S. G., Asnaani, A., Vonk, I. J. J., Sawyer, A. T. & Fang, A. (2012), 'The efficacy of Cognitive Behavioral Therapy: A review of meta-analyses', *Cognitive Therapy and Research* **36**(5), 427–440.

Imel, Z. E. & Wampold, B. E. (2008), The importance of treatment and the science of common factors in psychotherapy, *in* R. W. L. Steven D Brown, ed., 'Handbook of Counseling Psychology', Vol. 4, John Wiley & Sons, chapter 15, pp. 249–266.

Insel, T. R. (2010), 'Rethinking Schizophrenia', *Nature* **468**(7321), 187–193.

Johnson, R. W. (2002), 'The concept of sickness behavior: A brief chronological account of four key discoveries', *Veterinary Immunology and Immunopathology* **87**(3-4), 443–450.

Joshi, A. & Phadke, K. (2018), *Rational Emotive Behaviour Therapy Integrated*, SAGE Publishing India, New Delhi, India and Thousand Oaks, California.

Kästner, L. (2018), Identifying causes in psychiatry, *in* 'PSA 2018: The 26th Biennial Meeting of the Philosophy of Science Association (Seattle, WA; 1-4 November 2018)'. presented at PSA 2018 in Seattle.

Kelly, G. (1991), *The Psychology of Personal Constructs, Volume One: Theory of Personality*, Routledge, New York, NY.

Kendler, K. S. (2005), 'Toward a philosophical structure for psychiatry', *American Journal of Psychiatry* **162**(3), 433–440.

Kendler, K. S. (2008), Introduction: Why does psychiatry need philosophy?, *in* K. S. Kendler & J. Parnas, eds, 'Philosophical Issues in Psychiatry: Explanation, Phenomenology, and Nosology', Johns Hopkins University Press, pp. 1–16.

Kessler, R. C., Angermeyer, M., Anthony, J. C., De Graaf, R., Demyttenaere, K., Gasquet, I., De Girolamo, G., Gluzman, S., Gureje, O., Haro, J. M. et al. (2007), 'Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's World Mental Health Survey Initiative', *World Psychiatry* **6**(3), 168–176.

Kincaid, H. & Sullivan, J. A., eds (2014), *Classifying Psychopathology: Mental Kinds and Natural Kinds*, MIT Press, Cambridge, MA and London, England.

Kinderman, P. & Cooke, A. (2017), 'Mind your language! A guide to language about mental health and psychological wellbeing in the media and creative arts'. **URL**: *https://livrepository.liverpool.ac.uk/3007765/1/mind%20your%20language%20v6.pdf*. Last accessed on 07/19/2019.

Kizilhan, J. I. (2018), 'PTSD of rape after IS ("Islamic State") captivity', *Archives of Women's Mental Health* **21**(5), 517–524.

Kraines, S. H. (1957), *Mental Depressions and Their Treatment*, Macmillan.

Lacasse, J. R. & Leo, J. (2005), 'Serotonin and depression: A disconnect between the advertisements and the scientific literature', *PLoS Medicine* **2**(12), 1211–1216.

Lakatos, I. (1970), 'History of science and its rational reconstructions', *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* pp. 91–136.

Lambert, M. J. & Barley, D. E. (2001), 'Research summary on the therapeutic relationship and psychotherapy outcome', *Psychotherapy: Theory, Research, Practice, Training* **38**(4), 357–361.

Lambert, M. J. & Vermeersch, D. A. (2002), The effectiveness of psychotherapy, *in* M. Hersen & W. Sledge, eds, 'Encyclopedia of Psychotherapy', Vol. 1, Academic Press, San Diego, CA, pp. 709–714.

Lane, D. A. & Corrie, S. (2007), *The Modern Scientist-Practitioner: A Guide to Practice in Psychology*, Routledge.

LaVaque-Manty, M. (2006), 'Dueling for equality: Masculine honor and the modern politics of dignity', *Political Theory* **34**(6), 715–740.

Leslie, S.-J. & Lerner, A. (2016), Generic generalizations, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', winter 2016 edn, Metaphysics Research Lab, Stanford University.

Lewis, D. K. (1986), *Philosophical Papers: Volume II*, Oxford University Press.

Lichtenberg, P., Heresco-Levy, U. & Nitzan, U. (2004), 'The ethics of the placebo in clinical practice', *Journal of Medical Ethics* **30**, 551–554.

Lilienfeld, S. O. & Marino, L. (1995), 'Mental disorder as a Roschian concept: A critique of Wakefield's "harmful dysfunction" analysis', *Journal of Abnormal Psychology* **104**(3), 411–420.

Lilienfeld, S. O., Sauvigné, K. C., Lynn, S. J., Cautin, R. L., Latzman, R. D. & Waldman, I. D. (2015), 'Fifty psychological and psychiatric terms to avoid: A list of inaccurate, misleading, misused, ambiguous, and logically confused words and phrases', *Frontiers in Psychology* **6**, 1–15.

Lindert, J., von Ehrenstein, O. S., Grashow, R., Gal, G., Braehler, E. & Weisskopf, M. G. (2014), 'Sexual and physical abuse in childhood is associated with depression and anxiety over the life course: Systematic review and meta-analysis', *International Journal of Public Health* **59**(2), 359–372.

Losada, A., Montorio, I., Knight, B. G., Márquez, M. & Izal, M. (2006), 'Explanation of caregivers' distress from the cognitive model: The role of dysfunctional thoughts', *Psicología Conductual* **14**(1), 115–128.

Machamer, P., Darden, L. & Craver, C. F. (2000), 'Thinking about mechanisms', *Philosophy of Science* pp. 1–25.

Mahon, J. E. (2016), The definition of lying and deception, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', winter 2016 edn, Metaphysics Research Lab, Stanford University.

Maletic, V., Robinson, M., Oakes, T., Iyengar, S., Ball, S. & Russell, J. (2007), 'Neurobiology of depression: An integrated view of key findings', *International Journal of Clinical Practice* **61**, 2030–2040.

McCrae, R. R. & Costa Jr, P. T. (1999), A five-factor theory of personality, *in* O. P. J. Lawrence A Pervin, ed., 'Handbook of Personality: Theory and Research', 2 edn, Guilford Press, pp. 139–153.

McFall, R. M. (1991), 'Manifesto for a science of clinical psychology', *The Clinical Psychologist* **44**(6), 75–88.

McLaughlin, P. (2001), *What Functions Explain: Functional Explanation and Self-Reproducing Systems*, Cambridge University Press.

Menzel, C. (2017), Possible worlds, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', winter 2017 edn, Metaphysics Research Lab, Stanford University.

Miller, G. A. (2003), 'The cognitive revolution: A historical perspective', *Trends in Cognitive Sciences* **7**(3), 141–144.

Milton, J., Polmear, C. & Fabricius, J. (2011), *A Short Introduction to Psychoanalysis*, 2 edn, SAGE Publications.

Moore, G. E. (1988), *Principia Ethica*, Prometheus Books, Amherst, MA and New York, NY.

Moore, M. T. & Fresco, D. M. (2012), 'Depressive realism: A meta-analytic review', *Clinical Psychology Review* **32**(6), 496–509.

Moores, D. F. (2010), 'Partners in progress: The 21st International Congress on Education of the Deaf and the repudiation of the 1880 Congress of Milan', *American Annals of the Deaf* **155**(3), 309–310.

Morawska, A. & Sanders, M. (2007), 'Concurrent predictors of dysfunctional parenting and maternal confidence: Implications for parenting interventions', *Child: Care, Health and Development* **33**(6), 757–767.

Murphy, D. (2006), *Psychiatry in the Scientific Image*, MIT Press.

Murphy, D. (2010), 'Explanation in psychiatry', *Philosophy Compass* **5**, 602–610.

Nathan, P. E. & Gorman, J. M., eds (2015), *A Guide to Treatments That Work*, Oxford University Press.

Nestler, S. (2010), 'Belief perseverance: The role of accessible content and accessibility experiences', *Social Psychology* **41**, 35–41.

Nordenfelt, L. (2007), *Rationality and Compulsion: Applying Action Theory to Psychiatry*, International Perspectives in Philosophy and Psychiatry, Oxford University Press.

Norvell, N. & Forsyth, D. R. (1984), 'The impact of inhibiting or facilitating causal factors on group members' reactions after success and failure', *Social Psychology Quarterly* **47**(3), 293–297.

Obsessive Compulsive Cognitions Working Group (1997), 'Cognitive assessment of Obsessive-Compulsive Disorder', *Behaviour Research and Therapy* **35**(7), 667–681.

Okiishi, J., Lambert, M. J., Nielsen, S. L. & Ogles, B. M. (2003), 'Waiting for supershrink: An empirical analysis of therapist effects', *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice* **10**(6), 361–373.

Oxford Centre for Evidence-based Medicine (2009), 'Levels of evidence', **URL**: *https://www.cebm.net/2009/06/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/*. Last accessed on 09/11/2019.

Papa, A. & Follette, W. C. (2015), *Dismantling Studies of Psychotherapy*, American Cancer Society, pp. 1–6.

Piccinini, G. & Craver, C. (2011), 'Integrating psychology and neuroscience: Functional analyses as mechanism sketches', *Synthese* **183**(3), 283–311.

Plante, T. G. (2005), *Contemporary Clinical Psychology*, 2 edn, John Wiley & Sons, Hoboken, NJ.

Plous, S. (1993), *The psychology of judgment and decision making*, Mcgraw-Hill Book Company.

Plunkett, D. & Cappelen, H. (forthcoming), A guided tour of conceptual engineering and conceptual ethics, *in* H. Cappelen, D. Plunkett & A. Burgess, eds, 'Conceptual Engineering and Conceptual Ethics', Oxford University Press, Oxford, England. Draft of July 30, 2018.

Porta, M., ed. (2014), *A Dictionary of Epidemiology*, 6 edn, Oxford University Press, New York, NY.

Poston, T. (2019), ‘Internalism and externalism in epistemology’, *The Internet Encyclopedia of Philosophy* . **URL**: *https://www.iep.utm.edu/int-ext/.* Last accessed on 09/17/2019.

Priest, G. (1998), ‘What is so bad about contradictions?’, *The Journal of Philosophy* **95**(8), 410–426.

Rachman, d. S., De Silva, P. & Röper, G. (1976), ‘The spontaneous decay of compulsive urges’, *Behaviour Research and Therapy* **14**(6), 445–453.

Rachman, S. (1971), ‘Obsessional ruminations’, *Behaviour Research and Therapy* **9**(3), 229–235.

Rachman, S. (1976), ‘The modification of obsessions: A new formulation’, *Behaviour Research and Therapy* **14**(6), 437–443.

Rachman, S. & de Silva, P. (1978), ‘Abnormal and normal obsessions’, *Behaviour Research and Therapy* **16**(4), 233–248.

Rao, U., Chen, L.-A., Bidesi, A. S., Shad, M. U., Thomas, M. A. & Hammen, C. L. (2010), ‘Hippocampal changes associated with early-life adversity and vulnerability to depression’, *Biological Psychiatry* **67**(4), 357–364.

Ravenscroft, I. (2019), Folk psychology as a theory, *in* E. N. Zalta, ed., ‘The Stanford Encyclopedia of Philosophy’, summer 2019 edn, Metaphysics Research Lab, Stanford University.

Robertson, D. (2010), *The Philosophy of Cognitive-Behavioural Therapy (CBT): Stoic Philosophy as Rational and Cognitive Psychotherapy*, Karnac Books, London, England.

Roth, M. & Cummins, R. C. (2014), ‘Two tales of functional explanation’, *Philosophical Psychology* **27**(6), 773–788.

Rudy-Hiller, F. (2018), The epistemic condition for moral responsibility, *in* E. N. Zalta, ed., ‘The Stanford Encyclopedia of Philosophy’, fall 2018 edn, Metaphysics Research Lab, Stanford University.

Salkovskis, P. M. (1985), ‘Obsessional-compulsive problems: A cognitive-behavioural analysis’, *Behaviour Research and Therapy* **23**(5), 571–583.

Salkovskis, P. M. (1991), 'The importance of behaviour in the maintenance of anxiety and panic: A cognitive account', *Behavioural and Cognitive Psychotherapy* **19**(1), 6–19.

Salkovskis, P. M. (1996), Cognitive Therapy and Aaron T. Beck, *in* P. M. Salkovskis, ed., 'Frontiers of Cognitive Therapy', Guilford Press, New York, pp. 531–539.

Salkovskis, P. M. (1999), 'Understanding and treating Obsessive-Compulsive Disorder', *Behaviour Research and Therapy* **37**, S29–S52.

Salkovskis, P. M. (2019), 'Power Threat Meaning Framework presentation: What happened to me, why it seems threatening and what it means', **URL**: *https://psychonoclast.wordpress.com/2019/07/13/power-threat-meaning-framework-presentation-what-happened-to-me-why-it-seems-threatening-and-what-it-means/.* Last accessed on 09/05/2019.

Salkovskis, P. M. & Forrester, E. (2002), Responsibility, *in* 'Cognitive Approaches to Obsessions and Compulsions', Elsevier, chapter 4, pp. 45–61.

Salkovskis, P. M., Forrester, E. & Richards, C. (1998), 'Cognitive-behavioural approach to understanding obsessional thinking', *The British Journal of Psychiatry. Suppl.* **173**(S35), 53–63.

Salkovskis, P. M. & Warwick, H. M. (1985), 'Cognitive Therapy of Obsessive-Compulsive Disorder: Treating treatment failures', *Behavioural and Cognitive Psychotherapy* **13**(3), 243–255.

Samuels, R., Stich, S. & Bishop, M. (2002), Ending the rationality wars: How to make disputes about human rationality disappear, *in* 'Common Sense, Reasoning, and Rationality', Oxford University Press, pp. 236–268.

Scarantino, A. & de Sousa, R. (2018), Emotion, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', winter 2018 edn, Metaphysics Research Lab, Stanford University.

Schwitzgebel, E. (2019), Belief, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', fall 2019 edn, Metaphysics Research Lab, Stanford University.

Seligman, M. E. (1972), 'Learned helplessness', *Annual Review of Medicine* **23**(1), 407–412.

Shedler, J. (2010), 'The efficacy of psychodynamic psychotherapy', *American Psychologist* **65**(2), 98–109.

Söderfeldt, Y. (2013), *From Pathology to Public Sphere: The German Deaf Movement 1848–1914*, Vol. 9, Transcript, Bielefeld, Germany.

Stengler-Wenzke, K., Kroll, M., Matschinger, H. & Angermeyer, M. C. (2006), 'Subjective quality of life of patients with Obsessive-Compulsive Disorder', *Social Psychiatry and Psychiatric Epidemiology* **41**(8), 662–668.

Stinson, C. (2016), 'Mechanisms in psychology: Ripping nature at its seams', *Synthese* **193**(5), 1585–1614.

Tanner, J. (2006), 'The naturalistic fallacy', *The Richmond Journal of Philosophy* **13**, 1–6.

Teasdale, J. D. & Barnard, P. J. (1993), *Affect, Cognition and Change: Remodelling Depressive Thought*, Psychology Press, London, England.

Tee, J. & Kazantzis, N. (2011), 'Collaborative empiricism in Cognitive Therapy: A definition and theory for the relationship construct', *Clinical Psychology: Science and Practice* **18**, 47–61.

Tolin, D. F. (2010), 'Is Cognitive-Behavioral Therapy more effective than other therapies?: A meta-analytic review', *Clinical Psychology Review* **30**(6), 710–720.

Tracey, T. J. G. (2003), 'Concept mapping of therapeutic common factors', *Psychotherapy Research* **13**(4), 401–413. PMID: 21827252.

Tversky, A. & Kahneman, D. (1974), 'Judgment under uncertainty: Heuristics and biases', *Science* **185**(4157), 1124–1131.

Vita of Professor Paul Salkovskis (2017), **URL**: *http://www.hmc.ox.ac.uk/people/professor-paul-salkovskis/*. Last accessed on 04/07/2019.

Voigt, P. & von dem Bussche, A. (2017), 'The EU General Data Protection Regulation (GDPR): A practical guide'.

Vosgerau, G. & Soom, P. (2018), 'Reduction without elimination: Mental disorders as causally efficacious properties', *Minds and Machines* **28**(2), 311–330.

Wakefield, J. C. (1992), '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. (1999), 'Evolutionary versus prototype analyses of the concept of disorder', *Journal of Abnormal Psychology* **108**(3), 374–399.

Wakefield, J. C., Horwitz, A. V. & Schmitz, M. F. (2005), 'Are we overpathologizing the socially anxious? Social Phobia from a harmful dysfunction perspective', *Canadian Journal of Psychiatry* **50**(6), 317–319.

Watrin, J. P. & Darwich, R. (2012), 'On behaviorism in the cognitive revolution: Myth and reactions', *Review of General Psychology* **16**(3), 269–282.

Weishaar, M. E. (1993), *Aaron T. Beck*, SAGE Publications, London, England and Thousand Oaks, CA and New Delhi, India.

Weissman, A. N. & Beck, A. T. (1978), 'Development and validation of the Dysfunctional Attitude Scale: A preliminary investigation'. Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27—31, 1978).

Wells, A. (2011), *Metacognitive Therapy for Anxiety and Depression*, Guilford Press, New York, NY.

Williams, B. (2006), *Ethics and the Limits of Philosophy*, Routledge, London, England.

Wilson, G. & Shpall, S. (2016), Action, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', winter 2016 edn, Metaphysics Research Lab, Stanford University.

Wittchen, H.-U. & Hoyer, J. (2011), *Klinische Psychologie & Psychotherapie (Lehrbuch mit Online-Materialien)*, 2 edn, Springer.

Woodward, J. F. (2003), *Making Things Happen: A Theory of Causal Explanation*, Oxford University Press.

World Health Organization (2009), *Global Health Risks: Mortality and Burden of Disease Attributable to Selected Major Risks*, World Health Organization, Geneva, Switzerland. **URL**: *http://publichealthwell.ie/node/9612*. Last accessed on 09/12/2019.

Wright, L. (1973), 'Functions', *Philosophical Review* **82**(2), 139–168.

Appendices

Appendix A

Diagnostic Criteria from the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*

A.1 Diagnostic Criteria of a *Manic Episode*

A. A distinct period of abnormally and persistently elevated, expansive, or irritable mood and abnormally and persistently increased goal-directed activity or energy, lasting at least 1 week and present most of the day, nearly every day (or any duration if hospitalization is necessary).

B. During the period of mood disturbance and increased energy or activity, three (or more) of the following symptoms (four if the mood is only irritable) are present to significant degree and represent a noticeable change from usual behavior:

1. Inflated self-esteem or grandiosity.
2. Decreased need for sleep (e.g., feels rested after only 3 hours of sleep).
3. More talkative than usual or pressure to keep talking.
4. Flight of ideas or subjective experience that thoughts are racing.
5. Distractability (i.e., attention too easily drawn to unimportant or irrelevant external stimuli), as reported or observed.
6. Increase in goal-directed activity (either socially, at work or school, or sexually) or psychomotor agitation (i.e., purposeless non-goal-directed activity).

7. Excessive involvement in activities that have a high potential for painful consequences (e.g., engaging in unrestrained buying sprees, sexual indiscretions, or foolish business investments).

C. The mood disturbance is sufficiently severe to cause marked impairment in social or occupational functioning or to necessitate hospitalization to prevent harm to self or others, or there are psychotic features.

D. The episode is not attributable to the physiological effects of a substance (e.g., a drubg of abuse, a medication, other treatment) or to another medical condition.
Note: A full manic episode that emerges during antidepressant treatment (e.g., medication, electroconvulsive therapy) but persists at a fully syndromal level beyond the physiological effect of that treatment is sufficient evidence for a manic episode, and, therefore, a bipolar I diagnosis.

Note: Criteria A–D constitute a manic episode. At least one lifetime manic episode is required for the diagnosis of bipolar I disorder.
(American Psychiatric Association 2013, p. 124)

A.2 Diagnostic Criteria of *Dysruptive Mood Dysregulation Disorder*

A. Severe recurrent temper outbursts manifested verbally (e.g., verbal rages) and/or behaviorally (e.g., physical aggression toward people or property) that are grossly out of proportion in intensity or duration to the situation or provocation.

B. The temper outbursts are inconsistent with developmental level.

C. The temper outbursts occur, on average, three or more times per week.

D. The mood between temper outbursts is persistently irritable or angry most of the day, nearly every day, and is observable by others (e.g., parents, teachers, peers).

E. Criteria A–D have been present for 12 or more months. Throughout that time, the individual has not had a period lasting 3 or more consecutive months without all of the symptoms in Criteria A–D.

F. Criteria A and D are present in at least two of three settings (i.e., at home, at school, with peers) and are severe in at least one of these.

G. The diagnosis should not be made for the first time before age 6 years or after age 18 years.

H. By history or observation, the age at onset of Criteria A–E is before 10 years.

I. There has never been a distinct period lasting more than 1 day during which the full symptom criteria, except duration, for a manic or hypomanic episode have been met.

   Note: Developmentally appropriate mood elevation, such as occurs in the context of a highly positive events or its anticipation, should not be considered as a symptom of mania or hypomania.

J. The behaviors do not occur exclusively during an episode of major depressive disorder and are not better explained by another mental disorder (e.g., autism spectrum disorder, posttraumatic stress disorder, separation anxiety disorder, persistent depressive disorder [dysthymia]).

Note: The diagnosis cannot coexist with oppositional defiant disorder, intermittent explosive disorder, or bipolar disorder, though it can coexist with others, including major depressive disorder, attention-deficit/hyperactivity disorder, conduct disorder, and substance use disorders. Individuals whose symptoms meet criteria for both disruptive mood dysregulation disorder and oppositional defiant disorder should be given the diagnosis of disruptive mood dysregulation disorder. If an individual has ever experienced a manic or hypomanic episode, the diagnosis of disruptive mood dysregulation disorder should not be assigned.

K. The symptoms are not attributable to the physiological effects of a substance or to another medical or neurological condition.

(American Psychiatric Association 2013, p. 156)

A.3 Diagnostic Criteria of *Major Depressive Disorder*

A. Five (or more) of the following symptoms have been present during the same 2-week period and represent a change from previous functioning; at least one of the symptoms is either (1) depressed mood or (2) loss or interest or pleasure.

   Note: Do not include symptoms that are clearly attributable to another medical condition.

   1. Depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad, empty, hopeless) or observation made by others (e.g., appears tearful).
      (**Note:** In children and adolescents, can be irritable mood.)
   2. Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation).
   3. Significant weight loss when not dieting or weight gain (e.g., a change of more than 5% of body weight in a month), or decrease or increase in appetite nearly every day.
      (**Note:** In children, consider failure to make expected weight gain.)
   4. Insomnia or hypersomnia almost every day.
   5. Psychomotor agitation or retardation nearly every day (observable by others, not merely subjective feelings of restlessness or being slowed down).
   6. Fatigue or loss of energy nearly every day.
   7. Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick).
   8. Diminished ability to think or concentrate, or indecisiveness, nearly every day (either by subjective account or as observed by others).
   9. Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide.

B. The symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.

C. The episode is not attributable to the physiological effects of a substance or another medical condition.

Note: Criteria A–C represent a major depressive episode.
Note: Responses to a significant loss (e.g., bereavement, financial ruin, losses from a natural disaster, a serious medical illness or disability) may include the feelings of intense sadness, rumination about the loss, insomnia, poor appetite, and weight loss noted in Criterion A, which may resemble a depressive episode. Although such symptoms may be understandable or considered appropriate tho the loss, the presence of a major depressive episode in addition to the normal response to a significant loss should also be carefully considered. This decision inevitably requires the exercise of clinical judgement based on the individual's history and the cultural norms for the expression of distress in the context of loss.

D. The occurrence of the major depressive episode is not better explained by schizoaffective disorder, schizophrenia, schizophreniform disorder, delusional disorder, or other specified and unspecified schizophrenia spectrum and other psychotic disorders.

E. There has never been a manic episode or a hypomanic episode.
Note: This exclusion does not apply if all of the manic-like or hypomanic-like episodes are substance-induced or are attributable to the physiological effects of another medical condition.

(American Psychiatric Association 2013, p. 160-161)

A.4 Diagnostic Criteria of *Social Anxiety Disorder* (*Social Phobia*)

A. Marked fear or anxiety about one or more social situations in which the individual is exposed to possible scrutiny by others. Examples include social interactions (e.g., having a conversation, meeting unfamiliar people), being observed (e.g., eating or drinking), and performing in front of others (e.g., giving a speech).
   Note: In children, the anxiety must occur in peer settings and not just during interactions with adults.

B. The individual fears that he or she will act in a way or show anxiety symptoms that will be negatively evaluated (i.e., will be humiliating or embarrassing; will lead to rejection or offend others).

C. The social situations almost always provoke fear or anxiety.
   Note: In children, the fear or anxiety may be expressed by crying, tantrums, freezing, clinging, shrinking, or failing to speak in social situations.

D. The social situations are avoided or endured with intense fear or anxiety.

E. The fear or anxiety is out of proportion to the actual threat posed by the social situation and the sociocultural context.

F. The fear, anxiety, or avoidance is persistent, typically lasting for 6 months or more.

G. The fear, anxiety, or avoidance causes clinically significant distress or impairment in social, occupational, or other important areas of functioning.

H. The fear, anxiety, or avoidance is not attributable to the physiological effects of a substance (e.g., a drug of abuse, a medication) or another medical condition.

I. The fear, anxiety, or avoidance is not better explained by the symptoms of another mental disorder, such as panic disorder, body dysmorphic disorder, or autism spectrum disorder.

J. If another medical condition (e.g., Parkinson's disease, obesity, disfigurement from burns or injury) is present, the fear, anxiety, or avoidance is clearly unrelated or is excessive.

Specify if:

Performance only: If the fear is restricted to speaking or performing in public.

(American Psychiatric Association 2013, p. 202-203)

A.5 Diagnostic Criteria of *Obsessive-Compulsive Disorder*

A. Presence of obsessions, compulsions, or both:
Obsessions are defined by (1) and (2):

1. Recurrent and persistent thoughts, urges, or images that are experienced, at some time during the disturbance, as intrusive and unwanted, and that in most individuals cause marked anxiety and distress.
2. The individual attempts to ignore or suppress such thoughts, urges, or images, or to neutralize them with some other thought or action.

Compulsions are defined by (1) and (2):

1. Repetitive behaviors (e.g., hand washing, ordering, checking) or mental acts (e.g., praying, counting, repeating words silently) that the individual feels driven to perform in response to an obsession or according to rules that must be applied rigidly.
2. The behaviors or mental acts are aimed at preventing or reducing anxiety or distress, or preventing some dreaded event or situation; however, these behaviors or mental acts are not connected in a realistic way with what they are designed to neutralize or prevent, or are clearly excessive.
Note: Young children may not be able to articulate the aims of these behaviors or mental acts.

B. The obsessions or compulsions are time-consuming (e.g., take more than 1 hour per day) or cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.

C. The obsessive-compulsive symptoms are not attributable to the physiological effects of a substance (e.g., a drug of abuse, a medication) or another medical condition.

D. The disturbance is not better explained by the symptoms of another mental disorder (e.g., excessive worries, as in generalized anxiety disorder; preoccupation with appearance, as in body dysmorphic disorder; difficulty discarding or parting with possessions, as in hoarding disorder;

> hair pulling, as in trichotillomania [hair-pulling disorder]; skin picking, as in excoriation [skin-picking] disorder; stereotypies, as in stereotypic movement disorder; ritualized eating behavior, as in eating disorders; preoccupation with substances or gambling, as in substance-related and addictive disorders; preoccupation with having an illness, as in illness anxiety disorder; sexual urges or fantasies, as in paraphilic disorders; impulses, as in disruptive, impulse-control, and conduct disorders; guilty ruminations, as in major depressive disorder; thought insertion or delusional preoccupations, as in schizophrenia spectrum and other psychotic disorders; or repetitive patterns of behavior, as in autism spectrum disorder).

(American Psychiatric Association 2013, p. 237)

A.6 Diagnostic Criteria of *Posttraumatic Stress Disorder* for adults, adolescents and children older than six years

A. Exposure to actual or threatened death, serious injury, or sexual violence in one (or more) of the following ways:

1. Directly experiencing the traumatic event(s).
2. Witnessing, in person, the event(s) as it occurred to others.
3. Learning that the traumatic event(s) occurred to a close family member or close friend. In cases of actual or threatened death of a family member or friend, the event(s) must have been violent or accidental.
4. Experiencing repeated or extreme exposure to aversive details of the traumatic event(s) (e.g., first responders collecting human remains; police officers repeatedly exposed to details of child abuse).
   Note: Criterion A4 does not apply to exposure through electronic media, television, movies, or pictures, unless this exposure is work related.

B. Presence of one (or more) of the following intrusion symptoms associated with the traumatic event(s), beginning after the traumatic event(s) occurred:

1. Recurrent, involuntary, and intrusive distressing memories of the traumatic event(s). Note: In children older than 6 years, repetitive play may occur in which themes or aspects of the traumatic event(s) are expressed.
2. Recurrent distressing dreams in which the content and/or affect of the dream are related to the traumatic event(s).
   Note: In children, there may be frightening dreams without recognizable content.
3. Dissociative reactions (e.g., flashbacks) in which the individual feels or acts as if the traumatic event(s) were recurring. (Such reactions may occur on a continuum, with the most extreme expression being a complete loss of awareness of present surroundings.)
   Note: In children, trauma-specific reenactment may occur in play.

4. Intense or prolonged psychological distress at exposure to internal or external cues that symbolize or resemble an aspect of the traumatic event(s).
5. Marked physiological reactions to internal or external cues that symbolize or resemble an aspect of the traumatic event(s).

C. Persistent avoidance of stimuli associated with the traumatic event(s), beginning after the traumatic event(s) occurred, as evidenced by one or both of the following:

   1. Avoidance of or efforts to avoid distressing memories, thoughts, or feelings about or closely associated with the traumatic event(s).
   2. Avoidance of or efforts to avoid external reminders (people, places, conversations, activities, objects, situations) that arouse distressing memories, thoughts, or feelings about or closely associated with the traumatic event(s).

D. Negative alterations in cognitions and mood associated with the traumatic event(s), beginning or worsening after the traumatic event(s) occurred, as evidenced by two (or more) of the following:

   1. Inability to remember an important aspect of the traumatic event(s) (typically due to dissociative amnesia, and not to other factors such as head injury, alcohol, or drugs).
   2. Persistent and exaggerated negative beliefs or expectations about oneself, others, or the world (e.g., “I am bad,” “No one can be trusted,” “The world is completely dangerous,” “My whole nervous system is permanently ruined”).
   3. Persistent, distorted cognitions about the cause or consequences of the traumatic event(s) that lead the individual to blame himself/herself or others.
   4. Persistent negative emotional state (e.g., fear, horror, anger, guilt, or shame).
   5. Markedly diminished interest or participation in significant activities.
   6. Feelings of detachment or estrangement from others.

7. Persistent inability to experience positive emotions (e.g., inability to experience happiness, satisfaction, or loving feelings).

E. Marked alterations in arousal and reactivity associated with the traumatic event(s), beginning or worsening after the traumatic event(s) occurred, as evidenced by two (or more) of the following:

   1. Irritable behavior and angry outbursts (with little or no provocation), typically expressed as verbal or physical aggression toward people or objects.
   2. Reckless or self-destructive behavior.
   3. Hypervigilance.
   4. Exaggerated startle response.
   5. Problems with concentration.
   6. Sleep disturbance (e.g., difficulty falling or staying asleep or restless sleep).

F. Duration of the disturbance (Criteria B, C, D and E) is more than 1 month.

G. The disturbance causes clinically significant distress or impairment in social, occupational, or other important areas of functioning.

H. The disturbance is not attributable to the physiological effects of a substance (e.g., medication, alcohol) or another medical condition.

Specify whether:

With dissociative symptoms: The individual's symptoms meet the criteria for posttraumatic stress disorder, and in addition, in response to the stressor, the individual experiences persistent or recurrent symptoms of either of the following:

1. **Depersonalization:** Persistent or recurrent experiences of feeling detached from, and as if one were an outside observer of, one's mental processes or body (e.g., feeling as though one were in a dream; feeling a sense of unreality of self or body or of time moving slowly).

2. **Derealization:** Persistent or recurrent experiences of unreality of surroundings (e.g., the world around the individual is experienced as unreal, dreamlike, distant, or distorted).

Note: To use this subtype, the dissociative symptoms must not be attributable to the physiological effects of a substance (e.g., blackouts, behavior during alcohol intoxication) or another medical condition (e.g., complex partial seizures).

Specify if:

With delayed expression: If the full diagnostic criteria are not met until at least 6 months after the event (although the onset and expression of some symptoms may be immediate).

(American Psychiatric Association 2013, p. 669-670)

A.7 Diagnostic Criteria of *Narcissistic Personality Disorder*

A pervasive pattern of grandiosity (in fantasy or behavior), need for admiration, and lack of empathy, beginning by early adulthood and present in a variety of contexts, as indicated by five (or more) of the following:

1. Has a grandiose sense of self-importance (e.g., exaggerates achievements and talents, expects to be recognized as superior without commensurate achievements).

2. Is preoccupied with fantasies of unlimited success, power, brilliance, beauty, or ideal love.

3. Believes that he or she is "special" and unique and can only be understood by, or should associate with, other special or high-status people (or institutions).

4. Requires excessive admiration.

5. Has a sense of entitlement (i.e., unreasonable expectations of especially favorable treatment or automatic compliance with his or her expectations).

6. Is interpersonally exploitative (i.e., takes advantage of others to achieve his or her own ends).

7. Lacks empathy: is unwilling to recognize or identify with the feelings and needs of others.

8. Is often envious of others or believes that others are envious of him or her.

9. Shows arrogant, haughty behaviors or attitudes.

(American Psychiatric Association 2013, p. 669-670)

Appendix B

Interview Material

B.1 Information for Participants

Information für Teilnehmer/innen

Interviewleitung: Julia Pfeiff, M.Sc., B.A.
julia.pfeiff@philos.uni-hannover.de
Institut für Philosophie
Leibniz Universität Hannover

Sehr geehrte Damen und Herren,

Was ist Ziel der Studie?
Im Rahmen des Graduiertenkollegs „Die Integration von theoretischer und praktischer Wissenschaftsphilosophie“ wird an der Leibniz-Universität Hannover ein Projekt zur wissenschaftsphilosophischen Erforschung klinisch-psychologischer Erklärungsmodelle durchgeführt. Der Fokus dieses Projekts sind die kognitive Prozesse und die sozialen Praktiken, welche die Konstruktion sowie die psychotherapeutische Anwendung kognitiver, klinisch-psychologischer Modelle beeinflussen. Im Rahmen dieses Projekts soll anhand von Interviews mit Experten und Literaturstudien geklärt werden, wie die Struktur klinisch-psychologischer Erklärungen durch diese verschiedenen praktischen Ansprüche beeinflusst wird.

Wie sieht der Ablauf der Studie aus?
Die Studie besteht in einem etwa zweistündigen Interview. Zu Beginn dieser Sitzung werden Sie ausführlich über Ziele, Zweck und Ablauf der Studie aufgeklärt, und es werden erste Angaben zu Ihrer Person erhoben. Daraufhin beginnt das eigentliche Interview.

Ergeben sich aus der Teilnahme an der Studie für Sie zusätzliche Risiken?
Es ergeben sich keine Risiken, wenn Sie an dieser Studie teilnehmen. Im Falle dessen, dass Sie einzelne Fragen nicht beantworten können oder wollen, entstehen Ihnen keine Nachteile und das Fortführen der Studie ist weiterhin möglich.

Welche Maßnahmen werden zur Vermeidung von Risiken und Unannehmlichkeiten getroffen und kann ich von der Studie zurücktreten?
Ihre Teilnahme an dieser Studie ist freiwillig. Sie können während des kompletten Zeitraums ohne Angaben von Gründen Ihre Teilnahme beenden oder das Beantworten einer Frage verweigern. Daraus entstehen Ihnen keinerlei Nachteile. Falls Sie den Wunsch haben Ihre Teilnahme zu beenden, wenden Sie sich bitte an die aufgeführte Studienleitung. Alle Ihre persönlichen Daten werden dann gelöscht.

Welchen Nutzen hat die Studie für Sie?
Sie tragen einen bedeutenden Teil zur Erforschung von klinisch-psychologischen Erklärungen und Erklärungsmodellen sowie deren Anwendung bei.

Hinweise zum Datenschutz
In dieser Studie werden persönliche Daten von Ihnen erfasst. Alle erhobenen Daten werden unter strenger Beachtung der gesetzlichen Regelungen zum Datenschutz aufbewahrt. Die Projektleiterin ist verantwortlich für die Einhaltung der nationalen und internationalen Richtlinien zum Datenschutz in dieser Studie. Sie können jederzeit Auskunft über Ihre gespeicherten Daten verlangen. Sie haben das Recht, fehlerhafte Daten zu berichtigen oder Daten löschen zu lassen, und Sie haben das Recht zu jeder Zeit die Einwilligung zur Verarbeitung Ihrer personenbezogenen Daten zu widerrufen. Bitte kontaktieren Sie hierfür die verantwortliche Studienleiterin, Frau Julia Pfeiff (Mail: julia.pfeiff@philos.uni-hannover.de, Tel.: 0173 - 4958976).

Es werden nur personenbezogene Daten erhoben, die für das Erreichen des Studienziels erforderlich sind (Vor- und Familienname, auditive Aufnahme des Interviews). Ihre wissenschaftlichen Daten werden zunächst in pseudonymisierter Form elektronisch abgespeichert. Sie sind nur an der Studie beteiligten Fachleuten in kodierter Form zur wissenschaftlichen Auswertung zugänglich. Pseudonymisierung bedeutet, dass ein Dokument erstellt wird, das Ihren Namen mit den anderen Studiendaten verbindet. Dieses Dokument wird an einem separaten Ort aufbewahrt und ausschließlich dem verantwortlichen Studienleiter zugänglich gemacht. Sobald die Datenauswertung im September 2019 abgeschlossen ist, wird dieses Dokument vernichtet. Ab diesem Zeitpunkt ist eine Auskunft, Berichtigung oder Löschung Ihrer Daten nicht mehr möglich. Alle anderen Daten, welche nicht mit Ihrer Person in Zusammenhang gebracht werden können, werden aufbewahrt. Ihr Name wird in keiner Weise in Berichten oder Publikationen, die aus der Studie hervorgehen, veröffentlicht.

Für Fragen im Zusammenhang mit dieser Studie können Sie sich gerne an die Studienleitende wenden:

Julia Pfeiff, M. Sc., B. A.
DFG-Graduiertenkolleg 2073
Institut für Philosophie
Leibniz Universität Hannover
Am Klagesmarkt 14-17
30159 Hannover
Raum 511
Tel. +49 (0) 511 / 762 - 14505
julia.pfeiff@philos.uni-hannover.de

Information for participants

Interviewleitung: Julia Pfeiff, M.Sc., B.A.
julia.pfeiff@philos.uni-hannover.de
Institut für Philosophie
Leibniz Universität Hannover

Dear Sir or Madam,

What is the goal of the study?
I conduct a research project in the philosophy of science investigating clinical-psychological explanatory models of mental disorders. This project is part of the research conducted within the research training group "Integrating Ethics and Epistemology of Scientific Research", funded by the *Deutsche Forschungsgemeinschaft* (DFG). The focus of this project are the cognitive processes and social practices which influence the construction and the psychotherapeutic application of cognitive clinical-psychological models. In this project, we want to clarify how the structure of clinical-psychological explanation is influenced by these different practical demands by conducting literature research and interviewing experts.

What will happen during the study?
The study consists in a qualitative interview which will take approximately two hours. At the beginning of this session, you will be informed about the goals, the purpose and the procedure of the study. The interview will start after that.

Are there any additional risks due to participating in this study?
The information available to me strongly supports the view that there are no risks from participating in this study. Refusal or inability to answer individual questions will not result in disadvantages for you and completing of the study will still be possible.

Which procedures are in place in order to minimize risk and inconvenience and how can I withdraw from the study?
Your participation in this study is voluntary. During the entire course of the interview, you can end your participation in your study or refuse to answer a question without giving reasons. This will not result in any disadvantages. If you wish to end your participation in the study, please turn to the interviewer. All of your personal data will be destroyed by September 2019.

Which benefits will I have from participating?
You will contribute to research on clinical-psychological explanations and explanatory models as well as their application in scientific research and medical practice.

Data security
In this study, personal data will be gathered. All of this data will be kept with due regard to the legal regulations on data security. The interviewer is responsible to adhere to the national and international guidelines for data security. You have the right to always obtain information about your stored data.
You have the right to correct flawed data or to let data be deleted. You also have the right to revoke your agreement for the analysis of your personal data. To do so, please contact the responsible interviewer, Julia Pfeiff (eMail: julia.pfeiff@philos.uni-hannover.de, Tel.: 0173 - 4958976).
We will only gather personal data which is necessary for accomplishing the goals of the study (first name, surname, auditive recording of the interview). Your scientific data will first be stored electronically in pseudonymized form. It

is only accessible to the professionals who are involved in the study in pseudonymized form. Pseudonymization means that a document will be created which connects your name with other data in the study.

.

This document will be stored at a separate place from the other data. It will only be accessible to the responsible head of the study. When the analysis of the data is finished in September 2019, this document will be destroyed. From this point on, it will not be possible for you to obtain information about your data, to correct or to erase it. All of the data which cannot be associated with your person will be stored. Your name will not be mentioned in reports or publications which will be a result of the study.

For any questions concerning this study, please consult:

Julia Pfeiff, M. Sc., B. A.
DFG graduate training group 2073
Institute for Philosophy
Leibniz University Hanover
Am Klagesmarkt 14-17
30159 Hannover
Room 511
Tel. +49 (0) 511 / 762 - 14505
julia.pfeiff@philos.uni-hannover.de

B.2 Consent Form

Einwilligungserklärung

Forschungsprojekt „Erklärungsmodelle psychischer Störungen“

Interviewleitung: Julia Pfeiff, M.Sc., B.A.
julia.pfeiff@philos.uni-hannover.de
Institut für Philosophie
Leibniz Universität Hannover
Am Klagesmarkt 14-17
30159 Hannover
Raum 511
Tel. +49 (0) 511 / 762 - 14505

Hiermit erkläre ich mich bereit, im Rahmen des oben genannten Forschungsprojekts freiwillig an einem etwa zweistündigen Interview teilzunehmen. Ich bin in einem persönlichen Gespräch ausführlich und verständlich über Ziele, Bedeutung und Zweck des Forschungsprojekts aufgeklärt worden. Ich hatte die Gelegenheit zu einem Beratungsgespräch. Alle meine Fragen wurden zufriedenstellend beantwortet. Ich kann jederzeit neue Fragen stellen.

Ich hatte ausreichend Zeit, mich für oder gegen die Teilnahme an diesem Interview zu entscheiden. Mir ist bekannt, dass ich jederzeit und ohne Angabe von Gründen meine Einwilligung zur Teilnahme zurückziehen kann (mündlich oder schriftlich) sowie die Beantwortung einzelner Fragen verweigern kann, ohne dass mir daraus Nachteile entstehen.

Ich habe verstanden und bin damit einverstanden, dass meine studienbezogenen Daten zunächst pseudonymisiert (d.h. kodiert ohne Angabe von Namen, Anschrift, Initialen oder Ähnliches) erhoben, auf Datenträgern gespeichert und ausgewertet werden. Insbesondere bin ich damit einverstanden, dass das durchgeführte Interview auditiv aufgezeichnet, transkribiert, gespeichert und ausgewertet wird.
Die Weitergabe an Dritte einschließlich Publikation erfolgt ausschließlich in anonymer Form, d.h. kann nicht meiner Person zugeordnet werden. Für den Fall, dass ich die Studienteilnahme widerrufe, werden meine bereits erhobenen personenbezogenen Daten umgehend gelöscht.

Ein Exemplar der Einwilligungserklärung habe ich erhalten, gelesen und verstanden.

Ort, Datum, Unterschrift Teilnehmer/in

Ich habe das Aufklärungsgespräch geführt und die Einwilligung des Teilnehmers eingeholt.

Ort, Datum, Name der Versuchsleitung in Druckbuchstaben und Unterschrift

Consent form

Research project „Explanatory models of mental disorders“

Interviewer: Julia Pfeiff, M.Sc., B.A.
julia.pfeiff@philos.uni-hannover.de
Institute for Philosophy
Leibniz University Hanover
Am Klagesmarkt 14-17
30159 Hannover
Room 511
Tel. +49 (0) 511 / 762 - 14505

I hereby agree to participate in a two-hour interview which is a part of the research project mentioned above. I have been informed about the goals, the meaning and the purpose of this research project. I have had the opportunity for consultation. All of my questions have been answered satisfactorily. I can ask new questions at any time.

I have had enough time to decide for or against participating in this interview. I know that I may withdraw my participation at any time during the interview without consequences of any kind or loss of benefits. I also know that I may refuse to answer any questions I do not want to answer. There is no penalty if I decide to withdraw from the study.

I understand and agree to my interview data being stored in pseudonymized form (it will be coded without mentioning my name, address, initials or the like), that it will be stored and analyzed. In particular, I agree that the interview will be recorded, transcribed, stored on a password-protected computer, and analyzed. The data will only be given to third parties (including publication) in anonymized form, meaning that it cannot be traced back to me. In case I withdraw my consent to participate in this study, the data which has already been gathered will be destroyed immediately.

I have been given a copy of this document. I have read and understood it.

Place, Date, Signature of the Participant

I have informed the participant about the study and obtained his or her consent.

Place, Date, Signature of the Interviewer

Appendix C

List of Figures

C.1 Beck and Bredemeier's 2016 Model of Depression: Predisposition to the Syndrome

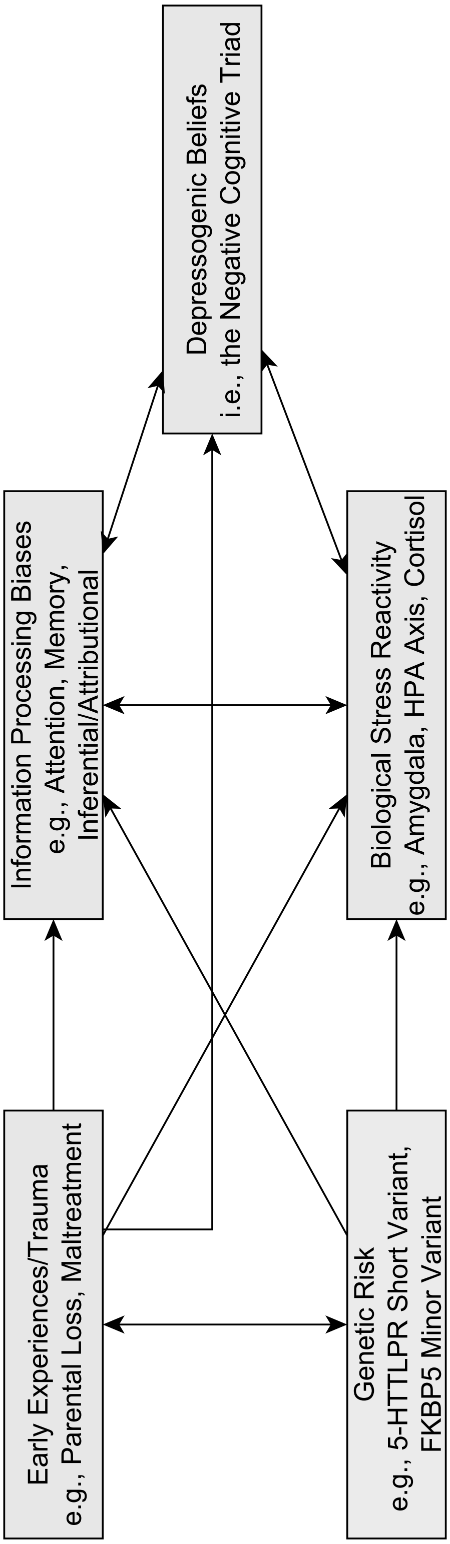
Depressogenic Beliefs
i.e., the Negative Cognitive Triad
Information Processing Biases
e.g., Attention, Memory,
Inferential/Attributional
Biological Stress Reactivity
e.g., Amygdala, HPA Axis, Cortisol
Early Experiences/Trauma
e.g., Parental Loss, Maltreatment
Genetic Risk
e.g., 5-HTTLPR Short Variant,
FKBP5 Minor Variant

Figure C.1: Factors underlying the predisposition for MDD, put forward by Beck & Bredemeier (2016), slightly adapted.

C.2 Beck and Bredemeier's 2016 Model of Depression: Maintenance of the Syndrome

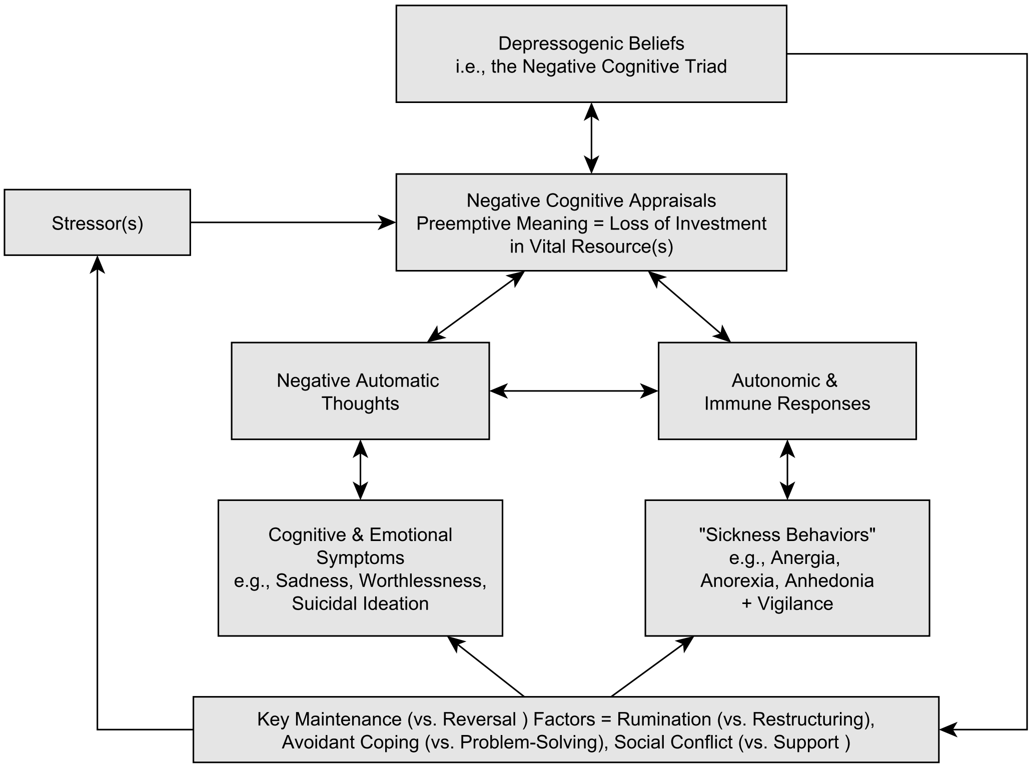
Depressogenic Beliefs
i.e., the Negative Cognitive Triad
Stressor(s)
Negative Cognitive Appraisals
Preemptive Meaning = Loss of Investment
in Vital Resource(s)
Negative Automatic
Thoughts
Autonomic &
Immune Responses
Cognitive & Emotional
Symptoms
e.g., Sadness, Worthlessness,
Suicidal Ideation
"Sickness Behaviors"
e.g., Anergia,
Anorexia, Anhedonia
+ Vigilance
Key Maintenance (vs. Reversal) Factors = Rumination (vs. Restructuring),
Avoidant Coping (vs. Problem-Solving), Social Conflict (vs. Support)

Figure C.2: MDD as due to the execution of an evolved program and maintenance factors stabilizing these symptoms, put forward by Beck & Bredemeier (2016), slightly adapted.

C.3 Salkovskis' 1985 Model of OCD

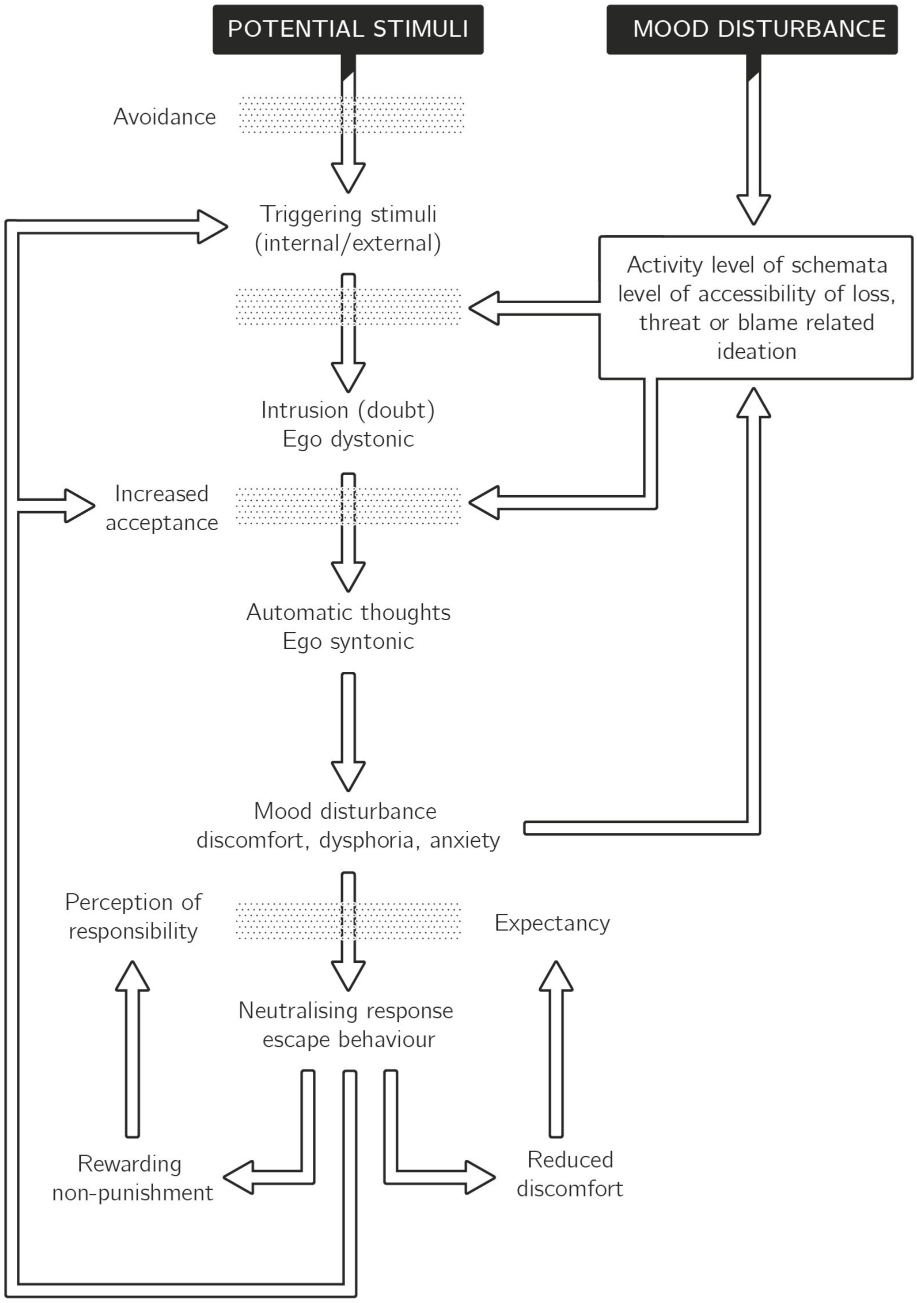

Figure C.3: Cognitive model of the origins and maintenance of Obsessive-Compulsive Disorder, put forward by Salkovskis (1985). Slightly adapted.

C.4 Salkovskis et al.'s 1998 Model of OCD

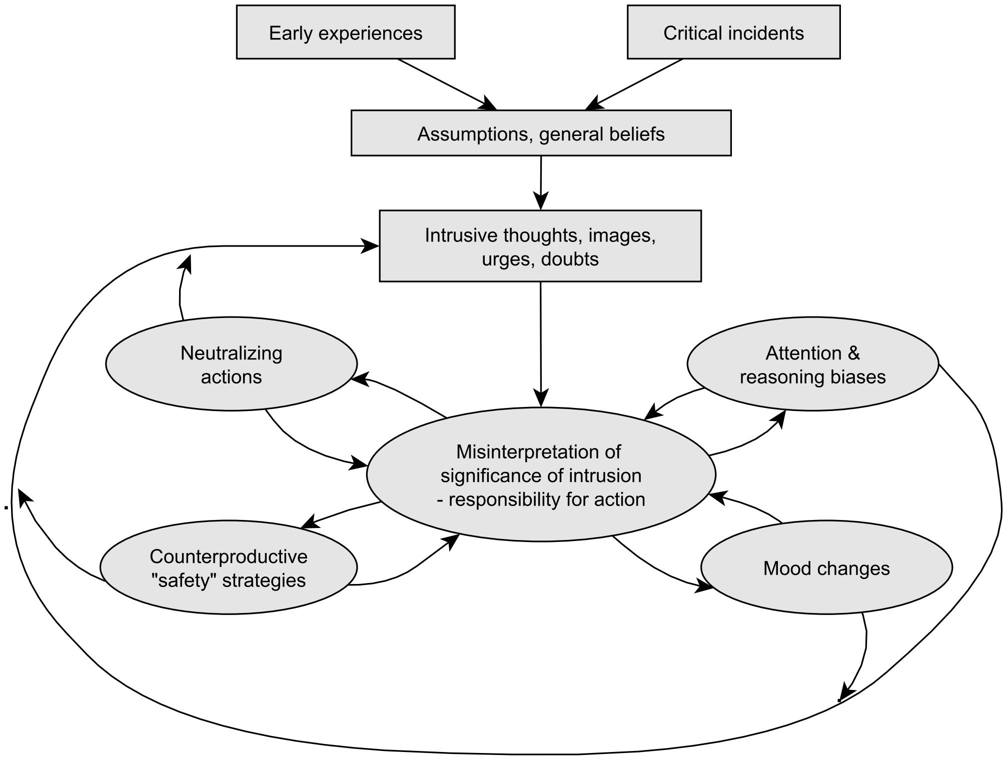
Early experiences
Critical incidents
Assumptions, general beliefs
Intrusive thoughts, images, urges, doubts
Neutralizing actions
Attention & reasoning biases
Misinterpretation of significance of intrusion - responsibility for action
Counterproductive "safety" strategies
Mood changes

Figure C.4: Cognitive model of the origins and maintenance of Obsessive-Compulsive Disorder, put forward by Salkovskis et al. (1998), slightly adapted.

C.5 Baddeley's Model of Working Memory

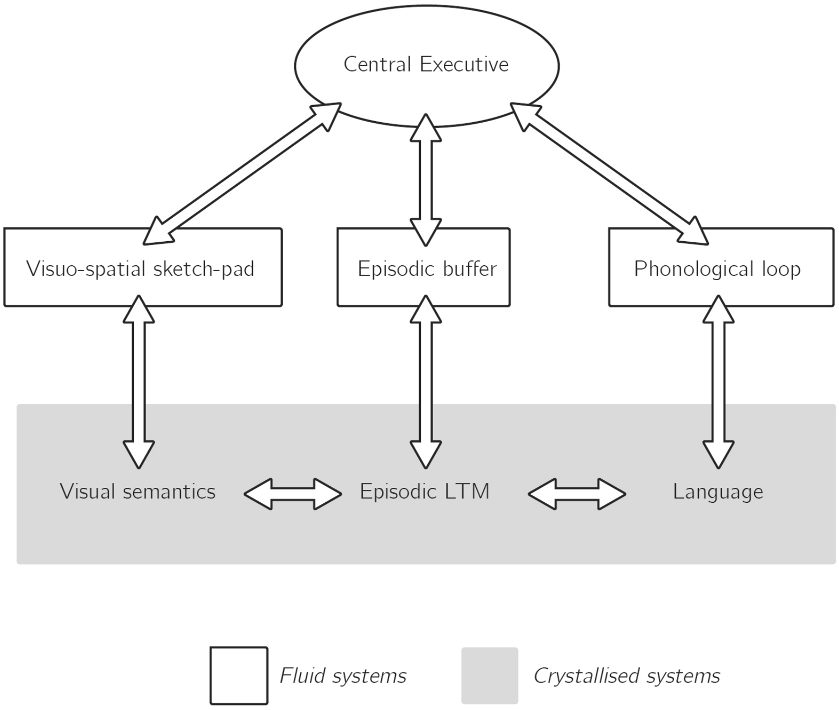

Figure C.5: Model of the working memory as presented by Baddeley (2012), slightly adapted.