Autoengano e delírio
Dois ensaios sobre crença e racionalidade

José Eduardo Freitas Porcher

Orientado por
Paulo Francisco Estrella Faria

Dissertação apresentada como requisito parcial
à obtenção do grau de Mestre em Filosofia

Universidade Federal do Rio Grande do Sul
Instituto de Filosofia e Ciências Humanas
Programa de Pós-Graduação em Filosofia

Porto Alegre, Brasil

Junho de 2011

Conteúdo

Agradecimentos **3**

Introdução **5**

1 Belief and imagination in the explanation of self-deception **8**
1.1 Introducing self-deception 8
1.2 What's wrong with self-deceptive belief 14
1.3 Self-deception as pretense 15
1.4 Assessing pretense 17
1.4.1 Velleman on belief and Gendler's appropriation 18
1.4.2 The motivational roles of belief and imagination 22
1.4.3 Distinguishing belief and imagination 26
1.4.4 Context and practical ground 27
1.4.5 The practical ground of self-deception 30
1.5 Conclusion 33

2 The tenability of a dispositional account of delusional belief **35**
2.1 Introducing delusion 35
2.2 What's wrong with delusional belief 39
2.3 A dispositional approach to delusions 44
2.3.1 In-between believing 44
2.3.2 A phenomenal, dispositional account of belief 46
2.3.3 Bayne and Pacherie's appropriation 50
2.4 Can dispositionalism save doxasticism? 53
2.4.1 Deviations, excuses, and explanations 53
2.4.2 Context-dependency 55
2.4.3 Revisiting the objections to the doxastic account 57
2.5 Conclusion 62

Conclusão **66**

Bibliografia **69**

Ao meu amigo Thiago Sebben

Agradecimentos

Este estudo não teria sequer sido iniciado sem o apoio do professor Paulo Faria. Ao levar até ele a proposta de estudar um tema admitidamente inusitado, sua reação foi a melhor que eu poderia esperar. Ele provou ser um orientador perfeito, não só deixando que eu desenvolvesse minha pesquisa sem qualquer tipo de obstáculo, mas me apoiando energicamente por todo o percurso que resultou neste trabalho. Meu débito de gratidão se deve também à sua leitura minuciosa dos textos que compõem esta dissertação, e aos seus comentários valiosos.

Durante o percurso que me trouxe até aqui tive o prazer de ter tido ótimos professores, aos quais se deve em grande parte o fato de eu ter decidido prosseguir na vida acadêmica. Possuo um débito de gratidão particular ao professor Fernando Fleck, cujo exemplo de rigor, paciência e humildade que testemunhei em seus seminários ao longo de vários semestres foi muito importante para o meu desenvolvimento intelectual e pessoal. Agradeço também aos professores Lia Levy, Sílvia Altmann, Tiago Falkenbach e Renato Fonseca, cada um dos quais teve um impacto positivo e duradouro sobre a minha formação.

Aos meus queridos amigos Pedro Pricladnitzky, Thiago Dihl Perin, Thiago Sebben e José Ademar Arnold, agradeço pela amizade, pelo carinho e pelo apoio sem o qual não teria sido tão fácil percorrer os últimos sete anos, que se passaram com alegria apesar de todas as crises pelas quais qualquer estudante sério de filosofia há de passar. Ao meu amigo Thiago Sebben, a quem esta dissertação é humildemente dedicada, agradeço em especial por depositar em mim um voto de confiança que estou muito aquém de merecer e que motiva minha vida intelectual a todo o momento.

À minha namorada, amiga e companheira Magda Togni, a quem nenhuma palavra de agradecimento seria suficiente e a quem eu devo muito mais do que seria capaz de descrever, agradeço pelas inúmeras vezes em que me emprestou os seus ouvidos, pelo apoio e confiança incondicionais, pela paciência inigualável, pelo conforto em momentos difíceis, por se alegrar genuinamente diante de cada etapa concluída, e por ser tudo e mais do que eu mereço.

Pelo amor e amparo incondicionais durante toda a minha vida, agradeço aos meus queridos pais Maria Teresa e Paulo Luiz, e aos meus avós Lídia e Carlos Guilherme, sem os quais nada disso seria possível. Agradeço também

à minha segunda família, os Togni, especialmente aos meus sogros Maria e Sérgio e à minha cunhada Pakisa, pelos cinco anos de acolhimento e afeição.

Aos colegas César Schirmer, Danilo Fraga e Marcos Klemz, agradeço por comentários e conversas sobre alguns dos assuntos desenvolvidos nesta dissertação. Aos colegas Cleber Corrêa, Marco Aurélio Alves e Mitieli Seixas, agradeço pela gentileza de me apoiar com o envio de artigos até então inacessíveis para mim. Aos professores Alfred Mele, Eric Schwitzgebel, John Campbell, Jordi Fernández e Maura Tumulty, agradeço pelo compartilhamento de materiais de difícil acesso ou, em alguns casos, cuja publicação ainda não havia sido realizada. Aos professores Tamar Gendler, George Graham e Neil Van Leeuwen agradeço pela sua atenção e amabilidade, e pela leitura e comentários aos ensaios que compõem esta dissertação.

À comunidade de desenvolvedores do sistema de preparação de documentos LaTeX, agradeço pelas ferramentas que me possibilitaram produzir esta dissertação utilizando excelentes recursos tipográficos e de editoração disponíveis gratuita e livremente.

Finalmente, agradeço cordialmente a todos os cidadãos brasileiros que, por dois anos, financiaram minha pesquisa mediante uma bolsa de mestrado concedida pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Introdução

A presente dissertação é composta por dois capítulos redigidos em inglês, que serão posteriormente submetidos a publicação como artigos separados.* Apesar de autossuficientes, estes ensaios possuem um objeto comum: o estudo dos limites de aplicação do conceito de crença mediante o exame de fenômenos que desafiam a atribuição de crenças aos sujeitos em que se manifestam. Os fenômenos escolhidos—o autoengano e os delírios estudados pela psiquiatria—são exemplos de estados mentais cuja inaptidão a desempenharem plenamente o papel funcional tido como essencial para que um estado mental seja uma crença põe em dúvida tanto o estatuto de crença (doravante: estatuto ou caráter 'doxástico') desses estados, quanto os critérios para a atribuição desta categoria psicológica.

Todavia, a mera anomalia desses estados, por si só, não constitui razão suficiente para concluir, sem argumento adicional, que tais estados não são crenças. Não obstante, ambos os lados do debate concordam que o papel funcional desempenhado por esses estados difere daquele desempenhado por crenças ordinárias ou paradigmáticas. Por um lado, aqueles que tendem a rejeitar explicações doxásticas mantêm que por mais que o papel funcional de tais estados se assemelhe àquele de crenças genuínas, ele não se assemelha o suficiente para tomarmos tais estados por crenças. Por outro lado, aqueles que tendem a aceitar explicações doxásticas apontam que essas diferenças não são tão importantes a ponto de excluir tais estados da extensão do conceito de crença. Porém, cabe àquele que deseja sustentar que certas anomalias *não* são incompatíveis com a atribuição de estatuto doxástico a um estado mental mostrar que as restrições tradicionais sobre a atribuição de crenças são problemáticas.

Mesmo supondo que o conceito de crença seja suficientemente preciso, poder-se-ia indagar *por que* deveríamos nos importar se o autoengano e os delírios constituem crenças anômalas, ou outro tipo de estado mental (e.g. imaginação), ou mesmo um estado mental misto. Poder-se-ia sugerir, com certa razão, que o que realmente importa para muitos propósitos é a questão de *qual* papel funcional os estados anômalos relevantes desempenham. Toda-

* Como faculta a Resolução no. 093/2007, da Câmara de Pós-Graduação da Universidade Federal do Rio Grande do Sul.

via, aquele que aborda esses fenômenos com o objetivo ulterior de incorporá-los a uma teoria compreensiva da crença encontrará relevância no exercício de resposta à questão discutida na presente investigação. Espero tornar claro, até o final desta, que perseguir uma resposta a essa questão é proveitoso, no mínimo, porque expõe o quão imprecisa é a aplicação das categorias da psicologia do senso comum (a assim-chamada '*folk psychology*'), cujo exame pode, assim, contribuir para a determinação das revisões necessárias com vistas a uma psicologia mais adequada. (A utilidade óbvia, por outro lado, é a melhor compreensão dos fenômenos anômalos em questão.) Em suma, fenômenos como o autoengano e os delírios são interessantes, entre outras razões, porque tornam manifesto que a extensão do conceito de crença está longe de ser homogênea e, portanto, sugerem uma análise mais sofisticada das condições de atribuição de atitudes proposicionais.

O que dá às crenças em especial seu papel central na filosofia da mente, na filosofia da ação, e na epistemologia é o fato de que são elas que norteiam nossas ações, nossas reações, e nossas cognições, incluindo tanto nossas hipóteses e inferências espontâneas quanto nossas reflexões mais ponderadas. Tomando emprestada a famosa metáfora de Frank Ramsey (1931, 238): crenças são mapas pelos quais nos guiamos. Endossar sinceramente uma proposição é um modo importante pelo qual nos guiamos, mas não é por si só suficientemente importante para justificar dar a um conceito definido exclusivamente nesses termos o papel central que tem, na filosofia e na psicologia, o conceito de crença. O que mais importa é como nos guiamos *em geral*. E no caso de fenômenos anômalos como o autoengano e os delírios, o modo como nos guiamos é marcadamente confuso e tudo indica que não há *um único mapa* coerente.

O interesse em propor uma investigação paralela do autoengano e do delírio provém do reconhecimento de que os desafios que esses fenômenos suscitam para os critérios mais seguidamente encontradiços na literatura filosófica e psicológica para a atribuição de crenças são similares. Além disso, enquanto patologias da cognição, o autoengano e os delírios evocam muitos desafios comuns para aqueles que buscam compreendê-los, e uma investigação concomitante de ambos traz à tona novas questões: esses fenômenos são inteiramente distintos, ou podem algumas formas de autoengano ser qualificadas como deliróides? Em qual medida podem modelos de explicação desses fenômenos compartilhar elementos comuns? Quais são as diferenças no modo como a emoção e a motivação influenciam a formação do autoengano e de delírios? Essas são questões importantes para a compreensão mais profunda de cada um desses fenômenos, e o seu tratamento é muito recente (cf. Bayne e Fernández 2009). Todavia, os ensaios que compõem esta dissertação pretendem dar continuação a esse debate focando-se em uma questão mais fundamental, a saber: que espécie de estado mental constituem esses fenômenos? Estamos autorizados a classificar como crenças estados anômalos que ostensivamente violam as normas da

racionalidade?

A pesquisa que resultou neste trabalho tinha originalmente como objeto os problemas de explicação do autoengano, e apenas estendeu-se ao estudo do delírio a partir da percepção de que a nova teoria do autoengano proposta por Tamar Gendler (2007) constitui uma adaptação da teoria do delírio avançada por Gregory Currie e colaboradores (2001, 2002). Paralelos continuaram a ser encontrados a partir da observação de que a crítica à teoria de Currie por parte de Tim Bayne e Elisabeth Pacherie (2005) tem consequências para a teoria de Gendler. Ainda, o uso que Bayne e Pacherie fazem da teoria disposicionalista da crença de Eric Schwitzgebel (2001, 2002) tem aplicação não só aos delírios, mas também ao autoengano (e a todos os outros fenômenos que desafiam a atribuição de crenças).[†] Finalmente, esta dissertação não constitui uma defesa ou um ataque às concepções doxásticas do autoengano ou do delírio, mas antes uma análise de argumentos recentes em favor de uma e outra teoria, focando-se em explicações cuja análise ou não foi ainda empreendida, o está sendo simulaneamente à produção deste trabalho.

[†] O primeiro desses paralelos é explorado no capítulo 1 desta dissertação. O segundo, no capítulo 2. Veja-se a Conclusão desta dissertação para direções para pesquisa futura com base no terceiro paralelo.

Capítulo 1

Belief and imagination in the explanation of self-deception

1.1 Introducing self-deception

Self-deception is a psychological phenomenon with which every human being is familiar (with the exception, perhaps, of those who are too good at it). There is no shortage of examples, but maybe it will help if we confine ourselves to those which are relatable and less intricate.[1] Consider the following:

> Martha has good evidence that her son has been killed in the war. His apparently lifeless body was sighted by a fellow-soldier 3 years ago, and Martha has not heard from her son despite the fact that the war ended a year ago. Yet Martha continues to insist that her son is still alive.
>
> Last year Justin left his wife of 40 years for his 25 year-old secretary. He says that his marriage had been on the rocks for decades, and that his wife will actually be happier without him. Those who know Justin and his wife well say that their marriage had turned sour only recently, and that his wife is devastated by the fact that she has been left for another, and younger, woman.
>
> Sonia has cancer, and has been told by doctors that she has 1 month to live. She avoids talking about the diagnosis, and continues to live as though her illness is merely temporary. She is saving money for a trip to see her son in one year; and refuses to put her affairs

[1] I was first drawn to the subject of self-deception through literature, especially Ibsen's *The Wild Duck* and O'Neill's *The Iceman Cometh*, and recommend these works as a source of beautifully rendered, ostensible self-deception. For analyses of these plays within investigations of self-deception, see Martin (1986, 110-6) and Neu (2000).

> in order despite the requests of her friends and family to do so. (Bayne and Fernández 2009, 2)

Notwithstanding our familiarity with this type of situation, and although the folk-psychological concept of self-deception is used by us every day, we have yet to come up with a successful explanation of all its aspects. And while it may not be obvious why philosophers (as opposed to experimental psychologists, neuroscientists and behavioral economists) should devote themselves to analyzing in depth the notion of self-deception, the simple explanation is that, while the experience is common and real enough, the concept of self-deception still eludes us.

The phenomenon itself has been portrayed since the earliest texts that have come down to us. The oldest allegory of self-deception I know is in the Hebrew Bible. Part of the second Book of Samuel narrates the weaknesses and failures of David's kingship, two of which were adultery and murder. Upon falling for Bathsheba, the beautiful wife of a righteous soldier named Uriah, David sends him out to the front where the ongoing war is fiercest. Uriah subsequently dies in combat, fulfilling David's plan, and as soon as Bathsheba is done mourning her husband's death, David brings her to his house and she becomes his wife. Displeased with this, God sends Nathan, the court prophet, to reprimand David. He does this by telling him a story:

> There were two men in a certain town, one rich and the other poor. The rich man had a very large number of sheep and cattle, but the poor man had nothing except one little ewe lamb he had bought. He raised it, and it grew up with him and his children. It shared his food, drank from his cup and even slept in his arms. It was like a daughter to him. Now a traveler came to the rich man, but the rich man refrained from taking one of his own sheep or cattle to prepare a meal for the traveler who had come to him. Instead, he took the ewe lamb that belonged to the poor man and prepared it for the one who had come to him. (2 Sam 11-12)

Upon hearing this, David is incensed and demands to know who the rich man is, and proclaims that he must pay for that lamb four times over and be sentenced to death. To what Nathan predictably replies that David himself was the man in the allegory, and he finally sees the light.

There are other ancient mentions of self-deception in such texts as Plato's *Cratylus* or Paul's *Epistle to the Galatians*, but Augustine was perhaps the first writer to actually develop the subject in the course of examining his own ways before conversion in the *Confessions*. He was the forerunner of a tradition of Christian writers who pursued the theme of self-ignorance—a tradition that continued with Pascal's *Pensées* down to many Seventeenth and Eighteenth Century British thinkers such as Daniel Dyke, Richard Baxter,

Joseph Butler, and John Mason.[2] The subject eventually got a more secular, if cursory, treatment in Adam Smith's *Theory of Moral Sentiments*. Despite the importance of the aforementioned works, however, none of them treated self-deception as a puzzling condition. To my knowledge, Kant was the first philosopher to recognize that there is something amiss with the very concept of self-deception. He formulated the now widely known puzzle attached to this concept in the second part of *The Metaphysics of Morals*:

> It is easy to show that man is actually guilty of many inner lies, but it seems more difficult to explain *how they are possible*; for a lie requires a second person whom one intends to deceive, whereas to deceive oneself on purpose seems to contain a contradiction. (1797/1996, 183, my emphasis)

This puzzle, however, did not give rise to a specialized debate up until the translation of Sartre's *Being and Nothingness* into English. In his discussion of bad faith, Sartre recognizes and elaborates on the same contradiction:

> I must know in my capacity as deceiver the truth which is hidden from me in my capacity as the one deceived. Better yet, I must know the truth very exactly *in order* to conceal it more carefully—and this not at two different moments, which at a pinch would allow us to reestablish a semblance of duality—but in the unitary structure of a single project. (1949/1957, 49)

Perhaps it's easier to envisage now why self-deception would hold interest for analytic philosophers (especially since Raphael Demos published his paper titled 'Lying to oneself' in 1960). As David Pears succintly put it, 'self-deception is an irritating concept. Its supposed denotation is far from clear and, if its connotation is taken literally, it cannot really have any denotation' (1984, 25). Which is to say that, apart from the very difficulty of arriving at a consensual definition, the very word 'self-deception' carries with it an air of impossibility if we take it to mean exactly what it seems to mean. On close inspection, two puzzles arise from a literal interpretation, each of which is derived from one of two lexical assumptions:

1. By definition, person A deceives person B (where B may or may not be the same person as A) into believing that p only if A knows, or at least believes truly, that $\neg p$ and causes B to believe that p.

[2] The major works on the subject by these authors are, respectively, *The Mystery of Selfe-Deceiuing: Or, a Discourse and Discouery of the Deceitfulnesse of Mans Heart* (1633), *On the Mischiefs of Self-Ignorance, and the Benefits of Self-Acquaintance* (1662), the sermons 'Upon the Character of Balaam' and 'Upon Self-Deceit' in *Fifteen Sermons* (1726), and *Self-Knowledge: A Treatise, shewing the Nature and Benefit of That Important Science, and the Way to attain it* (1745).

2. By definition, deceiving is an intentional activity: nonintentional deceiving is conceptually impossible. (Mele 2001, 6)

The first puzzle, then, arises from the recognition that if I deceive myself in the same manner in which I deceive someone else, it seems that I am in an impossible state of mind, namely, that of believing two contradictory propositions p and $\neg p$ simultaneously. (This is, of course, not to say that self-deceivers believe a contradition, but only that they have a pair of beliefs the content of which is logically incompatible.) The second puzzle, on the other hand, arises from the recogntion that if I literally deceive myself, it seems that I engage in the impossible process of intentionally bringing myself to believe something that I myself believe to be false. Alfred Mele (1987, 2001) has christened these problems the *static* and *dynamic* paradoxes of self-deception, respectively.[3]

In response to the aura of paradox that results from a lexical interpretation of 'self-deception,' some philosophers—such as Haight (1980), Kipp (1980) and Gergen (1985)—have become convinced of the impossibility of self-deception itself. And as much as we might think that the strategy of denying the existence of a commonly experienced phenomenon would have gone out with the days of literalistic ordinary language conceptual analysis, Steffen Borge has recently argued that 'there is no such thing as self-deception ... what has formerly been known as self-deception is rather a failure to understand, or lack of awareness of, one's emotional life and its influence on us' (2003, 1).

That said, there has always been a fairly general consensus in the literature that self-deception *does* exist. One of the most famous strategies that have been undertaken to explain literal self-deception has been developed by Pears (1984) and Donald Davidson (1985). Both their views rest on the Freudian idea that the best way to account for the phenomenon is to *split* the person. Pears, for instance, argues that

> [There is a] subsystem ... built around the nucleus of the wish for the irrational belief and it is organized like a person. Although it is a separate center of agency within the whole person, it is, from its own point of view, entirely rational. It wants the main system

[3] In what follows I will focus on the first of these problems; after all, my intended scope is the import of the examination of borderline phenomena for the study of belief, and not a full account of the phenomena discussed. The study of the dynamic by which self-deception occurs inevitably engulfs one in action-theoretic debates about intentionality which I put aside for reasons of space and relevance. Suffice it to say that the intentional character of self-deception has been challenged, since unintentional deception, far from being conceptually impossible, is commonplace (see Barnes 1997). Lying, on the other hand, *is* intentional by definition, and the second lexical assumption mentioned above stems exactly from a confusion of 'deception' and 'lying'. See Carson (2009) for a thorough treatment of definitions of lying and deception (and the related concepts of withholding information, keeping someone in the dark, bullshit, spin, and half-truths.

> to form the irrational belief and it is aware that it will not form it, if the cautionary belief [i.e., the belief that it would be irrational to form the desired belief] is allowed to intervene. So with perfect rationality it stops its intervention. (1984, 87)

It's easy to see how this would solve the pending difficulties. Pears converts the problematic characterization 'A deceives A' that resulted from a lexical reading into the non-problematic 'A deceives B,' where A and B are different subsystems of agency within a presumably unified system, namely, the person. Because the roles of deceiver and deceived are played by different centers of agency, the aura of paradox disappears. However, this sort of explanation faces its own difficulties. As Mark Johnston observed:

> The homuncular explanation replaces a contradictory description of the self-deceiver with a host of psychological puzzles. How can the deceiving subsystem have the capacities to perpetrate the deception? ... Why should the deceiving subsystem be interested in the deception? Does it like lying for its own sake? Or does it suppose that it knows what it is best for the deceived system to believe? (1988, 64)[4]

On the other hand, Davidson proposed what we may call a *functional* division, which bypasses the aformentioned charges to homuncular explanations. His view is that all that is needed is a boundary between conflicting attitudes—there would be no contradiction in believing contradictory propositions if they didn't come in contact with each other. Davidson claims that it is the drawing of such a boundary between our inconsistent beliefs which constitutes the irrational step involved in self-deception, and that this step is assisted by the nonobservance of what Hempel and Carnap called the *requirement of total evidence for inductive reasoning* (a normative principle that enjoins us to give credence to the hypothesis most highly supported by all available relevant evidence when choosing among a set of mutually exclusive hypotheses). The following passage contains the core of Davidson's account:

> An agent A is self-deceived with respect to a proposition p under the following conditions. A has evidence on the basis of which he believes that p is more apt to be true than its negation; the thought that p, or the thought that he ought rationally to believe p, motivates A to act in such a way as to cause himself to believe the negation of p. (1985, 88)

[4] Also because it assimilates the intrapersonal to the interpersonal case, the homuncular explanation brings about problems concerning motivation: 'If one of the subpersons (truly) believes that p and does not believe that $\neg p$, and if that subperson is bothered by this and wishes it were not the case, why would she find it psychologically fruitful intentionally to bring someone *else* to believe the opposite?' (Gendler 2007, 235).

Despite the modest amount of division required to make sense of this proposal, most authors since Davidson have abandoned the literal reading altogether. Many have devised theories of self-deception in accordance with the spirit of Davidson's proposal, but they have chosen not to draw the unwanted consequence that the subject believes *both* propositions involved. For proponents of this view, self-deception is not to be understood as a reflexive form of deception, in the same manner that self-teaching is not understood as a reflexive form of teaching. The 'deception' in self-deception must be understood as a *metaphor* (an observation that originated as early as Canfield and Gustafson 1962). Such theorists are, however, left with the task of explaining *what* exactly the mental states involved in self-deception are, and *how* they are formed and maintained: in other words, the task of explaining what the metaphor stands for. In their efforts to make sense of non-literal self-deception, the main disagreement has been over what kind of attitude is the *product* of self-deception (i.e. the unwarranted proposition to which the self-deceiver is motivated to give assent), and also on whether self-deceived subjects retain a belief in the *doxastic alternative* (i.e. the warranted proposition the content of which the self-deceiver is motivated to evade).[5]

Both Robert Audi (1982) and Georges Rey (1988) have argued that self-deception does not bring about belief in the usual sense:

> It does not appear to me that S's being in self-deception with respect to p and unconsciously believing $\neg p$ entails his believing ... that p. All my view requires regarding S's positive attitude toward p is that S be disposed sincerely to *avow* it (1982, 138, my emphasis)

An 'avowal' or 'avowed belief' means a disposition or tendency to endorse a propositional content verbally (either privately or publicly). This view avoids the static paradox because it takes self-deception to be a conflict between different kinds of attitudes, namely, central beliefs and mere avowals. It tries to explain how a self-deceived subject might sincerely speak about (say) not being cheated by an unfaithful spouse, while retaining a belief in the doxastic alternative. However, it does so by denying that self-deception entails other properties of proper beliefs, such as their deep connections to non-verbal action.

Finally, among the main approaches adopted to explain self-deception, the most widely espoused nowadays has been that which is sometimes called *deflationary* (Mele 1997) for its rejection of the literalist interpretation (and of the mysterious, homuncular solutions that have been proposed to the puzzles it engenders), and for its choice of abandoning the ascription of unconscious, inaccessible belief in the doxastic alternative.[6] According to this view, the

[5] I here follow the terminology employed by Van Leeuwen (2007).

[6] See also Johnston (1988), Barnes (1997) and Van Leeuwen (2007).

mental state and product of self-deception is simply a form of *motivated false belief*—the subject has only one belief, namely, belief in the proposition the self-deception is about. That is, a subject in the hold of self-deception is seen as actually believing the false or unwarranted proposition that is the content of his desire[7] (and not believing the true or warranted proposition for which there is sufficient evidence available). Support for this view comes from the fact that self-deceivers are usually sincere in their assertions; that, upon reflection, they will assert that they believe the relevant proposition; and that they will often act (and not only react) on the basis of the content of their self-deceptions, sometimes with dire consequences—while, by contrast, the avowal view is only able to explain verbal behavior.[8]

In what follows I will refer to this as the 'doxastic conception' of self-deception and will ignore the subtleties of different doxastic explanations to pursue the more fundamental question of whether 'self-deceptive belief' is a tenable notion at all.

1.2 What's wrong with self-deceptive belief

Doxastic accounts of self-deception have recently been met with criticism from Tamar Gendler (2007). She follows David Velleman (2000) in rejecting explanations of belief that distinguish it from other cognitive attitudes solely on the basis of its role in the motivation of action. In his paper, Velleman produces a number of examples with the aim of showing that other cognitive attitudes, most importantly propositional imagination (i.e. imagining that p), can motivate action in the manner taken by some theorists to be exclusive of belief. His main claim is that, seeing that the motivational role of belief is shared by other attitudes, only the fact that it aims at the truth can successfully distinguish it from the other attitudes (whose aim falls short of truth). Differently from merely surmising or imagining that a proposition is true, believing a proposition consists in bearing the attitude that one does with the aim of thereby accepting a truth. In complete agreement, Gendler claims that belief is inherently 'reality-sensitive' and that it consists in a 'receptive', as opposed to a 'projective', attitude. Self-deception, on the other hand, seems to be anything but reality-sensitive. On this basis she advances the following argument:

[7] An explanatory dispute that won't be covered in the following investigation concerns the nature of this desire (Mele 2001, 54-5), whether it is always coupled with anxiety (Barnes 1997, 117), and whether a desire really must always be posited, as opposed to, say, emotions (Lazar 1999).

[8] Of course, there are also disadvantages to this approach. For instance, this view may fail to account for both the epistemic *tension* usually thought to be inherent in self-deception and the *avoidance* behavior characteristic of self-deceivers. But see Mele (2001, 52ff.) and Van Leeuwen (2007).

> If belief is the attitude of accepting a proposition with the aim of thereby accepting something true, then in a wide variety of circumstances — most strikingly in cases of self-deception — our thoughts are occupied and our actions are guided by contents that we do not believe. The correct response to this observation is not to relax our standards for belief, but to recognize that other attitudes may play its characteristic role. Belief is not, as Hume avers, 'that act of mind which renders realities more present to us than fictions, causes them to weigh more in the thought ... gives them superior influence on the passions and imagination ... and renders them the governing principles of all of our actions.' (2007, 247)

But note that her conclusion is actually understated. If belief is indeed essentially truth-oriented as she assumes throughout her paper, then it seems that not only do we not believe the content of our self-deceptions, but we *cannot* believe it. If her argument succeeds, then it rules out belief as the kind of state self-deception consists in and consequently does away with most explanations of the phenomenon. Because of its sweeping consequences and because it has not been thoroughly discussed as yet, we turn now to fleshing out some of the details of Gendler's account.

1.3 Self-deception as pretense

While belief is introduced by Gendler as a characteristically truth-directed attitude, self-deception is one of the most blunt (or at least the most widespread) examples of the malfunctioning of our rational capacities. From this tension emerge the challenges which we have briefly presented and, from the alleged inadequacy of the doxastic conception of self-deception, emerges the alternative view that will occupy us from now on.

In reaction to what she perceives as a fatal blow to the doxastic conception, Gendler proposes a novel account of the characteristic cognitive attitude of self-deception.[9] Her overarching thesis is that self-deception would be best explained by appeal to a form of propositional imaginative *pretense*.

[9] A cognitive attitude is one that treats its content as true; hence beliefs, surmisals, assumptions, acceptances in a context, and also imaginings are cognitive attitudes. Their counterparts are conative attitudes, which treat their contents as to be made true; hence desires and wishes are conative attitudes. (I follow Van Leeuwen 2009 in this formulation, who borrows it from Velleman and Shah 2005). As much as self-deception involves (in most, if not all of its instances) one or more conative attitudes in its formation and maintenance, its end-product isn't itself a conative, but a cognitive attitude. It isn't a representation of how the subject wants the world to be, but a representation of how the world actually is for the self-deceived, and thus it possesses a far-reaching role in both the theoretical and practical reasoning of such subjects.

Her proposal thus belongs to the family of accounts that maintain that the self-deceived holds a true, temporarily inaccessible belief, and another false or unwarranted attitude that is *not* itself belief—thus avoiding the static puzzle. In this, her main predecessors are Audi (1982) and Rey (1988).[10] However, the specifics of her view are unprecedented. Also, she has the further, parallel aim of unmasking what she deems a 'failure to recognize the philosophical significance of a crucial fact about the human mind, namely, the degree to which attitudes *other than belief* often play a central role in our mental and practical lives' (2007, 231, my emphasis). Gendler summarizes her proposal in the following manner:

> A person who is self-deceived about $\neg p$ pretends (in the sense of *makes-believe*, or *imagines* or *fantasizes*) that $\neg p$ is the case, often while believing that p is the case and not believing that $\neg p$ is the case. The pretense that $\neg p$ largely plays the role normally played by belief in terms of (i) introspective vivacity and (ii) motivation of action in a wide range of circumstances. (2007, 233-4)

Thus, Gendler's self-deceivers do not come to believe the content of their self-deceptions, but engage in a form of mental *simulation*: their motivation to avoid the recognition of some truth or other ($\neg p$) leads them to mentally escape the real world and intermittently inhabit a 'p-world', an imaginary environment which protects them from the inconvenient or undesired evidence. The usual tension displayed by self-deceivers is explained by allowing what Gendler terms a 'projective' attitude (namely, pretense) to play a role in a context in which rationality would mandate that a 'receptive' attitude (namely, belief) do the work. The avoidance behavior characteristic of self-deceivers, in turn, can be explained by the retention, albeit tacit, of the true or warranted belief. Finally, the threat of contradiction or paradox is relieved by the appeal to something other than belief.

However, whatever Gendler means by 'pretense', it can't just be something like the daydreaming kind of fantasy which would seldom, if ever dispose the subject to sincerely avow its content. In self-deception it is expected that the subject not only evade the harsh truth and a sufficient part of the body of evidence pointing to it—something Gendler's pretenders in their avoidance

[10] Audi, Rey and Gendler's self-deceivers resemble what Raymond Smullyan calls *peculiar reasoners*: 'We will call a reasoner *peculiar* if there is some proposition p such that he believes p and also believes that he doesn't believe p. (This strange condition doesn't necessarily involve a logical inconsistency, but it is certainly a psychological peculiarity!)' (1986, 344). However, it will be important for the avowal/pretense view that the self-deceiver also have the false second-order belief that she believes that $\neg p$. Smullyan's discussion did not capture this further peculiarity, but perhaps this is just as well, since it is a matter of dispute whether such reasoners do in fact exist. But see Moran (2001).

and mental flight fitfully exemplify—but they also must abide by their self-deceptive attitudes, whatever they are. Hence, as Michel and Newen recently pointed out, to play the explanatory role it is meant to play,

> [Gendler's] new concept of 'pretense' must be a hybrid that eats its cake and keeps it, too. It has to be belief-like in explaining [the subject's] p-behavior and p-confidence, while being sufficiently imagination-like so as not to conflict with [the subject's] knowledge that p is untrue. (2010, 736)

Unlike Michel and Newen, however, I have no problem with the promiscuous proliferation of kinds of cognitive states and have no principled objection to Gendler's mixed propositional attitude. Having briefly presented the portions of her original account[11] which are more immediately important to what will follow, it is worth noting that what Gendler advances is very different from, say, allowing a *role* for pretense in the explanation of the *process* of (some forms of) self-deception. Hers is a bold and sweeping claim. According to her, self-deception just *is* pretense, that is, the product of self-deception is pretense (a state that is belief-like but that falls short of constituting full-blown belief).

I now turn to presenting reasons to resist Gendler's reading of Velleman, and hence her principled argument against doxastic conceptions of self-deception. Then, following work done by Michael Bratman and Neil Van Leeuwen, I also present reasons to resist Velleman's claim that imaginative pretense can play exactly the same role as belief in the motivation of action.

1.4 Assessing pretense

However strong her conclusion, Gendler explicitly acknowledges the fact that 'there are numerous ways in which belief can obtain without its normal manifestations,' and that 'it is certainly possible for someone to have false or ill-grounded beliefs' (2007, 236). While this is all too obvious, it warrants the following question: if we accept that these statements are true, what prevents us from being able to characterize self-deception precisely in these terms (which are in fact the belief theorist's terms)?

One reply might be that in self-deception, the aim of the subject's cognitive attitude is not truth—that the self-deceived doesn't bear the attitude that she does with the aim of thereby bearing that attitude toward something true. However, by the same token, beliefs formed out of wishful thinking would seem not to be beliefs as well. On the other hand, for Gendler, what marks the bearing of belief, well- or ill-grounded, is a willingness to submit

[11] My presentation is at best cursory and does not at all do justice to Gendler's account. I especially commend the more positive portion of it (see section 4 of her paper).

it to rational scrutiny, which flows from the fact of truth-directedness, and avoidance to do so is characteristic of self-deceived subjects, not of wishful thinkers (it has been argued that that is exactly where wishful thinking turns into self-deception). This constraint on the characterization of belief is well exemplified in the writings of Gregory Currie and colleagues on the parallel issue of the cognitive attitude of delusional subjects, and is perfectly consonant with Gendler's Velleman-inspired claim:

> If someone says that he has discovered a kind of belief that is peculiar in that there is no obligation to resolve or even to be concerned about inconsistencies between these beliefs and beliefs of any other kind, then the correct response to him is to say that he is talking about *something other than belief*. (Currie and Ravenscroft 2002, 176, my emphasis)

I propose that Gendler's reading of the truth-directed character of belief is flawed. Just as Currie and Ravenscroft's consistency constraint on belief-ascription in the quote above, her view of belief confuses the ideal for the real.

1.4.1 Velleman on belief and Gendler's appropriation

First of all, there seems to be sufficient textual grounds to conclude that Velleman himself would not subscribe to Gendler's reading of truth-directedness. Indeed, he states in a telling footnote that a 'person's cognition of being Napoleon might ... remain under the control of truth-directed mechanisms, which were being *diverted* from their goal; and in that case, he would literally have deceived himself, by self-inducing a false belief' (2000, 281). Though I am forced to disagree with Velleman's terms, since he is not presenting an example of self-deception, but of full-blown delusion[12] (where plausibly there is misperception, not just self-induction), his assertion seems to oppose Gendler's own claim that self-deception cannot be characterized by false belief, since he clearly allows the possibility of self-induced false belief.

The key element of this passage is his willingness to allow that, though beliefs may be essentially truth-directed, they may be diverted from their goal. This fact must be acknowledged if one is to allow the possibility of any kind of

[12] There seems to be an important difference between what is going on in the mind of someone who steadfastly holds on to the unwarranted belief that his wife is not having an affair, or that his son is not using illicit drugs, and someone who goes around thinking he is a dead historical figure. Where in self-deception we have a clear case of epistemic irrationality, in delusion we arguably have something deeper, a breakdown for which the agent may be exempt of (epistemic) responsibility. I elaborate on this subject in the second essay of this dissertation.

misbelief, from simple error to full-blown delusion, with self-deception somewhere along the spectrum. In such cases, someone's cognition may deviate from and fall short of its goal for any number of reasons. The moral, I take it, is that it is *extrinsic* to the truth-directedness of belief whether it arrive at falsity by whichever means, be it negligence, bias, misinterpretation, etc. In Velleman's own words: 'Faulty or mistaken beliefs are the ones whose regulation has not succeeded in producing the kind of cognitions that it was designed to produce' (2000, 278). I suggest that self-deception just is one of many situations where regulation breaks down, that is, where epistemic rationality fails. Of course, this is not a new idea. The breakdown of regulation is one of the main aspects of self-deception that has interested epistemologists since Donald Davidson (1982, 1985).

Moreover, Velleman claims (from introspection) that when 'we discern a gap between a belief and the truth, the belief immediately becomes unsettled and begins to change' (2000, 278). Though I agree that this certainly happens (after all, we revise our beliefs constantly), and that it mirrors what Gendler has to say about reality-sensitivity and belief's willingness to be submitted to rational scrutiny, this unsettledness, through the operation of various defense mechanisms available to us, sometimes leads us astray. In cases of self-deception we are described as evading the recognition of the aforementioned gap between our wishful, unwarranted beliefs, and reality. Velleman finishes his thought stating that if it persists, 'we form another belief to close the gap, while reclassifying the recalcitrant cognition as an illusion or a bias' (ibid.). At least at first blush, I see no reason to reclassify the recalcitrant cognition as an *imaginative* state, just as I see no reason not to classify it as a false belief.

Velleman's description seems to fit well our common experience of what occurs when we finally realize that we are or have been wrong, which of course includes the particularly painful case of coming to terms with, and out of, self-deception. In fact, it seems more straightforward than Gendler's account of the related phenomenology. She describes it as involving 'the feeling of leaving a realm of make-believe where one has allowed one's thoughts and actions to be governed by some sort of fantasy, and returning to the realm of reality-receptiveness where one's thoughts and actions are responsive to things as they actually are' (2007, 245). As much as this might not be a bad *metaphor*, I cannot see how make-believe, as opposed to misbelief, naturally finds its way into such a description. Thus I think that in describing the related phenomenology we should keep the feeling of going from reality-insensitivity (instantiated in our biased inspection of the world) to reality-sensitivity, and replace fantasy references for misregulated false belief. Authors such as Ariela Lazar (1999) have indeed understood their own references to fantasy as ultimately pointing to motivated false belief on the part of the self-deceived.

In keeping with the claims I have examined so far, Gendler affirms that

pretense is the attitude of the self-deceived because 'if it were a receptive attitude (belief), the norms governing such attitudes would mandate that it be abandoned (on grounds of falsity)' (2007, 245). This would seem to imply that all irrational (or otherwise faulty) attitude formation involves some attitude (presumably pretense) other than belief. In a footnote, however, Gendler seems to point out that she does not aim at such a conclusion: 'Beliefs formed as a result of cognitive biases ... may well be defective in certain ways' (2007, 254 fn. 46). This seems to contradict her own problematic argument. She follows Lazar's distinction between attitudes formed out of bias and self-deceptive ones: 'cognitive biases are persistent patterns of biased reasoning. They are exhibited regardless of subject-matter. In contrast, self-deception is thematic: the content of the irrational belief is relevant to the explanation of its formation' (1999, 267). But Lazar oversimplifies, since a whole family of relevant cognitive biases, so-called 'hot' biases, is characteristically thematic.

Some biases are self-directed, which confers relevance to the content of the subsequent biased belief. Examples are all too familiar. The *self-serving bias* is one such behavioral pattern, in which people attribute their successes to internal or personal factors but attribute their failures to situational factors beyond their control (Miller and Ross 1975). The *Dunning–Kruger effect* is another such pattern, one whereby unskilled people make poor decisions and reach erroneous conclusions, but their incompetence denies them the metacognitive ability to appreciate their mistakes (Kruger and Dunning 1999). The unskilled therefore develop the further bias of *illusory superiority* (Hoorens 1993), rating their abilities as above average, sometimes much higher than they actually are, while the highly skilled underrate their own abilities, suffering from *illusory inferiority*. Driving a car, socializing and solving logic problems are common examples of such abilities. This bias provides an explanation to why actual competence may weaken self-confidence—or in Darwin's words, why 'ignorance more frequently begets confidence than does knowledge' (1871/1981, 3)—since competent individuals falsely assume that others have equivalent abilities. Kruger and Dunning conclude that 'the miscalibration of the incompetent stems from an error about the self, whereas the miscalibration of the highly competent stems from an error about others' (1999, 1127).

Gendler contrasts the biased subject to the self-deceived subject by pointing out that the latter 'will not readily accept that her attitude toward $\neg p$ is unjustified or illegitimate, even when this is pointed out to her' (2007, 254 fn. 46). The problem here seems to be that Gendler is comparing the self-deceiver to the irrelevant kind of biased reasoner, namely the 'cold' one. A person taken in by a motivated (hot) form of bias, such as illusory superiority will arguably exhibit the same, or at least a similar, kind of persistence since the subject-matter is indeed relevant to her attitude.[13] Ultimately, it seems Gendler is

[13] On the role that bot cold and hot biases play in the formation and maintenance of

faced with the following dilemma: to choose between either her hyperbolic conception of truth-directedness, or the possibility of beliefs formed out of motivational bias. If one should take the first route, however, it would follow that a person with a tendency to overestimate their positive qualities and abilities and underestimate their negative qualities would not be able to be characterized as really believing the content of her attitude. This further restriction of the domain of false belief seems even more implausible than claiming that the self-deceived don't believe the content of their self-deceptions.[14]

What can be learned from Gendler's reading of truth-directedness? While beliefs may be truth-directed and nevertheless be false in a number of ways, as she falteringly acknowledges, I would like to point out the importance of attending to the difference between belief (a state) and belief-formation (a process) when talking about truth-directedness. While one may hold that it is of the essence of beliefs to be reality-sensitive and represent our rational commitment to the world, it is patent that belief-formation can nevertheless find ways to go awry (some of them having something to do with the person's desires), as is recognizable in Gendler's dissonant affirmations. What this means is simply that some of our beliefs are not the product of ideal, perfect rationality—actually, *most* of them are not (Cherniak 1986). The existence of biased processes of belief-formation does not mean that truth-directedness is compromised, nor does it really have anything to do with it. Perfect regulation is no way to model our actual belief-formation processes. Furthermore, why would resistance to rational scrutiny be a hindrance toward belief-ascription? If anything, it should provide evidence to the ascription of irrationality. But as we have pointed out, Gendler does away with the very possibility of there being such a thing as irrational *beliefs*.

If what has been said is tenable, then a doxastic account of self-deception is not at all ruled out by accepting Velleman's claim of truth-directedness, but is perfectly consistent with it. I take this to be enough to reinstate belief as a plausible candidate. It does not follow from truth-directedness that beliefs can't be formed in non-truth-directed ways. Hence, it does not follow from truth-directedness that the product of self-deception cannot be, or is not, belief. However, this alone obviously does not speak against Gendler's pretense account. The success of Gendler's positive proposal partially hangs on whether her chosen attitude can do the required work left by giving up belief; if it does, then that would be enough independent sustenance for working with Gendler

self-deceptive belief, see Mele (2001, 25-31).

[14] Which is of no greater moment than the fact that mistaken subjects actually believe the contents of their misconceptions. Here the belief theorist's observation would be that Gendler's approach has the unwelcome consequence that her self-deceivers are never actually mistaken in their attitudes toward the self-deceptive content, which seems incongruent with our everyday experience and evaluation of self-deception in ourselves and in others.

in building a pretense account.[15] For it to work, however, one needs to show that imaginative pretense can have, as she puts it (following Hume's characterization of belief), the introspective vivacity and the motivational role of belief. While I don't deny that imaginative pretense can indeed be introspectively vivid, I will not take issue here with the *identity of introspective vivacity* between belief and imaginative pretense. I will, however, take issue with her claim that belief and imagination can and do share motivational role, jointly with desires, in the production of action. Gendler does not herself present any argument for the *identity of motivational role* between belief and pretense, but Velleman does—in fact, arguing for this thesis takes up most of his paper. Its importance is crucial in his argument for abandoning motivational role as a distinguishing feature of belief. For Gendler, the thesis is just as important: proving it wrong would constitute a heavy blow to her account. An examination of that which Gendler takes for granted will require us to delve deeper in Velleman's argument.

1.4.2 The motivational roles of belief and imagination

Beliefs, in conjunction with desires, cause and rationalize actions that will make the contents of the desires true, if the contents of the beliefs are true. If I believe there is Marsala wine in my cellar, I will go to it and pick up a bottle, provided I want to. Hence the conative contribution to action is pretty straightforward. What about the cognitive side? From experience, I know I have acted on many other, lesser kinds of commitment. I have gone to the store based on the mere surmisal that they sell mascarpone cheese. And I have talked out loud based on the mere imagining that I am discussing issues in my relationship with a close friend. So it is safe to say, to begin with, that other cognitive attitudes affect behavior, jointly with conative attitudes, in ways that are similar to the ways belief do.

Nevertheless, it is intuitive enough to think that they do not have the same role in producing action. Let us focus on imaginings: can an imagining that I have won a hundred million dollars in the lottery and a belief to the same effect be equal in their output? Do I behave in the same way as a consequence of holding each of these cognitive attitudes? It would seem absurd to think so. If I imagine, as in a daydream, that I have won the lottery, I may as a consequence imagine myself buying a luxurious apartment with an ocean view in Ipanema, as well as quitting my job, etc. Depending on the vividness of this mental simulation and the intensity of my desire, I might even go online and check out some real estate websites: not because I am about to actually

[15] To justify abandoning talk of belief altogether, however, there must be some advantage in giving it up. As we have seen, that advantage for Gendler is providing an explanation compatible with the view that belief essentially aims at the truth.

buy one, but because it might feel good to carry on daydreaming (that is, simulating the experience). On the other hand, if I really believe so, such a degree of conviction has markedly different consequences: not only may I entertain buying that house and quitting my job, I may actually call a real estate agent, make a deposit, announce to my family that I am moving, bid farewell to my colleagues, stop working on this paper, etc. What to make, then, of the following assertion by Velleman and Nishi Shah (in a follow-up paper to Velleman's 'On the Aim of Belief')?

> The question is how to differentiate the concept of belief from the concepts of other attitudes that involve regarding-as-true [i.e., cognitive attitudes]. The answer cannot be that belief plays a distinctive motivational role, because the motivational role of belief is one that it shares with other cognitive attitudes. Assuming that p and supposing that p resemble believing that p in that they dispose the subject to behave as if p were true; and even imagining that p may resemble belief in this respect. (Velleman and Shah 2005, 497-8)

At first sight my little scenario seems to contradict such claims. So far we have produced at least a distinction of degree between believing and imagining with respect to their role in motivating action, that is, we have intuitively created a hierarchy of motivational forces and placed belief higher than all other attitudes. Even if one accepts the identity thesis, beliefs patently are the standard background for our actions. This betrays an ambiguity in what 'motivational role' might actually mean. So far it is unclear what may be embedded in the word 'resemblance' in the quote above. After all, what is relevant to our discussion is not whether belief and imagining can have similar effects on behavior at a high level of abstraction. Velleman needs to abide by what Van Leeuwen has called the *identity of comprehensive motivational role* thesis. This need can be seen once one fleshes out Velleman's argument. Van Leeuwen (2009) provides a very clear formulation of the overall argument of 'On the Aim of Belief':

1. If belief cannot be distinguished from other cognitive attitudes by its role in action output, then it must be distinguished from them by etiology or cognitive input, i.e., regulation and production. [premise]
2. Belief and imagining have the same motivational role, i.e., 'conditional disposition to cause behavior,' a role shared by the other cognitive attitudes as well. [lemma argued for in the paper]
3. Therefore, belief cannot be distinguished from other cognitive attitudes by its role in action output. [from 2]

4. Therefore, belief must be distinguished from other cognitive attitudes by cognitive etiology, i.e., regulation and production. [from 1 and 3]
5. Aiming at truth, i.e., being regulated by mechanisms designed to produce truth in beliefs, is the best candidate among cognitive properties that could distinguish beliefs. [intuitive assumption, argued for briefly by Velleman in 'Answers to Objections']
6. '... truth-directedness is essential to the characterization of belief.' [from 4 and 5] (2009, 230-1)

Since the motivational roles of belief and imagining (and other attitudes) are shared, Velleman argues, we must turn to the notion of truth-directedness to distinguish belief from every other cognitive attitude. Velleman's strategy for demonstrating that 2 is true, as mentioned earlier, consists in pointing out, through a series of examples, that other attitudes besides belief have output in behavior and action. For example, there is make-believe, as when a child pretends that she is an elephant, waving her arm like a trunk, drinking from an imaginary pail of water, etc.; there is talking to oneself, as when we walk down the street discussing with an imaginary interlocutor or when we address an imaginary audience as we work on a conference paper; there are psychoanalytic examples, as in a number of cases catalogued and interpreted by Freud, in which a patient behaves and acts motivated by wishful fantasies, as when a jealous child symbolically throws out his baby brother, which was how Freud interpreted Goethe's earliest childhood memory of throwing crockery out a window and watching it smash in the street (1917/1958); and, finally, there is expressive behavior, as in Hume's famous example of the dangling cage in *A Treatise of Human Nature* (I.3.13), where someone who is suspended at a great height trembles with fear and holds on to the bars of the cage, despite acknowledging that she is securely supported.[16]

However, logically, the step from 2 to 3 cannot be made unless it can be shown that belief and imagining have the same motivational role, that is, the same disposition to cause behavior, in a comprehensive sense, in which case 'motivational role' would mean all characteristic effects an attitude of a given kind has on behavior (Van Leeuwen 2009, 232). If Velleman does mean the comprehensive sense, it follows that, other things being equal, imagining that p will cause the same behaviors as believing that p. So the fact that he can logically extract 3 from 2 is of no help: his argument may be logically valid, but it hangs on a premise which is easily shown to be absurd.

[16] Gendler (2008) describes a similar case taking the behavior expressed by tourists when walking the Grand Canyon 'Skywalk'. She coins a new term, 'alief', and argues that this is a different kind of attitude from the attitudes discussed so far in the literature.

On the other hand, it is possible that Velleman means 'motivational role' to be read in a minimum and highly abstract way, which Van Leeuwen calls the *vanilla* sense, in which 'the motivational role of a belief is to cause behavior that will satisfy conations if the belief is true' (ibid.). His examples indeed serve well the purpose of showing that there are some circumstances in which we act in ways that would make our wants satisfied if the contents of our imaginings were true. But is that enough to demonstrate the truth of 2 (taking the vanilla sense into consideration)? While that is an interesting question, even a positive answer to it would not be enough to save his argument, since it would not be logically valid: from a vanilla interpretation it follows that he has not ruled out the possibility that belief and imagining can be distinguished in terms of action output (something which only the comprehensive interpretation can grant). As Van Leeuwen puts it, if we read him as proposing only a vanilla sense of 'motivational role', Velleman wins a battle, but ultimately he loses the war.

What exactly are the consequences of this discussion to Gendler's explanation of self-deception? We have seen that she does not offer independent argument for the identity thesis. The question now is: does Gendler also need the identity thesis to be read in a comprehensive sense, or can her account succeed with appeal only to the vanilla sense? The answer to this question depends on the specifics of self-deception, that is, on what the output in action of whatever attitude self-deceived subjects hold actually is. It has been shown that identity of comprehensive motivational role is plainly false. But that would not represent a problem for Gendler provided that in claiming that pretense can play the role of belief in the motivation of action in a wide range of circumstances she abides by a minimum sense of 'motivational role'.

I think the absence of an all-inclusive clause in her formulation warrants us reading Gendler to mean something other than unrestricted identity of motivational role between belief and imagination. However, her clause 'in a wide range of circumstances' limits the subset of circumstances where identity holds, not the kind of identity itself. Where most theorists would ascribe full-fledged belief to a self-deceived subject (such as the one who acts on her self-deceptive beliefs about entrepreneurial success, to disastrous financial and personal consequences), Gendler does not. Hence, if she says pretense is what characterizes the mental state of the self-deceived, one can only conclude that, for her, self-deception is the type of case where an attitude besides belief plays the exact same role of, and attains comprehensive motivational role with relation to, belief.

A distinction presents itself. In order to salvage Gendler's formulation, it must be said that, while comprehensive identity across all settings is at the very best implausible, it can nevertheless hold in localized instances (a subset of which is self-deception). To assess whether or not this is tenable we must try

to peer at the essence of belief's relationship to all other cognitive attitudes.

1.4.3 Distinguishing belief and imagination

It is of course possible that, notwithstanding the shortcomings of Velleman's argument in proving it, truth-directedness is at least a causal, if not normative feature of beliefs.[17] That is not relevant for my purposes, however, since it has been established that truth-directedness does not rule out the possibility that self-deceived subjects actually believe the contents of their self-deceptions. Furthermore, it is perfectly possible that truth-directedness is true *and* that it is possible to distinguish belief from all other attitudes based solely on motivational role. So far we have seen that Velleman has not properly shown the latter to be false. My present purpose then is to assess whether it is true.

As a background to what follows we should briefly review the standard characterization of the motivational view of belief. In Velleman's understanding of it, 'All that's necessary for an attitude to qualify as a belief is that it disposes the subject to behave in ways that would promote the satisfaction of her desires if its content were true. An attitude's tendency to cause behavioral output is thus conceived as sufficient to make it a belief' (2000, 255). His many examples are enough to show that this view is overly simplistic and erroneous. Nevertheless, in the face of his main argument's lack of success, it would be prudent not to do away with the motivational view all at once.

The problem we are facing, the challenge of finding a way to properly distinguish belief and imagination, can be traced back at least to Hume's *An Enquiry Concerning Human Understanding*:

> Wherein, therefore, consists the difference between such a fiction and belief? It lies not merely in any peculiar idea, which is annexed to such a conception as commands our assent, and which is wanting to every known fiction. For as the mind has authority over all its ideas, it could voluntarily annex this particular idea to any fiction, and consequently be able to believe whatever it pleases; contrary to what we find by daily experience. (V.II)

Hume's solution to the problem, as I see it, would rest on a conception of belief directly attacked by Velleman and Gendler. On Hume's view, belief is 'that act of mind which renders realities more present to us than fictions, causes them to weigh more in the thought ... gives them superior influence on the passions and imagination ... and renders them the governing principles of all our actions' (ibid.). Gendler invokes this passage at the beginning of her paper and aims at making an example out of Hume, so to speak, by showing that

[17] See Engel (2004).

successfully demonstrating self-deception to be characterized by an imaginative attitude proves he was ultimately wrong: usurpers, says Gendler, 'do not always deserve the title of the one whom they usurp' (2007, 247). Furthermore, she thereby aims to denounce a widespread ignorance of the role played by a range of other attitudes in the cognitive economy of humans. While I strongly agree, being averse to any oversimplification in our understanding of human cognition, I think there are a couple of things that can be said in favor of a generally Humean view of belief and imagination. That is, the view that belief is, among other things, 'the governing principle of all our actions': a feature that distinguishes it from every other cognitive attitude; a role no usurper can ever play, but only mimic to a degree.

So far we have noticed that there is something missing in the standard characterization of the motivational view, and we know such a view can't be true. On the other hand, as Lucy O'Brien rightly notes, there is something missing also in Velleman's view that imaginings can play the motivational role of beliefs. 'The attitude of imagining that p, by itself and relative to a fixed background of desires, does not dispose the subject to behave in ways that would promote the satisfaction of his desires if its content were true' (2005, 58). That is to say, comprehensive identity of motivational role is false. Also, says O'Brien, 'it seems to be a quite general point that any 'regarding as true' [i.e. cognitive] states which are not beliefs, will require [a] kind of connection to the subject's beliefs about his actual world if they are to result in action' (ibid.). This observation sets the tone for the rest of the discussion in this section, namely, that what must be appended to the standard account is an account of the particular relation in which beliefs stand to other cognitive attitudes, a relation that (sometimes) confers upon the latter the capacity of producing output in behavior and action. Drawing on work by Bratman (1992), Van Leeuwen (2009) takes strides toward a unifying account that exhibits what is needed for an adequate explanation of what the motivational role of belief is.

1.4.4 Context and practical ground

The main problem for the motivational view is that every cognitive attitude is apt to play some role in the motivation of action. We do act on the basis of surmisals, as when we want to eat something and get up to look for it in the fridge, assuming that there is something there (but are not quite sure). It might be argued that this is nothing but a degree of belief this side of certainty. Even so, and more relevant to the present discussion, we also act on the basis of imaginings, as when we play a game of make-believe and swifly dodge an imaginary sword, despite the fact that our friend is just putting his hands together *as if* he were carrying a sword.

To provide an answer to Hume's problem, we may begin with Bratman's

examination of the difference between belief and what he calls 'acceptance in a context':

> An agent's beliefs provide the *default cognitive background* for further deliberation and planning ... this cognitive background is ... context independent. But practical reasoning admits adjustments to this default cognitive background, adjustments in what one takes for granted in the specific practical context... To be accepted in a context is to be taken as given in the adjusted cognitive background for that context. (1992, 10-11)

In a previous article, Van Leeuwen suggested extending this idea of Bratman's to include all other non-belief cognitive attitudes, resulting in the statement that 'non-belief cognitive attitudes require specific contexts in order to function as the background of deliberation for the constitution of action' (2007, 434). This implies that whenever an imagining prompts action, it does so only *in the context* of a game of make-believe (or whatever kind of imaginative play is at work). Likewise, whenever a surmisal prompts action, it does so only in the context of an investigation. Beliefs, however, are *context-independent*, and for this reason, they are the default cognitive background for the constitution of action.

An extension of Bratman's idea, Van Leeuwen's *practical ground thesis* states that belief is the practical ground of all other non-belief cognitive attitudes in circumstances wherein the latter prompt action. This is to say that, while one may have the impression that non-belief cognitive attitudes prompt action on their own (as Velleman compellingly argued), they do so only in virtue of being grounded on belief. Beliefs are, to use Ramsey's metaphor, maps by which we steer (1931, 238). Furthermore, Van Leeuwen adds that beliefs determine if one is in the right *setting* for acting on the basis of another attitude or not. Thus, he observes that 'beliefs are not the only maps by which we steer the ship; they are also maps by which other maps are chosen and appraised' (2009, 239).

Van Leeuwen presents the practical ground relation as the conjunction of three types of relation that can hold between classes of cognitive attitudes:

1. Attitudes of type X are available for motivating actions across all practical settings, while attitudes of type Y depend on the agent's being in a certain practical setting to be effective in influencing action.
2. Attitudes of type X represent the *practical setting* one is in such that one acts on attitudes of type Y on account of being in that setting.

3. Attitudes of type X are the cognitive input into choosing to act with attitudes of type Y as input into practical reasoning, when one does so choose. (2009, 226)

Therefore, X is the practical ground of Y just in case all three relations hold, and this is precisely the case for the ordered pair <Belief, Acceptance in a Context> and, by extension (Van Leeuwen claims), for the ordered pair <Belief, Imagining>—but never for <Acceptance in a Context, Belief>, <Imagining, Belief> etc.

Van Leeuwen does us the favor of fleshing out his argument for the practical ground thesis in explicit and logically clear form. It begins with the already argued-for premise that the identity of comprehensive motivational role thesis is false, from which it follows that there are practical settings in which the behavioral consequences of imaginings differ from that of beliefs.[18] Moreover, Van Leeuwen argues that these differences in output are non-accidental, since practical setting typically influences behavior in conjunction with cognitions. It follows that they are caused, at least in part, by differences in the very practical settings in which they occur, which implies that either the psychomechanical efficacy of imagining is practical setting-dependent, or the psychomechanical efficacy of belief is, or both.[19] Given that the psychomechanical efficacy of belief is not context-dependent (as Bratman has shown), the psychomechanical efficacy of imagining is. From this pair of statements Van Leeuwen extracts the first of three lemmas in his argument:

> ***Lemma 1***. Beliefs are effective in practical reasoning and motivating actions in a practical setting-independent way, while imagining depends on practical setting to be effective in influencing action.

Short of a 'magical connection' (2009, 237 fn. 19) between being in a practical setting and having that practical setting activate a specifically adjusted cognitive background for it, it must be assumed that the fact of practical setting-dependence mandates that the agent *represent* the practical setting she is in. Given the practical setting-dependence of imagining, it follows that an agent who acts on the basis of imaginings must have a representation of the practical setting she is in. Now, whatever they are, representations of practical setting are certainly cognitive, as opposed to conative, attitudes. However, given that non-belief attitudes are themselves practical setting-dependent, representations of practical setting must be practical setting-*independent* (since otherwise we would have an infinite regress of representations of practical setting, something which contradicts the simple fact that humans are finite). This,

[18] Here and for the remainder of the argument, 'imaginings' is meant as a shorthand for 'imaginings and other non-belief cognitive attitudes'.

[19] 'Psychomecanical efficacy' is Van Leeuwen's term for the property of being effective in influencing action.

in conjunction with the first lemma, yields a second lemma in Van Leeuwen's argument:

> ***Lemma 2***. There are beliefs that represent the practical setting an agent is in, on which the psychomechanical efficacy of imagining is dependent.

Now, one of the important conclusions he derives from a story about him and his childhood friend Chris (2009, 227-9)—which illustrates the third type of relation needed for the practical ground relation to hold—is that acting with imagining as the adjusted cognitive background is a *choice* (whereas being taken in by imaginative play is often involuntary). In short: while playing a game of make-believe in the mud, his friend gets stuck. Initially, the protagonist believes that this is part of the game and carries on. But once his friend informs him that he is really stuck, he turns to acting on other beliefs, for instance, the belief that by helping him he would become unstuck, as a consequence of which he goes over to get Chris unstuck. Believing, then, that Chris was successfully unstuck, and that the game of make-believe could be resumed, the protagonist makes the choice (against the background of these beliefs) to resume it. Given that attitudes of the kind that represent the practical setting an agent is in are cognitive inputs into choices the agent makes in those settings, a third and final lemma is derived, whence the conclusion of the argument is finally extracted:

> ***Lemma 3***. Beliefs are cognitive inputs into choosing to act with imaginings as the adjusted cognitive background, when one does so choose.
>
> ***Practical ground thesis***. Belief is the practical ground of imagining. [from lemma 1, lemma 2, lemma 3, and the definition of practical ground ...] **QED**

This concludes the paraphrasis of Van Leeuwen's argument. I now turn to filling in the gaps in Gendler's account by looking at how Van Leeuwen's results affect it.

1.4.5 The practical ground of self-deception

Can imaginative pretense, or fantasy, or make-believe sometimes play the role that belief does in the motivation of action? So far we have seen that pretense (and other non-belief cognitive attitudes) can and do play a role in motivating action in a wide range of cases. On the other hand, we have learned from Bratman and Van Leeuwen that that doesn't warrant ascribing them the precise same role. If the practical ground thesis is true, that is, if belief is the practical ground of imagining, then every time we perform an action on the basis of an

imagining, we have beliefs that represent the practical setting we are in, on which the psychomechanical efficacy of said imaginings depends.

This presents a problem for Gendler's account, since self-deception by definition requires a certain (variable) level of ignorance of one's own situation. Provided that Gendler does claim that imaginings can be psychomechanically effective (and are so in cases of self-deception), she is thereby pushed to accept that, in self-deception, the agent has beliefs that represent this peculiar type of practical setting. While this is of course compatible with a general theory of make-believe, for Gendler's account it means that self-deceived agents, to act (as they do) on their supposed pretenses, are required to believe that they are pretending, or fantasizing, or making-believe. However, it is plainly impossible to be motivated to act on our self-deceptions if not only do we not believe their content, but we simultaneously believe that we only pretend that such content is true (or, to use Gendler's terms, that we are in a world where what we want to be true is true).

Another difficulty arises by the application of Van Leeuwen's framework to Gendler's account, since acting with imagining as the adjusted cognitive background is supposedly a choice. I take it that this does not imply that we cannot cry while watching a film—imagining that James Stewart's character in *It's a Wonderful Life* really is delivering a beautiful speech—or mumble something as we walk down the street—imagining that we are finally letting our friend know what we think of his drinking habit. No, these are things we do while daydreaming or so, and we do them involuntarily. Someone who voluntarily speaks his mind out loud while walking down the street is not in the grasp of imaginative pretense, but something altogether different.

What Van Leeuwen claims, on the other hand, is that voluntary action in the adjusted cognitive background of imagining, such as when we make-believe we are conductors directing Beethoven's Ninth—as a consequence of which we try to wave our hands in the fashion of a great conductor—is a choice. The question then is: does Gendler have in mind the mumbling-down-the-street, involuntary type of action when she describes the workings of self-deception? Or does she mean the waving-hands-as-if-conducting, voluntary type of action? While one hardly could mean the first sense, it is worth delving briefly into everyday cases of self-deception to see more clearly why one could not.

First, let us picture a businessman who overestimates his own skills, underestimates (and often just ignores) the failure of his past and current enterprises, and who is very much in debt. He often talks about opening up new businesses, much to the dismay of his family, and has a knack for finding improbable ventures and locations, such as selling car batteries in a beach town whose population drops from 30,000 in the summer to 3,000 throughout the rest of the year. Nevertheless, each time he seems firm in his conviction that his business ideas will prove successful and lucrative (and each and every time

he is proven wrong). He is not just stubborn, but adamant, and won't listen to reason, won't extract from the evidence the same conclusion anyone else would, won't heed to the advice of many friends who cannot bear to witness him come close to the bottom. He is self-deceived, and his most recent course of action while in the tight grasp of this state of mind was to request a considerable loan. Cases like these are widespread, this kind of thing happens every day. It is not in the best interest of people like the businessman to perpetuate and plunge even further into his already outstanding debt. But he deliberately does exactly that.

Second, I want to evoke a case where the misfortune lies not so much in acting, but in deliberately failing to act. A single mother who welcomes a man into her home and, despite the heaping evidence that her daughter's new stepfather is crossing the line between innocent and lustful affection, refuses to acknowledge that he might be sexually interested in her daughter. This is also a situation which is not uncommon and that can have appalling consequences. Let me elaborate. The mother is not subject to full-blown delusion. She does notice that the way her boyfriend treats her daughter is increasingly aggressive. She also notices that her daughter's face shows great distress when she is asked about the subject, and that her behavior has rapidly changed from that of a docile to that of an injured and indignant child. Again, anyone else, given the same amount of evidence she has, would be quick to conclude that if something grim has not already happened, it is about to. But the mother does nothing. She refrains from asking any more questions, since just thinking about the subject upsets her a great deal. However, she is not at all devoid of motherly love: she is just blindly in love with the man, who responds with violent indignation to the mere hint of her preoccupation. As with the businessman, her actions are the product of self-deception, but differently from my previous scenario, the mother sins by omission rather than by proactively doing anything.

Can deliberate courses of action like these be taken on the basis of mere imaginative pretense? Van Leeuwen (2007) has asked a similar question, but he did not aim it at pretense but rather avowal. As I said earlier, Gendler's account bears some resemblance to Audi's and Rey's; the reason I say so is that it takes the doxastic alternative to be the object of a tacit or unconscious belief, while claiming that the product of self-deception is not itself belief, but a weaker attitude. In that article, Van Leeuwen criticized the avowal view, arguing that it is absurd to propose that the action patterns as serious as those just now portrayed could happen as a consequence of holding something short of full-fledged belief. The way Audi and Rey define avowal, however, justifies Van Leeuwen's criticism in such a way that we cannot simply apply it to Gendler's view: for them, an avowal is different from a proper belief because it lacks belief's connection to action, whereas Gendler claims imaginative pretense *does*

have such a connection. So the avowal view simply cannot explain the import of an action such as taking out a loan at great personal risk instead of cutting losses.

Whether or not imaginative pretense can account for the kinds of action performed on the basis of self-deception, we gather from examples such as those we just now discussed that in forging a model of self-deception one *must* account for the actions that make self-deception a serious and potentially hazardous issue for the people involved. We have seen that one can easily choose to act, as in a game of make-believe, with imagining as the adjusted cognitive background. The relevant representation, which might be brought to consciousness with the form 'I believe that I am pretending that I am William Wallace' (or whatever) has consequences such as deliberately moving one's arms about as if one is carrying a heavy sword, yelling "FREEDOM!", etc. On the other hand, if Van Leeuwen's argument is right, then it follows that one simply *cannot* really act on self-deceptive, imaginative pretense. This is so because *choosing* to act on self-deceptive pretense would constitute nothing short of a self-defeating project, in the fashion of the old 'dynamic paradox', since imaginative pretense can motivate action only insofar as it is backed by a metarepresentation of (i.e. belief about) the setting. But the only way a self-deceived agent *could* act on her self-deceptions would be by being completely oblivious of the practical setting she is in. 'I believe that I am pretending that my girlfriend is faithful' would never have the kind of consequences it would have in cases where the subject is truly self-deceived, such as behaving as if nothing is wrong most of the time, asking her to move in (especially since the subject's central belief, according to Gendler, would be of the form 'I believe my girlfriend is unfaithful'). The only way one can come to act on the basis of self-deception is to falsely or unwarrantedly *believe* that its content is true, however unjustified that might be.

1.5 Conclusion

The main problem in Gendler's general attempt at explaining self-deception seems to lie in a confusion of *process* and *product*. First, her reading of belief's essential truth-directedness seems to stem from confusing the rationality of belief formation (a process) and the rationality of belief *per se* (its product). While belief as a cognitive attitude may be said to essentially aim at the truth, the processes by which beliefs are formed are obviously fallible, being subject to a great variety of missteps. While this is a trivial point, I believe attention to this fact may suffice to make clear that what happens in cases of biased cognition (be it hot, cold, or both), including self-deception, is not a violation of the truth-directedness constraint that would warrant the abandonment of belief talk altogether in this context. Truth-directedness can be held to be a

feature of belief even in the context of biased cognition, where belief is formed and maintained in ways that divert it from its ideal, rational aim.

Second, imaginative pretense may perfectly well figure in the process of forming and maintaining a self-deceptive belief and thus may be given a role in an explanatory account of self-deception. Spinoza, in the third part of his *Ethics*, already observed that 'it can easily happen that one who exults at being esteemed is proud and imagines himself to be pleasing to all, when he is burdensome to all' (1677/1996), 86). F.A. Siegler, in one of the very first articles in the analytic literature on self-deception, mentions, in passing, that we should resist attributing contradictory beliefs to the self-deceiver, since we would not attribute them to 'a man who was pretending or who was taken in by his own pretenses' (1962, 470). But Spinoza and Siegler could, and probably were, speaking figuratively (or at least referred to 'pretense' in a vague way). On the other hand, Mike W. Martin acknowledged 'self-pretense' among the many patterns of evasion seen in self-deception (1979; 1986, 8). It is in Martin's way that I think we should incorporate Gendler's core insight, namely, that 'just as you can deceive another ... by performatively pretending that $\neg p$ rather than p, so too you can deceive yourself ... by imaginatively pretending that $\neg p$ rather than p' (2007, 240). What I have tried to object here is to taking imaginative pretense to be the *product* of self-deception instead of one of the ways through which people can deceive themselves. And I have tried to do that by showing that Gendler's appropriation of Velleman's theory of truth-directedness is misguided and that, in addition, Velleman's theory of the motivational role of non-belief cognitive attitudes (on which Gendler's explanation depends) should be abandoned.

Capítulo 2

The tenability of a dispositional account of delusional belief

2.1 Introducing delusion

Similarly to self-deceptive states, delusional states have been traditionally thought of as beliefs. Indeed, they are referred to as beliefs almost everywhere in the psychiatric literature, as exemplified by the latest revision of what is perhaps its most influential handbook, the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV-TR). Here is its definition of delusion:

> A false belief based on incorrect inference about external reality that is firmly sustained despite what almost everybody else believes and despite what constitutes incontrovertible and obvious proof or evidence to the contrary. (American Psychiatric Association 2000, 821)

As we will see, this is not a very careful, nor a very helpful definition. Perhaps it was not even intended as a definition, but only as a gloss. Still, at close inspection, almost everything about it seems debatable.[1] On the bright side, I think walking through these difficulties is a good way to be introduced to the concept of delusion.[2]

[1] This is not of itself a form of criticism directed at the team of psychiatrists responsible for the compilation of the monumental DSM. The conceptual laxness with which the psychiatric literature is sometimes crafted constitutes fertile ground for philosophical analysis. This is true of all well-established sciences, which have given philosophers a multitude of notions to think about, distill, and refine. And it is also true of the so-called 'soft' sciences, the philosophical exploration of which is more recent.

[2] In what follows I discuss only epistemological points to be made concerning the DSM definition, but these are not its only source of criticism. For some, it fails to capture the important feature of disrupted functioning. For example, McKay, Langdon and Coltheart state: 'Essentially we view delusion as connoting both a dearth of evidential justification

'A *false* belief'

Delusions are not always (unqualifiedly) false. Reputedly, when during the 1973 Sinai talks Henry Kissinger accused Golda Meir of being paranoid for hesitating to grant further concessions to the Arabs, she quipped that 'even paranoids have enemies' (Graham 2010a, 195). In the same spirit, nothing precludes the husband of an unfaithful wife from having *delusional* jealousy. That is, delusional beliefs don't *need* to be false: it is not the truth-value of the proposition(s) believed by delusional subjects that is epistemologically interesting to the characterization of delusions, but the fact that their belief is 'sustained despite what constitutes incontrovertible and obvious proof or evidence to the contrary.'

'based on *incorrect inference*'

Given the anomalous experience arguably undergone by many if not all delusional patients, it is worth pondering over whether the conclusions drawn by delusional subjects are incorrect. Capgras delusion is a case in point—it usually involves the misidentification of the faces of loved ones, such as spouses or children—and it has been accounted for by a disruption of the system that subserves our normal affective responses to people's faces (Ellis et al. 1997). Hence there is room to argue that the odd conclusions of Capgras patients are not to be judged irrational, since that would call for normal affective responses. (It must be noted, however, that affective responses, be they normal or pathological, are not *premises* from which to reason, but at best experiential *input* prompting one to adopt premises from which to reason.) The implicit question here is whether delusions can be acquired rationally. Although Maher (1988) makes out delusions to be largely rational responses to anomalous experiences, authors like Langdon and Coltheart (2000) point out that no reasoning style can be as impervious to evidential override, and so speculate that a (neurological) *deficit* must be posited to account for this unusual resistance to counterevidence, whereas other authors suggest that a biased formed of reasoning (an *attributional style*) can be coupled with the experience to account for delusional beliefs in a way that prevents us from ascribing full rationality to delusional subjects (Stone and Young 1997). Mixed models have also been suggested (Davies and Coltheart 2000). Apart from the correctness of the inferences allegedly made by delusional subjects, it is also not self-evident that delusions are a product of inferential reasoning. Philip Gerrans objects:

> The inferential conception of delusion treats the delusional subject as a scientist in the grip of an intractable confirmation bias. She re-

and an element of day-to-day dysfunction. A person is deluded when he or she has come to hold a particular belief with a degree of firmness that is utterly unwarranted by the evidence at hand and that jeopardises their day-to-day functioning' (2009, 172).

> calls and attends selectively to evidence consistent with her biased hypothesis with the result that the delusions become ever more firmly woven into her Quinean web of beliefs... I propose instead that processes of selective attention and recall exert their effects, not on a process of hypothesis confirmation but of autobiographical narrative. Someone with a delusion is not a mad scientist but an unreliable narrator. (Gerrans 2009, 152)

'about *external* reality'

Perhaps some delusions are distinctly about external reality. Paranoid delusions sometimes involve the belief that one is being persecuted by government agents. De Clérambeault's syndrome (also known as erotomania) usually involves the belief that one is loved by a famous and distant person with whom one has never had any kind of contact. And intermetamorphosis involves the belief that familiar people in one's environment are constantly exchanging identities with each other while retaining the same appearance. Yet many delusions which may appear at first sight to be based solely on external factors do in fact revolve around the subject and, more importantly, so do the theoretical explanations of these delusions. Furthermore, the claim of externality is in direct contrast with the existence of delusions that do not seem to be about 'external reality' at all. Somatoparaphrenia, which involves the denial of ownership of one or more of one's limbs or sometimes an entire side of one's body, illustrates this well. The most blatant example, however, would probably be Cotard's syndrome, which involves delusional beliefs that paradigmatically concern the subject and the subject alone ('*I* am not alive'). On the other hand, if 'external reality' is meant to be synonymous to 'reality', then this portion of the definition is not ill-conceived, but only rather trivial.

'*firmly* sustained'

Not all delusions are firmly sustained, and the conviction of delusional subjects may fluctuate. Davies and colleagues observe that at least some delusional patients 'show considerable appreciation of the implausibility of their delusional beliefs' (2001, 149). Furthermore, clinicians have reported that some of their delusional patients actually entertain the possibility that they are *mistaken* in their beliefs (Bentall 2003, 324). Andrew Young provides a lively case:

> Capgras delusion patients can be ... able to appreciate that they are making an extraordinary claim. If you ask 'what would you think if I told you my wife had been replaced by an impostor,' you will often get answers to the effect that it would be unbelievable, absurd, an indication that you had gone mad. (1998, 37)

'despite what *almost everybody else* believes'

The fact that the DSM, in defining delusion, situates the delusional subject in an epistemic community (so to speak) is problematic. What is the relevance of the shared opinion of one's community in characterizing someone's mental state as delusional? Consider for a moment the following scenario: a delusional man believes that divine forces are preparing him for a sexual union with God by changing him into a woman.[3] Suppose he subsequently succeeds in convincing most of his peers that, however far-fetched his claims, they are true. His peers' new belief does not in the least make him less delusional—it only testifies to his outstanding rhetorical abilities. Of course, for the greater plausibility of this thought experiment I could certainly leave room for a few skeptics. It is nevertheless perfectly conceivable that the majority of an epistemic community be convinced of the reality of a fantastic claim. Although the assumption that probably underlies this portion of the DSM definition is comprehensible—that 'what almost everybody else believes' is more likely to be true than whatever idiosyncratic beliefs a particular individual may have—it is a poor standard from which to judge the delusional character of a given belief (in Galileo's days, *almost everybody* else believed that the Sun revolved around the Earth).Thus, I suggest that what other people believe should not be an integral part of any definition of delusion.

'despite what constitutes *incontrovertible and obvious proof or evidence to the contrary*'

As far as I know, this is indeed a matter of consensus. After all, it seems that all of the examples of delusion that we have conjured up so far are certain to meet a large body of counterevidence. But take the case of mirrored-self misidentification—the belief that one's reflection in the mirror is that of someone else. It sometimes is accompanied by the belief that whoever that person in the mirror is, he or she is following the subject around—a fact that helps us understand why sometimes the 'mirror sign' is at the onset of progressive dementing illness (Breen, Caine and Coltheart 2001). Now, are these patients in possession of 'incontrovertible and obvious proof or evidence' that, although they fail to identify the face in the mirror, it is nevertheless theirs? (The same could be asked about all other delusional misidentification syndromes.)

I think it is safe to say that they are, and that this particular bit of the DSM gloss is absolutely necessary for the characterization of delusional states. Just as not all hallucinatory symptoms lead to delusion, an otherwise normal subject (say you and me) presented with the anomalous experience of not recognizing oneself in the mirror, would not arrive at the belief that, say—although the

[3] This was one of the many florid delusions of German Supreme Court Judge Daniel Schreber (Schreber 1903/2000).

mirrored person is waving just like I am, wearing the same clothes, sporting the same hairstyle, etc.—the person in the mirror is not me. In addition to these overriding facts (which point to the great plausibility that there is something wrong with *me*), the testimony of each and everyone of one's epistemic peers would also weigh in heavily in the reasoning of a person whose thoughts did not mark the presence of some deficit and/or bias. This is the most puzzling characteristic of delusions, and the one which ties them to the notion of self-deception so that, in common parlance, 'to deceive oneself' and 'to delude oneself' are largely interchangeable expressions, meaning not only 'to have a false or unwarranted belief,' but 'to have it in the face of counterevidence.' Davidson succintly described the general problem—the underlying 'paradox of irrationality'— in accounting for these kinds of phenomena:

> If we explain it too well, we turn it into a concealed form of rationality; while if we assign incoherence too glibly, we merely compromise our ability to diagnose irrationality by withdrawing the background of rationality needed to justify any diagnosis at all. (1982, 303)

From what has been said it is not hard to predict that it will be for its insistence on ascribing full-blown belief to irrational subjects that the standard characterization of delusion will meet its most vexing objections. I now turn to a brief review of the main difficulties regarding the ascription of belief to of delusional subjects.

2.2 What's wrong with delusional belief

The implausibility of ascribing full-fledged belief to delusional subjects has been hinted at at least since the 1910s, when both Karl Jaspers' *General Psychopathology* and Eugen Bleuler's *Textbook of Psychiatry* were published. The set of objections against the traditional view forms an unignorable obstacle for doxastic accounts.[4]

Lack of content

The first objection—originally raised by Jaspers (1913/1963) and elaborated recently by German Berrios (1991) and Louis Sass (1994)—denies that delusions are contentful states. One may call this the *expressivist* (Gerrans 2001) or *non-assertoric* (Young 1999) account. This view is motivated by the fact that most (if not all) delusions appear obviously false or incoherent. The aforementioned Cotard delusion is a prized example of proponents of this view. Berrios,

[4] In listing these objections I largely follow the excellent survey provided in Bayne and Pacherie (2005). I also benefited from the discussion in Stephens and Graham (2004), Egan (2009) and Bortolotti (2010).

for example, states that when a patient who utters a verbal formula such as 'I am dead' or 'My internal organs have been removed' is questioned as to the real meaning of these assertions, she will not be able to coherently discuss them or their implications. 'Properly described,' says Berrios, 'delusions are *empty speech-acts* that disguise themselves as beliefs' (1996, 126, my emphasis). 'Their so-called content refers neither to world, nor self'. 'Delusions are so unlike normal beliefs that it must be asked why we persist in calling them beliefs at all' (1996, 114-5).

A wide variety of other cases besides Cotard's can be summoned in favor of such a view. Tim Bayne and Elisabeth Pacherie (2005) cite an intermetamorphosis patient who claimed that his mother changed into another person every time she put her glasses on (De Pauw and Szulecka 1988); another that had the delusion that there was a nuclear power station inside his body (David 1990); and a third that had the delusion of being both in Boston and in Paris at once (Weinstein and Kahn 1955).

Self-defeating content

One may not want to deny that delusional states possess content, and still object that it is difficult to see *how* the delusional patient themselves could believe such content. Again, Cotard patients are a fitting example. José Luis Bermúdez voices this concern in stating that there is 'something content-irrational about the belief ... that one is dead—because, to put it mildly, the belief is *pragmatically self-defeating*' (2001, my emphasis). Not only is it unclear that a self-defeating assertion such as 'I am dead' could be coherently expressed,[5] the question is open whether there can be self-defeating *beliefs* to begin with (as opposed to mere verbal utterances).

Lack of evidence

A distinctly rationalist objection consists in pointing out that delusional subjects appear, as opposed to self-deceived ones, to lack reasons or evidence for their delusional state. However faulty the reasons or flimsy (and biased) the evidence one may have to support some self-deceptive belief, there will be nevertheless *some* kind of support for such a belief. In contrast with this, John Campbell cites the well-known case of 'a patient who looked at a row of empty marble tables in a café and became convinced that the world was coming to an end' (2001, 95). Notwithstanding the DSM definition of delusions (that they are held 'despite what constitutes incontrovertible and obvious proof or

[5] Except, of course, in such contexts as that of the opening words of a will ('Now that I am dead...'), or of the hero's epitaph in Ezra Pound's *Mauberley* ('I was. And I no more exist; Here drifted. An hedonist.').

evidence to the contrary'), Campbell points out that it is difficult to understand (to put it midly) how an experience of marble tables could verify the proposition 'The world is ending'.[6] On the other hand, there is at any time a considerable body of evidence *against* the truth of the delusional content, to which the delusional subject seems utterly impervious. Furthermore, there are delusional patients that even recognize that they do not have evidence for their claims. A case in point is Young and Leafhead's Cotard patient, JK:

> We wanted to know whether the fact that JK had thoughts and feelings (however abnormal) struck her as being inconsistent with her belief that she was dead. We therefore asked her, during the period when she claimed to be dead, whether she could feel her heart beat, whether she could feel hot or cold. ... She said she could. We suggested that such feelings surely represented evidence that she was not dead, but alive. JK said that since she had such feelings even though she was dead, they clearly did not represent evidence she was alive. (Young and Leafhead 1996, 157-8)

This is a startling case of cognitive dissonance and resisting evidence, and suggests something along the lines of Andy Egan's observation that 'if we think that a certain responsiveness to evidence is essential to belief, then, in many cases, we'll be reluctant to say that delusional subjects genuinely believe the content of their delusions' (2009, 266). In other words, if there is a constitutive relationship between belief and evidence (even in the case of irrational belief and improper evidence), then it seems that delusional states do not warrant the ascription of delusional beliefs. This paves the way to what is perhaps the most forceful set of objetions to the doxastic conception: those which point to *bad integration* (Bortolotti 2010).

Theoretical reasoning

Many authors, like Velleman (2000), take belief to be somehow aimed at the truth. This may constitute an objection to the doxastic conception of delusion in a spirit similar to that of Gendler's (2007) objection to the doxastic conception of self-deception. It points to one of the ways in which delusional states present a degree of *circumscription* (Young 1999, 581) that speaks against their being properly taken as beliefs. Egan calls this property of delusional states

[6] Although not impossible. Paulo Faria (personal correspondence) suggests the following scenario: 'Suppose the Almighty (under cover, perhaps, of a burning bush, as in Exodus 3, 2-21) had told his prophet, call him Moses II, that the end was approaching, and that, as a warning signal to His chosen children, he would have Moses II run against a row of empty marble tables when entering a café. (We may suppose it would then be Moses II's duty to warn his brethren that the end had come.)'.

inferential circumscription (2009, 266). As Bayne and Pacherie neatly put it, proponents of such a view of belief point out that

> A subject will normally accept the obvious logical implications of her beliefs—at least when these are pointed out to her. And when she realizes that some of her beliefs are inconsistent, she will normally engage in a process of revision to restore consistency. In contrast, deluded patients often fail to draw the obvious logical consequences of their delusions and show little interest in resolving apparent contradictions between their delusion and the rest of their beliefs. (2005, 164)

This is the precisely the vein in which Currie and Ravenscroft affirm that

> If someone says that he has discovered a kind of belief that is peculiar in that there is no obligation to resolve or even to be concerned about inconsistencies between these beliefs and beliefs of any other kind, then the correct response to him is to say that he is talking about something other than belief. (2002, 176)

However, the majority of patients with the Capgras delusion, for example, do not draw the consequences the content of their delusion would usually mandate: their *worldview* does not seem to change at all as a consequence of supposedly adopting the belief that their spouses have been abducted and that the person they see in front of them is an impostor (Davies and Coltheart 2002). Whatever this state is, therefore, it seems that it is severely encapsulated, failing to be integrated with the subject's web of belief. But beliefs are the mainstay of theoretical and practical reasoning and, while one may ascribe false belief to subjects for any number of reasons, a state that fails to have the appropriate connections to the subject's other mental states may not be properly described as a belief.[7] As exemplified by Currie and collaborators, this view is especially espoused by authors who (tacitly or explicitly) endorse a consistency constraint on belief-ascription.

Indeed, authors such as Quine and Ullian (1970), as well as Fodor (1983), have argued that one of the attributes of a belief *qua* belief is its property of being inseparably connected with other beliefs of potentially widely diverse contents. Quine's answer as to why beliefs should be webbed or interconnected with other beliefs in a way that precludes severe encapsulation rests on the conditions of epistemic assessment of beliefs—for instance, whether I am warranted in believing that an acquaintance of mine lives in Chicago may

[7] I say 'appropriate' rather than 'necessary' because circumscribed (insulated) "beliefs" will usually stand in a number of (nonlogical) connections to the subject's other mental states: that of being simultaneously held to begin with, and then that of causing or being caused by other mental states.

depend on whether I believe that Chicago is a city and believe that cities are bigger than towns, etc. And for Quine, the conditions of epistemic assessment of beliefs are part of their functional role; beliefs are states or attitudes that are constituents in (what Fodor calls) the central processing that takes place in the mind.[8] Therefore, like-minded theorists will deny that delusional subjects are in the hold of belief.

Practical reasoning

Belief has profound connections to action, and many delusional subjects fail to act in ways expected of agents who really believed the content of their delusions. As Currie puts it, delusion 'exerts a powerful psychological force, absorbing inner mental resources, but it fails to engage behaviour in the way that genuine belief would' (2000, 175).[9] This seems likely due to the inferential circuscription noted above. Egan calls this characteristic of delusional patients *behavioral circuscription* (2009, 266). It was noted by Bleuler, who stated that his delusional patients 'rarely follow up the logic to act accordingly, as, for instance, to bark like a dog when they profess to be a dog. Although they may refuse to admit the truth, they behave as if the expression is only to be taken symbolically' (1916/1924). In the same manner, Capgras patients who (for all we can see) sincerely affirm 'This is not my wife' or 'My mother has been replaced by an impostor' do not as a consequence of this go looking for their missing loved ones, nor do they call the police to report the breaking and entering perpetrated by the person they claim to be an impostor.

Lack of appropriate affect

Finally, delusional patients often fail to exhibit the affective (i.e. emotional) responses one would expect of a person who believes the content of her assertions (Sass 1994, 23–24). We may call this *affective circumscription*, since what is observed is a failure of integration between the subjects' delusional state and their emotional lives. Capgras patients are more often than not unmoved by the fate of their relatives whom, according to the doxastic interpretation of

[8] Fodor is committed to the analogy between scientific confirmation and psychological fixation of belief, and states that 'the central processes which mediate the fixation of belief are typically processes of rational nondemonstrative inference and that, since processes of rational nondemonstrative inference are Quineian [i.e. the degree of confirmation assigned to any given hypothesis is sensitive to properties of the entire belief system] and isotropic [i.e. the facts relevant to the confirmation of a scientific hypothesis may be drawn from anywhere in the field of previously established truths], so too are central processes. In particular, the theory of such processes must be consonant with the principle that the level of acceptance of any belief is sensitive to the level of acceptance of any other and to global properties of the field of beliefs taken collectively' (1983, 110). [9] See also Sass (1994, 21) and Young (1999, 581).

this delusion, they believe to have been abducted. Why don't they exhibit the affective responses which the relevant beliefs would lead us to expect?

These last four objections are different aspects of a general *bad integration* objection. Thus Bortolotti observes that 'although it is possible for a belief system to have some internal tension, most philosophers resist the thought that subjects capable of having beliefs can have dissonant attitudes simultaneously activated and operative at the forefront of their minds' (2010, 62). Delusions lack the holistic character expected of beliefs and do not respect the notion of a coherent belief system whose adjustments to one belief implies adjustments to many others (Young 2000, 49). Belief-ascription in the context of delusion, then, is only admissible after explaining away these disparities between the roles that delusional states play in the overall cognitive economy of delusional patients and those roles we expect beliefs to play (following either folk-psychological intuitions or fully articulated theories of belief).

2.3 A dispositional approach to delusions

Having briefly introduced the concept of delusion and the problems faced by doxastic explanations, I now turn to the main purpose of this paper, which is to examine the tenability of a dispositional account of delusional belief. Instead of giving up doxasticism altogether in favor of some alternative explanation, such as Currie and colleagues' (2001, 2002) so-called *metacognitive* account, which relies on imagination rather than belief, the strategy I will rehearse consists in presenting a non-standard account of belief, and trying to show that deluded subjects meet the criteria it proposes for being in a belief-state. Our starting point will be an overview of the dispositionalist theory of belief proposed by Eric Schwitzgebel (2001, 2002), in order then to explore its subsequent appropriation in the literature on delusions by Bayne and Pacherie (2005).[10]

2.3.1 In-between believing

H.H. Price, in his famous series of lectures on belief, discussed the not uncommon phenomenon wherein a person may systematically feel himself to be and act as if he were fully committed to p in one set of circumstances, while systematically feeling and acting as if the opposite were true in others. He called this 'half-belief' (1960/1969, 302-14). More recently, Schwitzgebel (2001) recognized that there are countless cases in which a simple yes or no answer to the question 'Does S believe that p?' doesn't seem to be available, and that

[10] In their paper, Bayne and Pacherie indirectly defend the doxastic account by attacking the metacognitive account. My focus, however, will be solely on their direct defense of the doxastic account (especially in section 3 of their paper).

they can have a wide variety of causes. From these cases, Schwitzgebel draws the conclusion that

> For any proposition p, it may sometimes occur that a person is not quite accurately describable as believing that p, nor quite accurately describable as failing to believe that p. Such a person, I will say, is in an 'in-between state of belief' (2001, 76).

By way of illustration, he offers three examples stemming from three different causes, which are neither meant nor thought to be exhaustive. The first is *gradual forgetting.* It concerns the ubiquitous case in which someone forgets, say, an old colleague's last name. Years ago, you knew your colleague's full name. Now, you can only remember his first name (and, perhaps, the first letter of his last name). Years from now, you probably won't remember his name at all. So the belief that your colleague's name was Konstantin Guericke was fully present when you were in college, and will be fully absent when you are eighty years old. The question then is, what is the state you're in right now? Schwitzgebel asks: 'is it plausible to think that in the years between there was a discrete moment before which I absolutely had this belief and after which I absolutely did not? At some point during the course of forgetting, I must be *between believing and failing to believe* that his last name is 'Guericke' (or whatever)' (2001, 77, my emphasis). Arguably, we spend most of our lives in such an in-between state.

His second example is derived from our *failure to think things through.* Think of a school teacher who mentions prime numbers in her lessons, correctly listing the lower primes 2, 3, 5, 7, 11 etc. Now, when she is asked about or decides to offer the definition of 'prime number', she typically says that a prime number is any positive integer that can be divided evenly only by 1 and itself. This definition is not correct, however, since the number 1 is a positive integer evenly divisible only by 1 and itself, but it is not a prime number. On the other hand, if you asked the school teacher if 1 is a prime she would promptly answer that it isn't. So now the question is, does she believe that all positive integers which are evenly divisible only by themselves and 1 are prime? We have reasons to answer in the affirmative, for instance, she would never list 1 as a prime number. But we also have reasons to answer in the negative, for instance, the occasions on which she would be disposed to offer a correct definition of primes are few. For this reason, Schwitzgebel claims 'the most careful and accurate description of her would neither simply ascribe the belief to her nor simply deny it of her' (2001, 77).

Finally, there is *variability with context and mood.* Here, Schwitzgebel evokes a familiar example in the same vein as Price's famous case of the half-believing theist. Price suggests the case of someone who on Sundays bears all the subjective and objective marks of someone who believes that there is a

God, but who on weekdays bears none of them. Schwitzgebel, on the other hand, suggests the case of someone who, in certain moods and in certain contexts, bears all the subjective and objective marks, and who, in other moods and contexts, doesn't. (The latter spectrum may include circumstances from those of weakened confidence, as when someone thinks of God as 'a beautiful metaphor', to those where confidence is removed completely from recognition or memory.) Though he may be a regular Sunday churchgoer, he does not feel the urge to defend himself or his religion when, for example, his atheistic friends mock religious belief. In fact, at such moments (especially on weekdays), he may even find himself mildly convinced of the incongruousness of theistic dogma. How can we decide, then, whether he believes that God exists? Once again, Schwitzgebel makes the point that a simple yes or no answer would be misleading.

> One might say that his beliefs change from occasion to occasion—that as he is grousing about the church social, he does not believe that God exists; as he is rejoicing in the magnificence of spring, he does believe—but most of the time he is doing neither: he is eating breakfast or mowing or writing code and not giving the matter any thought. At such moments he may be simultaneously disposed to marvel at the wonder of creation if a robin were to fly past and to embrace atheism if Madge were unexpectedly to drop by. (2001, 78)

The widespread presence of problematic circumstances for belief-ascription such as these encourages the development of an account of belief that allows us to talk intelligibly about such in-between states—that allows us to say more than just that the subject 'sort of' believes something.

2.3.2 A phenomenal, dispositional account of belief

Given the notion that there is a continuum ranging from complete absence to complete presence of any given belief, a probabilistic treatment might be thought to manage cases of in-between believing. According to such an account, a person's beliefs would be characterized by a degree of confidence ranging from 0 (i.e. absolute confidence in the falsity of p) to 1 (i.e. absolute confidence in the truth of p), with 0.5 in between—perhaps representing suspension of judgment or a state of skeptical doubt.[11] Such an approach may

[11] 'Probabilistic' is the way I have chosen to put it. Schwitzgebel (2001, 2002) chooses the word 'Bayesian' but, as has been pointed out to me, there is nothing specifically bayesian about the view he describes. The notion that belief comes in different degrees of confidence is part of every probabilistic account of belief (Ramsey–DeFinetti's, for instance). However, since there is no mention of *conditional* probabilities (conditional, that is, upon prior beliefs), nothing warrants Schwitzgebel's choice of label. See Rowbottom (2007).

be thought to account for at least some of the cases because we could assign our half-believing theist, for example, with a degree of confidence of 0.7 or 0.8. However, this would consist in a gross oversimplification of the kind of uncertainty or wavering present in the cases discussed. The school teacher and the half-believing theist cannot be properly described as simply fluctuating between different degrees of confidence, since they are, 'at a single time, disposed quite confidently to assert one thing in one sort of situation and to assert its opposite in another' (Schwitzgebel 2001, 79). Nor can the process of gradually forgeting someone's last name be properly translated into a slow decline in one's confidence in the truth of some proposition. A purely probabilistic approach fails to capture the vast array of detail present in these cases.

Furthermore, it would seem that traditional *representational* accounts of belief cannot provide a way of successfully dealing with in-between belief states either. Indeed, to suggest that someone is in an in-between representational state appears even more unnatural than the probabilistic strategy would have it. Most talk of belief as representation makes out belief to be a *categorical* state—having a belief that p is something like having the sentence p inscribed in one's 'belief box' in the language of thought, according to one popular account. The metaphor must be pushed, though, if representationalists wish to embrace the very plausible presence of halfways states. Schwitzgebel points out that for that, however, they risk making a caricature of their own account by incorporating, say, explanations of gradual forgetting in terms of a sentence slowly 'losing its color', etc. To avoid the far-fetched claim that sentences either are or aren't inscribed in the belief box, then, Schwitzgebel claims that representationalists are left with the burden of coming up with helpful ways of describing in-between cases in representational terms.[12]

Schwitzgebel opts for pursuing a more flexible explanation of the nature of belief and belief-ascription by appeal to a revision of Gilbert Ryle's dispositionalism.

> I call the account of belief I am about to offer a *phenomenal, dispositional* account. I call it a *dispositional* account because it treats believing as nothing more or less than being disposed to do and experience certain kinds of things. I call it a *phenomenal* account because, unlike dispositional accounts as typically conceived, it gives a central role to conscious experience, or 'phenomenology.' (2002, 250)

Ryle argued that to believe something is simply to be disposed to do and feel

[12] It has been pointed out to me that one might not necessarily have to resort to such metaphors as the one from Schwitzgebel's sketchy example. For instance, as a sentence encodes information, halfway states may plausibly enough be accounted for in terms of loss of information, along the lines, say, of the mathematical model of communication proposed by Shannon and Weaver (1963).

certain things in appropriate situations.[13] To use his own example, to believe that the ice you're skating on is dangerously thin is, in his words,

> to be unhesitant in telling oneself and others that it is thin, in acquiescing in other people's assertions to that effect, in objecting to statements to the contrary, in drawing consequences from the original proposition, and so forth. But it is also to be prone to skate warily, to shudder, to dwell in imagination on possible disasters and to warn other skaters. It is a propensity not only to make certain theoretical moves but also to make certain executive and imaginative moves as well as to have certain feelings. (1949, 134-5)

A person who has the dispositions described in Ryle's example matches what Schwitzgebel calls a *dispositional stereotype*. By a stereotype, he means a cluster of properties we are apt to associate with something—be it an object, a class, or a property. An example he adapts from Putnam (1975) is that of the stereotype of a tiger, whose properties include being striped and having four legs, among others.[14] This doesn't mean, of course, that a three-legged tiger without stripes is not a tiger. It only means that such a tiger wouldn't be a stereotypical one. Furthermore, the accuracy of stereotypes varies greatly in degree, so that the more or less objects instantiate their stereotypical properties, the more or less accurate the stereotype will be.

Schwitzgebel characterizes dispositions by means of conditional statements of the form 'If condition C holds, then object O will (or is likely to) enter (or remain in) state S' (2002, 250).[15] O's entering S is the *manifestation* of a disposition, whereas C is the *condition of manifestation*, and the event of C's obtaining is the *trigger*. Therefore, O will have the relevant disposition if and only if the corresponding conditional statement is true. Thus we may speak of

[13] Note that I don't mean to say that Ryle was behaviorist views, even if 'softer' ones. See Tanney (2009), especially section 8.

[14] It must be noted, however, that on Putnam's account these are not just properties 'we are apt to associate' with a tiger—they are criteria of tigerhood. Putnam's concept of a stereotype, in other words, is an epistemological one: a stereotype is the set of properties we use to identify something as an F, and it is to be sharply distinguished both from the (psychological) notion of 'what we are apt to associate' with an F and from the (metaphysical) notion of what makes an F an F. So the stereotype of water is: 'insipid, odorless, etc. liquid which quenches thirst, is found in lakes, rivers etc.' Yet that is not what water is: water is a chemical substance with the formula H_2O. See also Lecture III in Kripke (1980).

[15] Although grammatically in the indicative mood, I suspect a subjunctive conditional is what is meant here (as betrayed in the surface grammar by the use of 'will', implying necessity), namely: 'If condition C were to hold, the object O would (be likely to) enter (or remain) in state S'. The latter sentence states explicitly that there is a nomic connection between being in condition C and entering (or remaining) in state S, something which no strictly indicative conditional would have the force to express ('If P then Q' iff 'Not-P and/or Q': no law to be found there). For a classic discussion, see the first chapter in Goodman (1983).

dogs having the disposition to wag their tails when excited because when they are excited (the trigger), they wag their tails (the manifestation).[16]

A dispositional stereotype is simply a stereotype whose elements are dispositional properties. Many familiar stereotypes are dispositional, such as personality traits. For example, being impulsive is (something like) being disposed to act without thinking things through; being sympathetic is (something like) being disposed to easily putting oneself in someone else's position; etc. Just like having a personality trait is matching a stereotype, Schwitzgebel claims, so too is having a belief. As a consequence, the list of dispositions associated with a given belief is as indefinite as that of having a particular personality trait, and won't be linked to it explicitly by a conscious effort.[17] The most fruitful way of thinking about dispositional stereotypes is, rather, as consisting of clusters of dispositional properties (which we associate with particular stereotypes).

As regards the different kinds of dispositions belonging specifically to the stereotype of belief (for instance, the belief that there is beer in the fridge), Schwitzgebel identifies three main categories:

> The most obvious, perhaps, are *behavioral* dispositions, the manifestations of which are verbal and nonverbal behavior, such as, in the present case, the disposition to say that there is beer in the fridge (in appropriate circumstances) and the disposition to go to the fridge (if one wants a beer). Equally important, though rarely invoked in dispositional accounts of any sort, are what may be called *phenomenal* dispositions, dispositions to have certain sorts of conscious experiences. The disposition to say silently to oneself, 'there's beer in my fridge,' and the disposition to feel surprise should one open the fridge and find no beer are phenomenal dispositions stereotypical of the belief that there is beer in the fridge. Finally, there are dispositions to enter mental states that are not wholly characterizable phenomenally, such as dispositions to draw conclusions entailed by the belief in question or to acquire new desires or habits consonant with the belief. Call these *cognitive dispositions*. (2002, 252)

In Schwitzgebel's dispositionalism, a person who possesses *all* the dispositions

[16] Which, please note, does not mean that every dog is such that it is not excited and/or its tail wags—which would be true of a dog which is excited without wagging its tail—all indicative conditionals whose antecedent is true being alike trivially true. See Goodman (loc. cit.) for further detail.

[17] Which is not a shortcoming of the effort, but a logical property of subjunctive conditionals (again, as discussed in Goodman). A subjunctive conditional is true all other things being equal, that is, in possible worlds closest to the actual world. (I am indebted to Paulo Faria for comments and suggestions incorporated in the last four footnotes).

in the stereotype for belief that p will *always* accurately be described as believing that p. On the other hand, a person who possesses *none* of those dispositions will *never* accurately be so described. What is especially relevant to the present investigation, of course, are the cases in between those extremes—cases which, in all probability, account for most of our beliefs. As we have seen, those cases include gradual learning and forgetting and ignorance of related details, but also self-deception and, perhaps, even some cases of delusion. That is, cases where some but not all of the dispositions in a stereotype are present, and which help illustrate that having a disposition is not something which can always be ascribed in black-and-white terms. The core idea of this view with respect to belief, then, is that

> To believe that p ... is nothing more than to match to an appropriate degree and in appropriate respects the dispositional stereotype for believing that p. What respects and degrees of match are to count as 'appropriate' will vary contextually and so must be left as a matter of judgment. (2002, 253)

The ability of such a view of belief to handle the gray area of ascription is made clear by the postulate that no single disposition is either necessary nor sufficient for the possession of any belief (since dispositionalism links belief-ascription to clusters of dispositions). Schwitzgebel avoids the inflexibility of traditional accounts by admitting vagueness and context-dependency as integral to belief-ascription. As opposed to features that would entail undermining the value of ascription, they are supposed to provide the margin of safety we as belief-ascribers *need*. Finally, the problem Schwitzgebel intends to deal with is fundamentally the same we face when trying to ascribe belief to delusional patients, namely, that there are particular subjects of whom there is reason to think that they believe p but also reason to think that they do not believe p (or even that they believe $\neg p$).

2.3.3 Bayne and Pacherie's appropriation

The main goal of Bayne and Pacherie's article 'In Defence of the Doxastic Conception of Delusions' (2005) is to restore our ability—in the face of the reviewed objections and of alternative accounts—to make belief-ascriptions when talking about deluded subjects. As a motivation, they cite matters of practical importance such as the implications that theoretical speculation has for the treatment of delusions. They refer to cognitive behavioural therapy (CBT), an important form of therapy for delusions (Dickerson 2000), one of the essential components of which involves questioning the consistency and plausibility of the patient's delusions (Chadwick, Brichwood and Trower 1996). 'This form of therapy seems to accord with the doxastic account, in that the

therapist treats the delusional patient as a believer of p, and he or she gently invites the patient to question whether p is the thing that ought to be believed' (2005, 185).

Granted the important point that a sound methodology for the development of philosophical theories of the human mind should carefully attend to its compatibility with the relevant empirical data, the focus of my interest in their defense of doxasticism lies especially in their sketch of a theory of belief that, according to them, can elude the usual objections. In order to establish that at least some delusions qualify as beliefs they turn to Christopher Cherniak (1986) and Schwitzgebel (2002) for support. From Cherniak's account of 'minimal rationality' they derive the claim that the link between rationality and belief-ascription is much looser than classical theories of belief-ascription generally allow. Interpretationism—the view that we can gain an understanding of the nature of the mental by reflecting on the nature of interpretation, the process of ascribing propositional attitudes to an individual on the basis of what she says and does—is one such theory. It endorses a general rationality constraint that has been widely supported in the philosophical literature:

> When we are not [rational], the cases defy description in ordinary terms of belief and desire. (Dennett 1987, 87)

> If we are intelligibly to attribute attitudes and beliefs, or usefully to describe motions as behavior, then we are committed to finding, in the pattern of behavior, belief, and desire, a large degree of rationality and consistency. (Davidson 1974, 50)

If we were to accept this, it would follow that the difficulty we face in trying to explain and predict irrational behavior in intentional terms stems from the fact that irrational behavior *does not* support the ascription of intentional states with determinate content. Rationality would be emphasized to the exclusion of almost all other considerations.[18] However, upon closer inspection, the use of the folk-psychological concept of 'belief' (to focus on the intentional state we are presently concerned with) indicates that its extension is not homogeneous at all. As we have already seen, there are lots of kinds of different mental states we are prepared to call 'belief'. Hence, Bayne and Pacherie observe, even if they are right in that considerations of rationality play an important role in belief-ascription, theories like interpretationism obscure the heterogeneity of the set of states apt to be considered beliefs. As Cherniak urged, rationality

[18] It must be noted, however, that the strength of Davidson's and Dennett's constraint isn't the same. While Davidson leaves room for further inquiry by postulating the need for 'a large degree of rationality', Dennett's constraint seems impervious to qualification. Though a very interesting debate in itself, I must leave the tenability of rationality constraints on belief-ascription aside for reasons of space and scope. But see Bortolotti (2004, 2005).

constraints on belief-ascription should not be derived from a model of *ideal* rationality. 'Given our finitary predicament—the computational, memory, and time limitations we are subject to—it is actually irrational for us to aspire to ideal rationality' (Bayne and Pacherie 2005, 180). What this entails for the present discussion is that the cognitive attitudes of delusional patients are *continuous* with our ordinary beliefs—that is, there is no categorical difference between the abnormal mental states observed in delusions and the normal states that constitute our everyday cognitive economy.[19] If that is right, it constitutes an important first step toward a vindication of doxasticism about delusions.

More important to the present investigation, on the other hand, is Bayne and Pacherie's appeal to Schwitzgebel's dispositional account of belief in order to explain away deluded subjects' failure to manifest their beliefs in normal ways. From Schwitzgebel they derive the claim that beliefs are context-dependent in a number of ways:

> First, which dispositions are actualized is a function of several factors: (1) the way the long-term memory of the individual is structured, something that depends in turn both on the cognitive organization of the species and on the personal history of the individual, (2) the current external context, and (3) the current motivational and affective set of the individual. Second, belief-ascription is also context-dependent. According to Schwitzgebel, we have dispositional stereotypes for beliefs, specific clusters of behavioral, cognitive, and phenomenal dispositions we associate with given beliefs and expect to be manifested in standard situations. We attribute to a subject full belief that p if he conforms to the associated stereotype in standard situations and *if his deviations from the stereotype are readily explainable or excusable* by appeal to some non-standard feature of the situation in which they occur. When a deviation from the stereotype cannot be excused or explained in this way, whether or not the attributor ascribes the belief will depend on the context of the belief ascription and what her interests are. (2005, 181, my emphasis)

Dispositionalism does look like a very promising way to ensure that young children, the forgetful, the negligent, the weak-willed, the self-deceived and the deluded can be ascribed the beliefs they appear to have, despite the lapses in thinking and behavior that sometimes pose a challenge to these ascriptions. It makes it easy to secure permission to use belief-ascriptive language in our descriptions of these individuals. That is so because, for the dispositionalist, the question whether a subject should be ascribed the belief is not just a matter of

[19] See Bortolotti (2010) for an extended argument for this claim.

whether she manifests enough of the dispositions in the relevant cluster but also of whether her *not* manifesting some of these dispositions can be satisfactorily excused or explained by reference to non-standard aspects of her situation. Therefore, if deluded subjects' failure to manifest their beliefs in normal ways is excusable or explainable, their deviation from the dispositional stereotype associated with the relevant beliefs ceases to be a hindrance to belief-ascription. Nevertheless, Bayne and Pacherie's claim that the difficulties for the doxastic account can be resolved if belief-ascriptions are context-dependent is by no means beyond dispute. I now turn to a critical assessment of their attempt to make use of Schwitzgebel's dispositionalism in order to rescue doxasticism from the previously discussed objections.

2.4 Can dispositionalism save doxasticism?

Maura Tumulty (forthcoming) has been the first author to challenge Bayne and Pacherie's appeal to dispositionalism as a way to defend the doxastic conception of delusions. The core of her strategy consists in highlighting two features of Schwitzgebel's dispositionalism that Bayne and Pacherie overlook, and subsequently arguing that they can't use that view of belief to vindicate doxasticism.[20]

Tumulty accuses Bayne and Pacherie of underemphasizing Schwitzgebel's distinction between *excused non-manifestations of dispositions* and *explained dispositional absences*. In addition, she claims that they also underemphasize the *no-further-fact* clause in Schwitzgebel's account—that is, the observation that once a dispositional profile has been exhaustively specified, there is no further factual question as to whether or not a subject really believes a given proposition. Tumulty argues that dispositionalism is neither plausible nor distinctive without these points, and that these points clash with Bayne and Pacherie's aims. In this section, I present some of Tumulty's points and assess whether they damage Bayne and Pacherie's dispositionalist response to the objections discussed.

2.4.1 Deviations, excuses, and explanations

Bayne and Pacherie see explanations of manifestation-failures as the way to make room for ascriptions of belief in the context of delusions. However,

[20] In section 3 of her forthcoming paper, Tumulty also considers a strategy Bayne and Pacherie don't try—that of emphasizing the role of folk-psychological norms in individuating attitudes—but concludes that even this will not be any help in defending any full-blooded variant of doxasticism. I will leave this strategy aside for reasons of space and focus on her direct argument against Bayne and Pacherie.

Tumulty notes that they seem to miss Schwitzgebel's implicit distinction between excuses and explanations. Whereas an excuse ellucidates why someone fails to manifest a disposition while suggesting she in fact has the disposition, an explanation of an apparent manifestation-failure suggests not that a subject is inhibiting a manifestation of a disposition, but rather that the disposition in question is altogether absent. And, as we saw, such a subject is said to deviate from the dispositional profile for the relevant belief. It should be noted, however, that while we may seem to have evidence for the absence of a disposition whenever we observe a manifestation-failure, only a subset of those manifestation-failures are actually rooted in disposition-absences. Awareness of excusing conditions is necessary if one is not to incorrectly judge a subject to deviate from a dispositional stereotype she actually fits.

If we wish to use Schwitzgebel's dispositionalism in order to count delusions as beliefs, Tumulty suggests, we need to look for *excuses* for all the ways in which deluded subjects fail to look like ordinary believers with respect to the content of their delusions. Bayne and Pacherie (2005, 185) cite a fear of involuntary commitment, for example, to account for the failure of some patients to act on their alleged beliefs. While this does look like an excuse in the required sense—some deluded subjects may know that acting on their beliefs might result in hospitalization—Tumulty is unswayed by the excessive generality of other non-standard situational features to which they appeal in their attempt to account for other typical failures to manifest belief-appropriate dispositions.[21] Bayne and Pacherie, if they want to uphold a dispositionalist form of doxasticism, must answer the question of whether deluded subjects are inhibiting dispositions or rather lack them entirely. Tumulty argues that

> Examining exactly how these features result in missing action or cognition shows that in a number of these cases, if the relevant feature really is a significant factor, then the subject is failing to act or think in the relevant way because they lack the relevant disposition. They aren't manifesting the disposition because they don't have it in the first place. (forthcoming, 12)

Many deluded subjects act (or fail to act) in ways that make it likely they lack one or more of the important dispositions in the sterotype for belief in the content of their delusion. Of course, many non-deluded subjects also deviate from dispositional stereotypes for beliefs that they attribute to themselves, or that others may be tempted to attribute to them. But on the dispositionalist view of belief, the option of deciding that a deviation is not important is left open, and that option may be available with respect to subjects suffering from

[21] Bayne and Pacherie's use of non-standard features to answer some of the objections is discussed below in subsection 2.4.3.

mental illness and abnormal/maladaptive behavior, as well as with respect to normal subjects in the hold of irrational patterns of reasoning.

Hence, a dispositionalist about delusional belief may opt to argue that interpreters could decide that a particular deviation from the dispositional stereotype for a given delusion is relatively unimportant. That would preserve our ability to say that a patient really believes the content of her delusion. However, as Tumulty observes, that wouldn't preserve our ability to say that the sum total evidence of the patient's dispositions points toward her so believing, where her so believing is *a fact over and above* her having the dispositional profile she has. 'It only preserves our ability to refer to those dispositions without misleading our audience about them.' So while dispositionalism has the resources to generate the ascriptive claims Bayne and Pacherie want, 'it does so at the cost of not giving them the same weight they do.' (Tumulty, forthcoming, 25)

2.4.2 Context-dependency

As we have seen, Bayne and Pacherie appropriate Schwitzgebel's claim that belief-ascription is context-dependent. In cases when a deviation cannot be readily excused or explained, 'whether or not the attributor ascribes the belief will depend on the context of the belief ascription and what her interests are' (2005, 181). Tumulty questions the strength of this appeal as a way to achieve a robust vindication of doxasticism. To better understand her point we may briefly review the role of context in Schwitzgebel's account. In one of his examples he discusses

> a child studying for a test [who] reads, 'The Pilgrims landed at Plymouth Rock in 1620,' and remembers this fact. She is a bit confused about what Pilgrims are, though: she is unsure whether they were religious refugees or warriors or American natives. (2002, 257)

Clearly this child (call her Jane) doesn't fully fit the stereotype for believing that the Pilgrims landed on Plymouth Rock in 1620. She won't be disposed to infer, for example, that *Europeans* landed at Plymouth Rock in 1620. In a case like this there doesn't seem to be any available excuse that would render the ascription of the relevant belief uncontroversial. What (if anything) will determine the way we describe Jane's state are the practical matters with which we are concerned, such as her 'likely performance on a history dates quiz' (Schwitzgebel 2002, 257).

The lesson to be drawn, Tumulty argues, is that the introduction of belief-ascriptive language doesn't add to the information already available from the observation of the given dispositional profile, but only refers to it in a conve-

nient and perhaps helpful way.[22] Therefore, attending to the context of belief-ascription makes itself necessary when attributors need to decide whether the use of ascriptive shorthand will be helpful to their audience. It is in this way that Schwitzgebel's dispositionalism provides us with means to end disputes about how to describe in-between cases such as Jane's.

> We might wonder whether she believes 'the Pilgrims landed in 1620', or rather that '*some people* landed in 1620'. Schwitzgebel offers us a way out: specify her dispositions (to do well on the quiz, to imagine Native Americans climbing out of a boat onto Plymouth Rock), and let the communicative expectations decide if 'believing the Pilgrims landed in 1620' is a useful thing to say about Jane. (Tumulty, forthcoming, 8)

In the most difficult cases for ascription, however, the communicative demands on the attributor may not successfully determine whether or not it is appropriate to describe the subject as believing the content of (say) their self-deception, or of their delusion. Cases like these, in which the set of ascribable dispositions available to the interpreter is such a 'mixed bag', leave us only with the option of specification—that is, describing how the subject's dispositions conform to the stereotype for the belief in question and how they deviate from it. There will be times, then, when withholding the use of ascriptive language is going to be preferable so as not to mislead one's audience. Such cases are those in which the observable deviations raise questions regarding both the content of the subject's attitude, and the nature of the attitude itself.

So if there is no way to decide whether something is determinately a case of belief, our move should be to allow *some* indeterminacy in our belief talk, for fear that we should abandon it altogether. This is where Tumulty's observations meet Schwitzgebel's (forthcoming) own remarks on how to handle delusional states—or at least those which defy ascriptive language and practice. He suggests that 'believes that p' should be treated as a vague predicate admitting of vague cases:

> In in-between cases of canonically vague predicates like 'tall', the appropriateness of ascribing the predicate varies contextually, and

[22] Compare Schwitzgebel (2002, 252 fn. 6): 'It may turn out that for you, with your stereotypes, it is accurate to describe me as having a particular belief, while for someone else, with her stereotypes, it is not accurate to describe me that way. Although I suspect that differences in stereotype will tend not to be large enough to produce substantial differences in the appropriateness of belief ascriptions, at least among normal people with shared cultural background, relativism of this sort is not in any case as odious as it may at first appear. On the view espoused in this paper, dispositional facts are fundamental in matters of belief, and the language of belief is employed as a convenient way of grouping together dispositional properties that tend to co-occur.'

> often the best approach is to refuse to either simply ascribe or simply deny the predicate but rather to specify more detail (e.g., 'well, he's five foot eleven inches'); so too, I would argue, in in-between cases of belief. (forthcoming, 6)

Bayne and Pacherie want context-dependency, however, to support the view that *delusions are beliefs* (or at least that some of them are), whereas all it can really offer us is a pragmatic license to talk about delusions as beliefs whenever this is not apt to mislead our intended audience, and whenever there is no better alternative. Therefore, dispositionalism cannot grant them a definitive victory over competing accounts. Besides, it is conceivable that among the many cases that defy belief-ascriptive language there might be some cases of delusion that imagining-ascriptive language is better suited to describe (even if in localized instances, for the benefit of particular audiences). The fact that belief-ascriptive shorthand caters to the context and interests of the ascribers defeats Bayne and Pacherie's purpose of defending a full-blooded doxastic view of delusions by appeal to dispositionalism about belief.

2.4.3 Revisiting the objections to the doxastic account

Finally, in order to complete our assessment of Bayne and Pacherie's dispositionalist approach to delusions, I now turn to presenting and discussing their answers to the objections to the doxastic account I presented in section 2.3, which they separate into three classes: content-based, evidence-based, and commitment-based. Their answers to the first two families of objections do not depend on dispositionalism, but rather on the contention that their proponents exaggerate the connections between belief and logical possibility on the one hand, and between belief and evidence on the other.

The first and second objections presented earlier ('Lack of content' and 'Self-defeating content') converge in holding that what is allegedly believed by delusional subjects clashes with the nature of belief, because believability is taken to imply that the content be meaningful and logically possible. While not all delusional contents raise this kind of objection, those of Cotard and other delusions do. To which Bayne and Pacherie respond first by suggesting that 'issues of belief-ascription are best approached via the question of predictive leverage rather than claims about logical possibility' (2005, 182). Indeed, there seems to be no principled reason to deny that even the most bizarre delusions *lack* content only in the sense and to the extent that they are obviously false or incoherent. If 'I am dead' really were necessarily false or pragmatically self-defeating, a stronger case could be made. But is it either? Bayne and Pacherie argue that it is neither, at least in a number of very common usages—for example, those in which dying does not mean ceasing to exist. But in point of fact, even the stronger 'I don't exist' has been asserted and even argued within

mereological discussions (for instance, in Unger 1979).

The third objection presented earlier ('Lack of evidence') views the absence of support for that which is allegedly believed as conflicting with the nature of belief. In fact, as we saw, delusions typically are held in the face of overwhelming evidence to the contrary. Bayne and Pacherie do not view this as a real problem for doxasticism, denying that there is a *constitutive* connection between belief and evidence. Work by a multitude of researchers has established the existence—indeed, the prevalence—of non-rational elements in belief-formation and maintenance, such as cognitive biases and motivation (Nisbett and Ross 1980; Cherniak 1986; Kunda 1990). They also point out that there seems to be no principled reason to deny the possibility of beliefs being formed as a consequence of brain damage (perhaps not directly). Also, many proponents of so-called *bottom-up* accounts of delusion claim that at least some delusions *are* grounded in evidence of a sort (Bayne and Pacherie 2004). Such accounts suggest that first-person evidence—evidence gathered directly from one's own experience, as opposed to the views of other people or to general knowledge—is at the source of delusion-formation (Davies et al. 2001). It may be worth noting that at least some beliefs in articles of faith are based in what Bayne and Pacherie refer to as 'first-person evidence,' which may be as flimsy as a 'sense' that something is true.

Finally, it is the fourth, fifth, and sixth objections presented earlier ('Theoretical reasoning', 'Practical reasoning', and 'Lack of appropriate affect') that Bayne and Pacherie respond to by invoking dispositionalism. These objections view the deluded subject's circumscribed rational, behavioral and emotional responses as conflicting with the nature of belief. In the context of dispositionalism, the set of commitment-based objections can be thought of as asserting that delusional subjects who seem to believe that p deviate so much from the dispositional profile associated with the belief that p that the burden of demonstrating that we should think of them as believers of p falls squarely with those who wish to defend doxasticism. Their answers to the first two, and most substantive, objections in this group appeal to non-standard aspects of the delusional person's situation that they claim satisfactorily excuse (in the above-discussed, semi-technical sense) the occasional absence of manifestations relevant to the stereotypical dispositions expected to compose the profile.

Theoretical reason, revisited

The first non-standard features invoked by Bayne and Pacherie are the unusual perceptual and affective experiences of deluded subjects, and they are marshalled to meet the objection from theoretical reason. It should be noted that this answer depends on a bottom-up explanation of delusions. If, as bottom-up theorists claim, monothematic delusions are grounded in unusual experiences, 'these conditions may be thought to excuse the patient from man-

ifesting the cognitive dispositions stereotypically associated with their belief' (Bayne and Pacherie 2005, 184). Bayne and Pacherie illustrate this with the case of the Capgras delusion, for which there are strong bottom-up explanations (Bayne and Pacherie 2004; Pacherie 2009). The conclusion they extract from such accounts of Capgras is that the patients *are* manifesting the relevant cognitive dispositions, such as the inclination to test their delusion and consider evidence, and that their abnormal perceptual and affective experiences continually reinforce their delusional belief. But in order for this answer to add to the defense of doxasticism, distinct bottom-up accounts of all the other (monothematic) delusions must be developed. These accounts must show that the subjects of other monothematic delusions have perceptual abnormalities that precipitate their delusions, and also that those perceptual abnormalities cause the cognitive dispositions that are distinctive of belief.

However, hypothesis evaluation and verification isn't something one can point to in every case. For instance, in the case of delusions of thought-insertion or alien control, which involve the belief that one doesn't have ownership, and is not in control of, one's own thoughts and the actions that may emanate thereof (Pacherie, Green and Bayne 2006). While the hypothesis that someone is putting thoughts into one's head is antecedently highly implausible in the context of normal conceptions of causation, Bayne and Pacherie claim that the abnormal experience of agency to which some delusional persons are subjected precludes them from having normal ideas about causation and probability. The hypothesis of thought-insertion may then be formed to make sense of the anomaly, and given that there seemingly is no way to gather evidence for or against such a hypothesis, Bayne and Pacherie claim that its maintenance is understandable. They suggest that we are entitled to assume that the dispositions relevant for the delusional belief are not only present, but also being manifested (though, puzzlingly enough, this cannot be observed).

If such unusual conceptions of causation are at play—seeing that the subject fails to give up her aberrant state via the consideration of its extraordinary implausibility—then what entitles us to assume that her deviations are excusable? This question is rendered irrelevant by the recognition that the two arguments based on the non-standard features presented by Bayne and Pacherie actually aren't aimed at providing *excuses* at all. As Tumulty observes, 'They aim to show that we have failed to see subjects' activities as manifestations of those dispositions' (forthcoming, 16).

Practical reason, revisited

The third feature Bayne and Pacherie discuss is disrupted motivation, and it is invoked to answer the objection from practical reason. That is, it is meant to explain why deluded subjects don't manifest the behavioral dispositions (relevant to belief in the delusional content) that they may nevertheless have.

Before discussing motivation, however, they remind proponents of the objection that, as psychiatrists know all too well, delusion-generated action is not as rare as is often thought. Indeed, Cotard patients often become akinetic and stop everyday activities like eating and washing (Young and Leafhead 1996). Patients suffering from delusions of guilt or self-accusation, sometimes as a consequence of psychotic depression, often engage in physical self-punishment and many attempt suicide (Miller and Chabrier 1988). Patients of de Clérambault's syndrome—as a consequence of falsely believing that another person is secretely in love with them—write letters, make phone calls, send gifts, pay visits, and in a number of cases become violent toward the unwitting object of their obsession (Berrios and Kennedy 2002; McEwan 1997). The list could go on indefinitely, since it simply isn't true that delusional subjects are generally inert with respect to the delusional content (Bortolotti 2010, 162-167).

Nevertheless, as we have seen, there are many cases in which delusions don't cause people to act. In the idiom of dispositionalism: the absence of appropriate behavioral responses challenges belief-ascription. To account for these localized absences of behavioral manifestations, Bayne and Pacherie point out that action is not caused by cognitive states alone but by cognitive states in conjunction with motivational states, and that the motivation to act may not be acquired or not sustained in some cases. Bortolotti and Broome (forthcoming) argue that this may be due to avolition, to emotional disturbances, or to the fact that, given the peculiar content of some delusions, the surrounding environment does not support the agent's motivation to act.[23]

These causes are reflected in Bayne and Pacherie's example of choice, which is that of deluded patients who know that acting on their beliefs might result in hospitalization (Stone and Young 1997). As a consequence, they keep from acting in the way the dispositional profiles for the relevant delusional beliefs would lead us to expect. As Tumulty observes, such distinctive belief-desire pairs can function successfully as excuses: 'they explain why a subject fails to manifest a disposition without undermining the idea that he truly has the disposition. They leave open the possibility that he might manifest it in a different context, when her view of other features of her situation is different' (forthcoming, 16). Indeed, to a lesser extent than delusional subjects, normal subjects also act in ways that are inconsistent with some of their reported beliefs (e.g. hypocrisy), as well as fail to act on some of their beliefs for lack of motivation (e.g. weakness of the will).

The problem is that, to decisively answer the objection from agency, Bayne and Pacherie need some account of those subjects who *don't* have distinctive belief-desire pairs to rationalize their behavior (or lack thereof). One way

[23] Among the causes for disrupted motivation may also be the very content of a patient's delusion. Some contents, such as 'there is a nuclear power station inside my body' (David 1990), may not be conducive to any appropriate course of action.

to approach this would be to claim that since those subjects have generally disrupted motivation, they fail to act in the relevant way. However, regarding such a lack of motivation as sufficient to explain *all* of a subject's failures to manifest relevant behavioral dispositions would also open the possibility that the dispositions aren't there at all. That is, it would open the possibility that if the subject fails to act on her delusion it is because he does not *fully* believe the content of her delusion. In this spirit, Tumulty argues that in some cases, in which subjects fail to manifest the behavioral dispositions in the profile for belief (e.g. that their spouse has been abducted), the explanation is simply that the subject lacks these dispositions. This in turn is due to a persistent, if not permanent, lack of other relevant cognitive and phenomenal dispositions, which indicates that the subject falls short of full-fledged belief. Hence:

> Lack of motivation can't excuse the failure of many deluded subjects to act in ways that would be practically rational given the content of their delusion if their lack of motivation is consequent on their failures fully to believe the content of their delusion. Yet motivation and cognition are often entwined. *In at least some of the cases*, then, where lack of motivation is the explanation for a subject's failure to act in some belief-relevant way, that lack of motivation stems from the absence of dispositions in the profile for that belief. The lack of motivation, then, isn't a mere excuse. And while the presence of a condition like schizophrenia could explain those dispositional absences, the dispositions in question aren't unimportant. These dispositional deviances aren't compatible with *genuine* belief. (forthcoming, 20, my emphasis)

Lack of appropriate affect, revisited

Finally, Bayne and Pacherie's answer to the last objection presented earlier ('Lack of appropriate affect') doesn't appeal to any non-stardard features. They concede that many deluded subjects have deviant emotional and affective responses to the contents of their delusions (for example, some Capgras sufferers are unconcerned about the supposed impostor in their homes). They don't attempt to excuse this deviance but instead resist the thought that emotional and affective dispositions are constitutive elements of the belief stereotype. They also observe that the ascription of emotional states is far from straightforward, since the subjective experience of emotion can dissociate from the behavioral features of emotion—some depressive patients, for instance, have the former but not the latter (Bentall 2003, 225). Given the connection between some forms of delusion and depression, Bayne and Pacherie assert that some delusional patients may also have subjective experiences associated with certain emotions even when they lack the appropriate manifestations. 'If so,

the question of whether such individuals have a certain emotion might not admit of a definitive answer' (2005, 184).

Altough Schwitzgebel states that the dispositional profiles for some beliefs will include dispositions to be in certain emotional states, Bayne and Pacherie are, of course, free to endorse a version of dispositionalism in which some beliefs which would normally be accompanied by a particular emotional state should be present in the absence of that state, as Tumulty points out (forthcoming, 20). However, that doing so raises a new problem, namely, that reducing the number of dispositions in a profile increases the importance of the remaining dispositions. Suppose (for the sake of the argument) that the dispositional profile for a particular belief is taken to contain five dispositions. That means that a person who is missing one of them has four chances to match the profile in other ways, and hence other people have four dispositions to point to in assessing whether or not the person fits the profile. Now suppose that the dispositional profile for a particular belief is taken to contain only three dispositions. Now a person who is missing one of them has only two chances to match the profile in other ways.

Tumulty's concern about Bayne and Pacherie's claim that emotional dispositions never count as essential to any profile for belief, then, is that it leaves would-be believers with fewer dispositions in their respective profiles. That raises the probability, with each disposition they lack, that *that* lack will be the one that tips them over into no longer counting as believers of the relevant proposition (since they are no longer meeting the key portions of the profile). Given that the delusion-prone population has a lot of trouble matching profiles, giving them fewer chances to do so consists in a doubtful move. While their failure to have the 'right' emotion no longer counts against them, their failure to (say) have an appropriate disposition to action counts against them more than it otherwise would.[24] Furthermore, putting more weight on other kinds of dispositions is problematic because delusional subjects don't necessarily do a better job at maintaining those dispositions. Hence, removing the expectation that delusional subjects manifest the relevant emotions dispositions does not make it easier to count them as believers.

2.5 Conclusion

Tumulty's careful consideration of the features to which Bayne and Pacherie call attention—abnormal perceptual experiences, abnormal experiences of agency, and disrupted motivation—shows that appealing to these features does not really help us reach the conclusion at which they aimed to arrive, namely, that most delusional subjects in fact have the belief-relevant dispositions they

[24] I am indebted to Maura Tumulty for clarifying this point to me.

apparently fail to manifest. As far as I can see, Tumulty's results point to a disjunction: we must opt either for doxasticism, or for dispositionalism about delusions. Of course, if one were to opt for defending doxasticism, answers to the usual objections would have to be devised without Bayne and Pacherie's appeal to dispositionalism. An industrious attempt at such an undertaking has been recently presented by Lisa Bortolotti (2010). She does not, however, present us with good enough reasons to discard the kind of approach that has been the center of this investigation.

At first, Bortolotti (2010, 20-1) dismisses what she terms the 'sliding scale' approach on the questionable grounds that such an approach, by not giving a straightforward answer to the question 'Does the patient believe that p?', is unable to characterize precisely whether the patient's actions are intentional, which complicates issues of ethical and policy-guiding import. However, as Schwitzgebel rightly recognizes, apart from that not being nearly enough reason to discard an approach without more ado, its proponents might just as well suggest that 'in many cases of delusion it *shouldn't* be straightforward to assess intentionality, and that the ethical and policy applications *are* complicated, so that a philosophical approach that renders these matters straightforward is misleadingly simplistic' (forthcoming, 7). Ironically, toward the end of her book, Bortolotti hints at the in-between approach we have been discussing when she writes:

> Rarely do we have these clear-cut cases ... Most of the delusions we read about, and we come across, are integrated in the subject's narrative, to some extent, and with limitations. They may be excessively compartmentalised, for instance, or justified tentatively. That is what makes it so difficult to discuss the relationship between delusions, subjects' commitment to the content of the delusion, and autonomy. As authorship comes in degrees, so does the capacity to manifest the endorsement of the delusional thought in autonomous thought and action. (2010, 252)

As Schwitzgebel observes, from the fact that Bortolotti regards authorship and endorsement as necessary for belief,[25] it seems to follow that in the quoted passage she is acknowledging that many actual delusions are in-between cases of belief. This wavering on Bortolotti's part is symptomatic of an increasingly widespread, if latent, perception of which a recent formulation can be found in the words of Tim Bayne: 'there may not be enough determinacy in our ordinary conception of belief for there to be a fact of the matter as to whether many belief-like states are really beliefs or not' (2010, 332).[26] I would like to

[25] See Bortolotti (2010, 242), for instance.

[26] See also Hamilton (2007) for a defense of the absence of a 'fact of the matter' concerning the doxastic state of delusions; and Funkhouser (2009) for a similar conclusion concerning self-deception. Both authors have been influential on the concluding thoughts that follow.

conclude this inquiry with a few additional points in favor of pursuing a sliding scale approach along the lines of what may have been originally contemplated by Bayne and Pacherie, but this time definitively forfeiting the ambition to ascribe doxastic status to 'most' delusional states—a methodological approach to which it seems Bayne himself may now be open, and which is suggested both by the recognition that delusions resist unqualified ascription of doxastic status, and by the preceding analysis of the failed attempt at vindicating doxasticism by resorting to dispositionalism.

I want to claim that the difficulties I have surveyed concerning the ascription of belief to delusional subjects are not due to our limited epistemic perspective—by which I consciously imply that whatever indeterminacy we face in our attempts is a *real* indeterminacy in the phenomena. This claim concerns the nature of the folk-psychological notion of belief and the limits of its application. The underlying assumption in almost all discussions of the doxastic status of borderline phenomena (including self-deception, implicit bias, etc.) is that somehow there are necessary and suffcicient conditions for the application of the concept of belief, such that any given mental state can be determinately classified as either being, or failing to be, a belief. Such an assumption seems groundless for the simple reason that 'belief' is a *vague* concept. Undoubtedly it is a helpful tool in predicting and explaining behavior in ordinary circumstances, which happens when all (or most of) the plausible candidates for assessing its presence converge on the same result—or in dispositionalist terms, when the belief-relevant dispositions are manifested. However, this does not mean that there is *always* a fact of the matter as to whether a subject believes a given proposition. The appropriate response when ascription breaks down, and when persistent disagreement over how to describe a certain kind of mental state arises, is to recognize that no single set of rules is privileged by our ordinary (folk) practices.

I am aware that such an answer may seem unhelpful, and I agree that it should be accompanied by a positive lesson and a direction for future research, in keeping with the goal of attaining a better understanding of delusions and other puzzling phenomena. It is not easy to rest content with a conclusion such as that 'there is a proposition concerning which there is evidence that the subject believes it, and evidence that they do not, and that is the best that can be said' (Hamilton 2007). Indeed, I agree with Eric Funkhouser (2009) that we can be more informative than that. In order to adequately characterize the cognitive states of delusional and other deeply conflicted subjects, we should (in the appropriate contexts) abandon our simple folk-psychological classifications and descend to a lower level of description, namely, to the various dispositions that compose the profile of the subjects in the grasp of the relevant phenomenon. This chimes with Schwitzgebel's aforementioned suggestion that we treat 'belief' as a vague predicate and, when confronted with

difficulties of ascription stemming from its vagueness, turn to providing as much further detail as we may be able to come up with.

So my conclusion is not quite that, say, the Capgras patient *doesn't* believe that her loved one has been replaced by a double, or that the Cotard patient *doesn't* believe that she is dead. Rather, it is that the question as to whether these subjects believe the content of their delusions cannot be answered with a plain 'yes' or 'no'—which doesn't mean we should give up our efforts to understand delusion, but that we should shift our attention to what we *can* do. In this, I enthusiastically agree with George Graham that since 'delusions [are] messy, compound, and complex psychological states or attitudes (thoughts, feelings, and so on), defined more by how persons mismanage their content and fail to prudently act in terms of them, than by qualifying as beliefs,' a realistic picture of delusion 'should leave room for the clinical vagaries of delusional presentation and not try to funnel each case of delusion through the taxonomic filter of the propositional attitude of belief' (2010b, 337). The prolonged debate over how to characterize delusional states is predominantly due to participants using folk-psychological tools that simply can't handle the task.

Conclusão

Ao longo desta dissertação, espero ter-me desincumbido de pelo menos duas modestas tarefas. A primeira foi apresentar ao leitor duas famílias de fenômenos mentais, e as dificuldades que têm sido enfrentadas nos esforços de compreensão e explicação destes. A segunda, mais específica, foi expor uma dificuldade especial que tais fenômenos suscitam, a saber, que eles resistem, por várias razões, a serem descritos, sem mais, como 'crenças'.

No que diz respeito ao autoengano, foi analisado um argumento contra a opção de classificar como doxásticos os estados mentais característicos desse fenômeno. Tal argumento parte da premissa de que seja uma característica essencial das crenças que essas *visem à verdade*, e aponta que, qualquer que seja o tipo de estado produzido pelo autoengano, esse não satisfaz tal condição. Procurei mostrar uma deficiência fundamental nesse argumento, a saber, que a suposta teleologia da crença não acarreta a impossibilidade de estados que, apesar de essencialmente voltados a uma correspondência com o mundo, são impedidos de atingir o seu objetivo por processos extrínsecos a tal característica essencial. A partir disso, uma proposta positiva foi discutida, que substituiria a concepção doxástica do autoengano pela caracterização deste como uma forma de imaginação proposicional. Foi apontado que qualquer candidato a substituir a crença nesse contexto explanatório deveria, no mínimo, desempenhar o mesmo papel que a crença na produção de ações. Subsequentemente, foi examinada uma tentativa de equacionar os poderes motivacionais da crença e da imaginação, concluindo-se que tal tentativa está enredada em uma ambiguidade. Enfim, foi apresentado um argumento que pretende estabelecer que nenhuma atitude cognitiva possui papel motivacional sem a presença de uma crença de fundo—crenças não apenas, em conjunção com desejos, causam e racionalizam ações, mas também determinam se e quando se age com as outras atitudes como *inputs* cognitivos em escolhas e no raciocínio prático. A primeira investigação, portanto, consistiu na avaliação de razões a favor e contra a concepção doxástica do autoengano, e o resultado obtido foi que os argumentos e propostas examinados não são suficientes para suplantá-la. Ainda, se o papel motivacional de todas as atitudes cognitivas que não são crenças de fato depende da existência de uma representação doxástica do cenário prático, então a proposta de classificar o resultado do autoengano como um conteúdo imagi-

nado é derrotada, pois atribuir ao sujeito em autoengano uma representação do cenário prático relevante (que ele, justamente, *precisa* ignorar) conduziria a um paradoxo.

No que diz respeito aos delírios, foi analisada uma tentativa de fugir aos ataques à concepção doxástica desses estados por meio da adoção de uma teoria que flexibilizasse os critérios para a atribuição de crenças. A teoria analisada equipara crenças a conjuntos de disposições estereotípicas de três tipos principais (comportamentais, fenomenais e cognitivas), que podem manifestar-se em diferentes graus. Assim, crer que p nada mais seria que corresponder em um grau adequado, e em certos aspectos adequados ao estereótipo disposicional para crença em p, e uma característica importante dessa teoria é a postulação de que o que contará como 'adequado' dependerá do contexto de atribuição—o que, pode-se inferir, incluirá o público que desejamos informar com a atribuição relevante. Ademais, essa teoria propõe que atribuiremos a um sujeito crença em uma determinada proposição sempre que as maneiras como esse sujeito se afasta do estereótipo possam ser facilmente explicáveis ou desculpáveis por apelo a alguma característica anômala na situação em que ocorrem. Portanto, essa teoria parece acolher a concepção doxástica dos delírios, sob a condição de que encontremos explicações ou desculpas para as ausências de manifestação das disposições relevantes. Antes de analisar as características anômalas que marcam as circunstâncias em que delírios são formados e retidos, foi notado que a distinção implícita entre 'explicação' e 'desculpa' (empregados aqui como termos técnicos) é fundamental para o disposicionalismo apresentado.

Enquanto uma desculpa elucida por que alguém deixa de manifestar uma disposição sugerindo que a pessoa de fato possui tal disposição, uma explicação de uma aparente falha de manifestação sugere não que o sujeito está inibindo a manifestação de uma disposição, mas sim que a disposição em questão está totalmente ausente. Logo, para usar o disposicionalismo para defender a caracterização dos delírios como crenças, é necessário encontrar desculpas para todos os modos nos quais esses sujeitos deixam de manifestar as disposições estereotípicas de crença no conteúdo deliróide relevante (e.g. 'Eu estou morto'). Ademais, foi observado que o apelo à dependência em relação ao contexto não serve ao propósito de uma defesa da concepção doxástica, pois haverão contextos nos quais a atribuição de crenças genuínas não será a opção mais razoável—contextos nos quais um verdadeiro disposicionalista não verá problema algum em (ao invés de induzir seu público em erro) simplesmente especificar em maior detalhe as disposições manifestas pelo sujeito da atribuição. Finalmente, foram considerados possíveis candidatos às características anômalas e examinadas respostas possíveis às objeções contra a concepção doxástica dos delírios, tendo-se concluído que o disposicionalismo não auxilia a defesa desta, mas, antes, resulta em uma concepção empobrecida da atribuição de estados anômalos em geral. Na última seção dessa segunda investigação, ofereci

uma sucinta defesa da ideia de que a tentativa de usar categorias da psicologia do senso comum (*folk psychology*) para classificar estados anômalos é fútil e que devemos, antes, cuidar de detalhar nossa compreensão das disposições envolvidas nesses estados.

Ao término da preparação do segundo ensaio que compõe esta dissertação, meu olhar sobre as questões discutidas no primeiro ensaio já havia mudado. Enquanto o estudo dos problemas envolvidos em classificar como crença os estados mentais distintivos do autoengano suscitaram questões sobre as características supostamente essenciais da crença—a saber, sua teleologia e seu papel fundamental na orientação da ação—o estudo dos problemas envolvidos em assim classificar os delírios tornou-me muito mais cético quanto à possibilidade de descrever com exatidão as condições de aplicação de quaisquer conceitos da psicologia do senso comum. A moral que tiro disso não é exatamente que devamos perseguir um *eliminativismo* com respeito a todas as categorias psicológicas, mas sim que devemos buscar integrá-las em uma estrutura de explicação suficientemente flexível, que dê conta da imprecisão essencial desses conceitos e da consequente margem de indeterminação na prática de sua atribuição. A busca por adequação material na psicologia, portanto, implicará em vagueza na atribuição, porquanto esta espelha a vagueza *dos próprios conceitos* psicológicos empreendidos.

A tarefa de elucidar os estados e a dinâmica tanto de sujeitos em autoengano, quanto de sujeitos com cada um dos diversos tipos de delírio, já vem sendo empreendida proficuamente, ao lado dos esforços daqueles que buscam precisão na *classificação* desses estados. Minha principal direção para pesquisa futura consiste nessa busca por clareza conceitual com respeito aos fenômenos aqui examinados (e outros casos-limite de atribuição de atitudes proposicionais), e na subsequente integração destes em uma teoria compreensiva da mente, fundada na convicção articulada, ao longo da pesquisa que resultou nesta dissertação, de que nenhuma filosofia da mente humana conceptualmente organizada e normativamente informada pode ser construída sem atentar para seus transtornos e vicissitudes.

Bibliografia

American Psychiatric Association (2000) *Diagnostic and Statistical Manual of Mental Disorders*. Fourth edition. Text Revision. American Psychiatric Association.

Audi, R. (1982) Self-Deception, Action, and Will. *Erkenntnis* 18: 133–58.

Barnes, A. (1997) *Seeing through Self-Deception*. Cambridge University Press.

Bayne, T. (2010) Delusions as Doxastic States: Contexts, Compartments, and Commitments. *Philosophy, Psychiatry, & Psychology* 17 (4), 329-336.

Bayne, T. and Fernández, J. (2009) Delusion and Self-Deception: Mapping the Terrain. In T. Bayne and J. Fernández, eds., *Delusion and Self-Deception*. Psychology Press.

Bayne, T. and Pacherie, E. (2004) Bottom-up or top-down: Campbell's rationalist account of monothematic delusions. *Philosophy, Psychiatry & Psychology*, 11/1, 1–11.

Bayne, T. and Pacherie, E. (2005) In Defence of the Doxastic Conception of Delusions. *Mind & Language*, 20/2, 163–188.

Bentall, R. (2003) *Madness Explained: Psychosis and Human Nature*. Allen Lane.

Bermúdez, J.L. (2001) Normativity and rationality in delusional psychiatric disorders. *Mind & Language*, 16/5, 457–93.

Berrios, G.E. (1991) Delusions as 'wrong beliefs': A conceptual history. *British Journal of Psychiatry*, 159, 6–13.

Berrios, G.E. (1996) *The History of Mental Symptoms*. Cambridge University Press.

Berrios, G.E. and Kennedy, N. (2002) Erotomania: a conceptual history. *History of Psychiatry*, 13, 381-400.

Bleuler, E. (1916/1924) *Textbook of Psychiatry* (4th ed), trans. A.A. Brill. Macmillan.

Borge, S. (2003) The myth of self-deception. *Southern Journal of Philosophy* 41: 1–28.

Bortolotti, L. (2004) Can we interpret irrational behavior? *Behavior and Philosophy* 32(2), 359-375.

Bortolotti, L. (2005) Delusions and the background of rationality. *Mind & Language*, 20(2), 189-208.

Bortolotti, L. (2010) *Delusions and Other Irrational Beliefs*. Oxford University Press.

Bortolotti, L. and Broome, M. (forthcoming) Affective dimensions of the phenomenon of double bookkeeping in delusions. *Emotion Review*.

Bratman, M.E. (1992) Practical Reasoning and Acceptance in a Context. *Mind* 101 (401): 1-16.

Breen, N., Caine, D., and Coltheart ,M. (2001) Mirrored-self misidentification: two cases of focal onset dementia. *Neurocase* 7(3), 239-54.

Campbell, J. (2001) Rationality, meaning, and the analysis of delusion. *Philosophy, Psychiatry, & Psychology*, 8/2–3, 89–100.

Canfield, J.V. and Gustafson, D.F. (1962) Self-Deception. *Analysis* 23: 32–36.

Carson, T.L. (2009) Lying, Deception, and Related Concepts. In C.W. Martin, ed., *The Philosophy of Deception*. Oxford University Press.

Chadwick, P., Brichwood M., and Trower, P. (1996) *Cognitive Therapy for Delusions, Voices and Paranoia*. Wiley.

Cherniak, C. (1986) *Minimal Rationality*. MIT Press.

Currie, G. (2000) Imagination, delusion and hallucinations. In M. Coltheart and M. Davies, eds., *Pathologies of Belief*. Blackwell.

Currie, G. and Jureidini, J. (2001) Delusion, rationality, empathy. *Philosophy, Psychiatry & Psychology*, 8/2–3, 159–62.

Currie, G. and Ravenscroft, I. (2002) *Recreative Minds*. Oxford University Press.

Darwin, C. (1871/1981) *The Descent of Man, and Selection in Relation to Sex*. Princeton University Press.

David, A. (1990) Insight and psychosis. *British Journal of Psychiatry*, 156, 798–808.

Davidson, D. (1974) Psychology as philosophy. In S.C. Brown, ed., *Philosophy of Psychology*. Macmillan.

Davidson, D. (1982) Paradoxes of Irrationality. In R. Wollheim and J. Hopkins, eds., *Philosophical Essays on Freud*. Cambridge University Press.

Davidson, D. (1985) Deception and Division. In J. Elster, ed., *The Multiple Self*. Cambridge University Press

Davies, M. and Coltheart, M. (2000) Introduction: Pathologies of Belief. In M. Coltheart and M. Davies, eds., *Pathologies of Belief*. Blackwell.

Davies, M., Coltheart, M., Langdon, R. and Breen, N. (2001) Monothematic delusions: Towards a two-factor account. *Philosophy, Psychiatry & Psychology*, 8/2,3, 133–58.

De Pauw, K.W. and Szulecka, T.K. (1988) Dangerous delusions: Violence and the misidentification syndromes. *British Journal of Psychiatry*, 152, 91–96.

Demos, R. (1960) Lying to Oneself. *Journal of Philosophy*, 57: 588–95.

Dennett, D. (1987) Making sense of ourselves. In *The Intentional Stance*. MIT Press.

Dickerson, F.B. (2000) Cognitive behavioural psychotherapy for schizophrenia: A review of recent empirical studies. *Schizophrenia Research*, 43/2,3, 71–90.

Egan, A. (2009) Imagination, delusion, and self-deception. In T. Bayne and J. Fernández, eds., *Delusion and Self-Deception*. Psychology Press.

Ellis, H.D., Young, A W., Quayle, A.H. and de Pauw, K W. (1997) Reduced autonomic responses to faces in Capgras delusion. *Proceedings of the Royal Society: Biological Sciences* B264, 1085–92.

Engel, P. (2004) Truth and the Aim of Belief. In D. Gillies, ed., *Laws and Models in Science*. King's College Publications.

Fodor, J.A. (1983) *The Modularity of Mind*. MIT Press.

Freud, S. (1917/1958) A Childhood Recollection from *Dichtung und Wahrheit*. In J. Strachey et al., eds., *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Hogarth Press.

Funkhouser, E. (2009) Self-Deception and the Limits of Folk Psychology. *Social Theory and Practice* 35(1): 1–13.

Gendler, T.S. (2007) Self-deception as pretense. *Philosophical Perspectives* 21 (Philosophy of Mind), 231–258.

Gendler, T.S. (2008) Alief and Belief. *Journal of Philosophy* 105, 634–663.

Gergen, K. (1985) The Ethnopsychology of Self-Deception. In M. Martin, ed., *Self-Deception and Self-Understanding*. University of Kansas Press.

Gerrans, P. (2001) Authorship and ownership of thoughts. *Philosophy, Psychiatry & Psychology*, 8/2,3, 231–237.

Gerrans, P. (2009) Mad scientists or unreliable narrators? Dopamine dysregulation and delusion. In M. Broome and L. Bortolotti, eds., *Psychiatry as Cognitive Neuroscience: Philosophical Perspectives*. Oxford University Press.

Goodman, N. (1983) *Fact, Fiction, and Forecast*. Harvard University Press.

Graham, G. (2010a) *The Disordered Mind: An Introduction to Philosophy of Mind and Mental Illness*. Routledge.

Graham, G. (2010b) Are the Deluded Believers? Are Philosophers Among the Deluded? *Philosophy, Psychiatry, & Psychology* 17 (4), 337-339.

Haight, M. (1980) *A Study of Self-Deception*. Harvester Press.

Hamilton, A. (2007) Against the belief model of delusion. In *Reconceiving Schizophrenia*. Oxford University Press.

Hoorens, V. (1993) Self-enhancement and Superiority Biases in Social Comparison. *European Review of Social Psychology*, 4(1): 113–139.

Jaspers, K. (1913/1963) *General Psychopathology*, trans. J. Hoenig and M.W. Hamilton. Manchester University Press.

Johnston, M. (1988) Self-Deception and the Nature of Mind. In B. McLaughlin and A. Rorty, eds., *Perspectives on Self-Deception*. University of California Press.

Kant, I. (1797/1996) *The Metaphysics of Morals*, trans. M.J. Gregor. Cambridge University Press.

Kipp, D. (1980) On Self-Deception. *Philosophical Quarterly*, 30: 305–17.

Kripke, S. (1980) *Naming and Necessity*. Harvard University Press.

Kruger, J. and Dunning, D. (1999) Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments. *Journal of Personality and Social Psychology*, 77(6): 1121–34.

Kunda, Z. (1990) The case for motivated reasoning. *Psychological Bulletin* 108/3, 480–498.

Langdon, R. and Coltheart, M. (2000) The cognitive neuropsychology of delusions. In M. Coltheart and M. Davies, eds., *Pathologies of Belief*. Blackwell.

Lazar, A. (1999) Deceiving Oneself or Self-Deceived? On the Formation of Beliefs 'Under the Influence.' *Mind* 108: 265–90.

Maher, B. (1988) Anomalous experience and delusional thinking: The logic of explanations. In T.F. Oltmanns and B.A. Maher, eds., *Delusional Beliefs*. Wiley.

Martin, M.W. (1979) Self-Deception, Self-Pretence, and Emotional Detachment. *Mind* 88: 441-446.

Martin, M.W. (1986) *Self-deception and Morality*. Kansas University Press.

McEwan, I. (1997) *Enduring Love*. Vintage.

McKay, R., Langdon, R. and Coltheart, M. (2009) "Sleights of Mind": Delusions and Self-deception. In T. Bayne and J. Fernández, eds., *Delusion and Self-Deception*. Psychology Press.

Mele, A.R. (1987) *Irrationality*. Oxford University Press.

Mele, A.R. (1997) Real Self-Deception. *Behavioral and Brain Sciences* 20: 91–102.

Mele, A.R. (2001) *Self-Deception Unmasked*. Princeton University Press.

Michel, C. and Newen, A. (2010) Self-Deception as Pseudo-Rational Regulation of Belief. *Consciousness and Cognition* 19 (3): 731-744.

Miller, D.T. and Ross, M. (1975) Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin*, 82, 213-225.

Miller, F.T. and Chabrier, L.A. (1988) Suicide attempts correlate with delusional content in major depression. *Psychopathology*, 21(1): 34-7.

Moran, R. (2001) *Authority and Estrangement*. Princeton University Press.

Neu, J. (2000) Life-Lies and Pipe Dreams: Self-Deception in Ibsen's The Wild Duck and O'Neill's The Iceman Cometh. In *A Tear Is an Intellectual Thing*. Oxford University Press.

Nisbett, R. and Ross, L. (1980) *Human Inference*. Prentice-Hall.

O'Brien, L. (2005) Imagination and the motivational view of belief. *Analysis* 65 (285): 55–62.

Pacherie, E. (2009) Perception, emotion, and delusions: The case of the Capgras delusion. In T. Bayne and J. Fernández, eds., *Delusion and Self-Deception*. Psychology Press.

Pacherie, E., Green, M., and Bayne, T. (2006) Phenomenology and delusions: Who put the 'alien' in alien control? *Consciousness and Cognition*, 15, 566–77.

Pears, D. (1984) *Motivated Irrationality*. Oxford University Press.

Price, H.H. (1960/1969) *Belief*. George Allen and Unwin.

Putnam, H. (1975) The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science*, 7: 131–93.

Quine, W.V. and Ullian, J. (1970) *The Web of Belief*. Random House.

Ramsey, F.P. (1931) *The foundations of mathematics and other logical essays*. Routledge.

Rey, G. (1988) Toward a computational account of akrasia and self-deception. In B. McLaughlin and A. Rorty, eds., *Perspectives on Self-Deception*. University of California Press

Rowbottom, D.P. (2007) 'In-Between Believing' and Degrees of Belief. *Teorema* 26, 131-137.

Ryle, G. (1949) *The Concept of Mind*. Hutchinson.

Sartre, J.-P. (1949/1957) *Being and Nothingness*, trans. H. Barnes. Pocket Books.

Sass, L. (1994) *The Paradoxes of Delusion*. Cornell University Press.

Schreber, D.P. (1903/2000) *Memoirs of My Nervous Illness*. New York Review of Books.

Schwitzgebel, E. (2001) In-between believing. *Philosophical Quarterly*, 51, 76–82.

Schwitzgebel, E. (2002) A phenomenal, dispositional account of belief. *Noûs* 36/2, 249–275.

Schwitzgebel, E. (forthcoming) Mad Belief? *NeuroEthics*.

Shannon, C.E. and Weaver, W. (1963) *The Mathematical Theory of Communication*. University of Illinois Press.

Siegler, F.A. (1962) Demos on Lying to Oneself. *Journal of Philosophy*, 59: 469–475.

Smullyan, R.M. (1986) Logicians who reason about themselves. In *Proceedings of the 1986 Conference on Theoretical Aspects of Reasoning about Knowledge*. Morgan Kaufmann.

Spinoza, B. (1677/1996) *Ethics*, trans. E.M. Curley. Penguin Books.

Stephens, G.L. and Graham, G. (2004) Reconceiving delusions. *International Review of Psychiatry*, 16/3, 236–241.

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

Tanney, J. (2009) Gilbert Ryle. In Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Winter 2009 edition). URL = http://plato.stanford.edu/entries/ryle/

Tumulty, M. (forthcoming) Delusions and Dispositionalism about Belief. *Mind & Language*.

Unger, P.K. (1979) I Do Not Exist. In Graham F. Macdonald, ed., *Perception and Identity*. Cornell University Press.

Van Leeuwen, D.S.N. (2007) The Product of Self-Deception. *Erkenntnis* 67(3), 419–437.

Van Leeuwen, D.S.N. (2009) The Motivational Role of Belief. *Philosophical Papers* 38(2), 219–246.

Velleman, J.D. (2000) On the Aim of Belief. In *The Possibility of Practical Reason*. Oxford University Press.

Velleman, J.D. and Shah, N. (2005) Doxastic deliberation. *Philosophical Review* 114: 497–534.

Weinstein, E.A. and Kahn, R.L. (1955) *Denial of Illness*. Charles C. Thomas.

Young, A.W. (1998) *Face and Mind*. Oxford University Press.

Young, A.W. (1999) Delusions. *The Monist*. 82/4, 571–589.

Young, A.W. and Leafhead, K. (1996) Betwixt life and death: Case studies of the Cotard delusion. In P. Halligan and J. Marshall, eds., *Method in Madness*. Psychology Press.