RADICAL COGNITIVE SCIENCE IN PHILOSOPHICAL PSYCHOPATHOLOGY: THE CASE OF DEPRESSION

By

ALEXANDER JAMES IBBS MILLER TATE

A thesis submitted to the University of Birmingham for the degree of DOCTOR OF PHILOSOPHY

Department of Philosophy

School of Philosophy, Theology, and Religion

College of Arts and Law

University of Birmingham

January 2019

Radical Cognitive Science in Philosophical Psychopathology: The Case of Depression

Abstract

The principle purpose of this collection of papers is to explore and apply ideas from various kinds of non-traditional Cognitive Science, as well as comparing them with their more traditional counterparts, in order to reach a better understanding of the symptoms and features of depressive illness. By 'non-traditional' I mean to refer to Cognitive Science that makes minimal use of the notion of abstract, post-perceptual, and reconstructive mental representation, is computationally frugal, and treats the mind as fundamentally both embodied and environmentally embedded. This thesis in particular draws on insights from ecological psychology and action-oriented perception, embodied and situated cognition, and predictive processing. After introducing the subject matter, the first substantive paper argues that anhedonia is, in the general case, a disorder determined by disruption to affectively supportive elements of an individual's environment. The second proposes a predictive-processing approach to explaining the characteristic operation of motivational mental states. This paper supports the third, in which I argue that psychological, somatic, and (action-oriented) perceptual factors all contribute to depressed agents' struggles and failures to initiate and sustain action. I suggest that these problems should not all be thought of as disorders of motivation *per se*, but rather as broader kinds of action-oriented cognitive dysfunction. In the fourth paper, I reject Matthew Ratcliffe's argument for the claim that people with depression are not typically better able to empathise with other people with depression, though I find alternative evidence for this suggestion available to those happy to endorse a more mainstream view of empathy. Finally, I broaden the scope of my investigation to psychopathology in general, and argue that classical (neuro-centric and mechanical) explanations in Psychiatry have inadvertently resulted in psychiatric service users' subjection to a number of epistemic injustices. This suggests that non-classical theories of psychopathology are not just important for achieving accurate psychiatric explanation, but also for ensuring the ethical treatment of service users.

Dedication

For Becki Luscombe and Bryn Gough. Wise beyond their years, and gone too soon. Their voices inspired and guided this work. I hope they would have found something worth reading within its pages.

"It is quite often said that depression, or indeed any form of chronic mental unease, is like being followed by a cloud. I disagree. Clouds can be nice. If clouds form in a certain way they can sometimes resemble a family of turtles playing volleyball. Mental illness very rarely resembles even a SINGLE turtle playing volleyball."

Becki Luscombe, The Boxticker

Acknowledgments

I would never have completed this thesis without the support, both personal and philosophical, of an enormous number of people who I have had the good fortune to know for some or all of the last 26 years. What follows is probably an incomplete, but hopefully representative, list of the most important folks. In a very real sense, everything that comes next is their fault[1].

Mae Rohani, Ash Allen, Rachel Elkin, Chris Forster, Caroline Miller Tate, Philip Miller Tate, Nick Davidson, Sarah Dovey, Rosie Ellis, Amber Culley, Ben Puusta, Jen Elliott, Freya Watkins, Nikk Effingham, Malcolm Price, Farrell Baker, Robert Lydiard, Chris Harding, Sarah Davies, Lisa Bortolotti, Iain Law, Will Davies, Rachel Upthegrove, Joel Krueger, Michael Larkin, Matthew Parrott, Matilde Aliffi, Em Walsh, Federico Bongiorno, Michael Roberts, Kash Sunghuttee, Rachel Gunn, Maisie Gibson, Yingna Li, Alex Blanchard, Lauren Melchor, Marco di Natale, John Parry, Lauren Graham-Symonds, Sophie Stammers, Herjeet Marway, Anneli Jefferson, Elisabeth Muchka, Eugenia Lancellotta, Henry Taylor, Scott Sturgeon, Maja Spener, Ema Sullivan-Bissett, Matthew Parrott, Joel Krueger, Tom Davies, Tom Baker, Ray Schuur, Stew Yarlett, Casey Elliott, Femi Taiwo, Amy Conkerton-Darby, Helen Ryland, Will Sharp, Josh Brown, David Irvine, Jodie Neville, Kate MacKay, Matt Tugby, Andy Clark, Aidan McGlynn, Dave Carmel, Alistair Isaac, Till Vierkannt, Laura Jenkins, Emma Foster, Pete Kerr, Stephen Bates, Joanna Nowlan, Joe Cunningham, Philip Ellwood, Melvin Tiley, Andrew Watters, Shelagh Frawley, Mike Fitzgerald, Rob Nayman, Peter Lamb, Rory Scott, Tom Russell, Norma Robertson, James Bowker, Will Sandys, Matt Hewson, Alice Moneypenny, Ruth Samuel, Sophie White, Kathleen Murphy-Hollies, Julius Elster, Jonny Lee, Joe Dewhurst, Urtė Laukaitytė, Tamarinde Laura, Wout van Praet, María Jiménez, Dig Wilk, Mahi Hardalupas, Imke von Maur, Valentina Petrolini, André Grahle, Rebekka Hufendiek, François Jaquet, Andrea Valentino, Amica Nowlan, Dan Fisher, Jack Kirkby, Sam Boocock, Emma Graham, Ellie Price, Jamie Banurji, Inge Hertzog, Tim Liversage, Cat Bawtree, Emily Kitson, Katharine Jenkins, Bren Markey, Nora Berenstain, Nathaniel Adam Tobias ~~Coleman~~, Joseph Kisolo-Ssonko, Rachel Cooper, Hane Maung, Phoebe Friessen, Moujan Mirdamadi, Sabine Wantoch, Natalie Ashton, Nadia Mehdi, Cecily Whiteley, and Stephanie Harvey.

---

[1] Apart from the typos; those are on me.

Contents

Chapter 1: Motivating Novel Explanatory Strategies in Depression Research

1.1. Introduction

At our best guess, approximately 3.3% of the population of England are currently experiencing clinically significant depression, and around 7.8% are currently experiencing a clinically significant mixture of anxiety and depression (McManus et al 2016). If you are one of those people then, after you've been diagnosed, it might be natural to try and find out more about what depression is. If you seek an answer to this question via Google, you are likely to find an article on 'psychiatry.org' purporting to answer exactly this question (Parekh 2017).

In a section on risk factors for depression, Parekh lists 'biochemistry', 'genetics', 'personality', and 'environmental factors'. From this you would likely get the impression that the factors that explain your depression and its symptoms are mostly internal to you (biochemistry, genetics, and personality). Moreover, when you peruse the specifics of how 'environmental factors' are thought to dispose people to depression, you would be forgiven for thinking that the kinds of factors at issue are both relatively extreme and rare; "[c]ontinuous exposure to violence, neglect, abuse or poverty" (Parekh 2017). If you are continuously exposed to none of these things, then you might be forgiven for dismissing environmental factors from the explanation of your own depression altogether. Other sources likely to come high on a list of search results would largely validate this suspicion. An article from Harvard Health Publishing (HHP 2017) discusses in detail the role of the Amygdala, Thalamus, Hippocampus, various neurotransmitters, genes, and your 'temperament', on the onset of depression. Even when the author does dive into the topic of stressful life events, they emphasise the litany of physiological consequences such events bring about, and the topic quickly returns to matters of an individual's biology. You might well conclude that your depression is, first and foremost, a disorder explicable (almost) entirely in terms of factors internal to you.

Scrolling a little further through Parekh's article, or investigating elsewhere how your condition can be treated, you are likely to come across two main suggestions;

pharmacological treatment and 'talking therapy', usually Cognitive Behavioural Therapy (CBT) (Parekh 2017; HHP 2017; NHS 2016). Aside from once again emphasising how depression is internal to you, how CBT is said to work might suggest something else to you about depression; that it is grounded in thoughts and feelings about yourself and the world that are inappropriate or dysfunctional and need to be revised, or reframed. As Parekh writes;

> CBT helps to recognize *distorted* thinking and then change behaviors and thinking. (2017)

You might be forgiven for concluding from this that your depression stems from some kind of *error* in how you view the world. That it is an internal mistake or dysfunction in the processes and states that underpin your experience of reality.

Finally, you would probably be struck throughout your search at how most sources of information about depression provide little or no detailed description of what being depressed *is like*. In particular, there are very few such descriptions from people who have actually experienced depression. This is despite the fact that many sources are keen to distinguish depression from things like grief or sadness. Although a few distinguishing features are noted, such as prevalence of self-loathing or persistence of low mood (Parekh 2017), these give little positive insight into the experience of depression. This might be particularly frustrating if you are conducting this search on behalf of a loved one rather than yourself. Many significant resources on the topic are likely to feel somewhat detached from the subject matter.

Naturally there are exceptions to all of these trends, and some such exceptions might well come up even in an initial search. But, in general, the more the source of information is connected with the medical profession, rather than service users themselves, journalism or activist organisations (such as Mind or SANE), the more these generalisations will ring true. Moreover, they are acknowledged by many practicing and research psychiatrists (e.g. Bracken et al 2012). These trends, I shall suggest, reflect the predominance of a set of assumptions in psychiatric research and practice that are associated with what I shall call *traditional* cognitive science. The main theme of this collection of papers is to see how our understanding and

explanations of depression, as well as the wider constructs used to understand it and clinical practice in Psychiatry more generally, might change if we take the emphasis away from, or even flatly reject, these assumptions.

In this introductory chapter I provide an overview of the literature in both depression research and philosophy of cognitive science that forms the backdrop to the papers in this thesis. It does not aim to be a remotely comprehensive conceptual or historical tour of the contrasts between traditional and revisionary approaches to cognitive science, even in the specific context of psychiatry, or depression research. Rather, it aims to be a whistle-stop tour of the features of traditional and non-traditional cognitive science and depression research essential to fully appreciating the motivations and aims of this particular thesis.

I shall begin by giving a selective exegesis of traditional approaches in cognitive science. There are three specific assumptions built in to this tradition that I will pay particular attention to.

The first standard assumption is that the explanatory target of cognitive science is what occurs 'internally', between the input of sensory transducers and the output of motor activity. The second is that invoking algorithm-like procedures over detailed, reconstructive representations of the world is the best explanatory strategy in most cognitive science. The third is that constraints on actual explanation in cognitive science are primarily *functional*, in the sense that they promote explanation of how some cognitive capacity is realised, rather than *phenomenological*, in the sense of promoting explanation of what it is like to be in some cognitive state, or undergoing some cognitive process.

Next, I give an indicative overview of mainstream depression research, which demonstrates that these assumptions carry over into this area of applied cognitive science. After that, I explain why and how one might challenge each of the central assumptions listed above, and outline how these challenges will be deployed in the following thesis, to explain and/or enhance understanding of certain features and symptoms of depression.

Having identified the role of traditional cognitive science in depression research, I explain how a particular non-traditional approach to cognitive science, Predictive Processing, approaches the study of cognition, and how this challenges tradition in a variety of ways. This provides the backdrop to understanding the approach to motivation, a key construct in explanations of depression, that I provide in Chapter 3. Finally, I suggest that given the nature of Psychiatry, challenging mainstream explanatory assumptions and strategies is likely to have important ramifications for ensuring ethical clinical practice. This gives some background to the final chapter of the thesis. I close by giving an overview of the five main papers that compose this thesis.

1.2. Traditional Cognitive Science

1.2.1. Representations and Internalism

The traditional discipline of cognitive science owes much to a specific computational metaphor; that the mind is software running on the hardware of the brain. Since computation is, effectively, symbol manipulation, cognition is typically framed as the computational manipulation of symbols that represent the world. As Paul Thagard summarises,

> The central hypothesis of cognitive science is that thinking can best be understood in terms of *representational* structures in the mind and computational procedures that operate on those structures…Most work in cognitive science assumes that the mind has mental representations analogous to computer data structures, and computational procedures similar to computational algorithms." (Thagard 2018: §3)

I shall not concern myself in this chapter with the significant literature dedicated to understanding the nature of mental representation (but see Pitt 2018 for a very helpful overview). It will suffice to note that not only does traditional cognitive science posit that minds *have* such representations and procedures, but that it makes very heavy use of these theoretical posits in explaining cognitive phenomena. As Larry Shapiro puts it,

> …cognitive scientists typically view cognitive processes as computational. Commensurate with this view is the idea that *cognition consists in the manipulation of symbols*, where these manipulations often involve the application of rules for the purpose of deriving conclusions…Because cognitive operations begin with the receipt of symbolic inputs and end with the production of symbolically encoded outputs, the subject matter of cognitive science lays *nestled between the peripheral shells of sensory organs and motor systems*…(Shapiro 2011: 28; *emphasis mine*)

Two aspects of Shapiro's overview are particularly important here. Firstly, traditional cognitive science takes cognition to *consist in* the procedural manipulation of representational symbols. That is, not only is symbol manipulation explanatorily central to cognitive science, but, more strongly than that, anything that is not symbol manipulation does not count as cognition. This position makes representational symbols, and the computational procedures to manipulate them, explanatorily *indispensable* to traditional cognitive science. That is, the kind of explanations of cognitive phenomena that marks cognitive science out as a discipline are ones that proceed by way of positing the computational manipulation of mental representations. Moreover, many (if not the majority) of the relevant representations are typically assumed to be non-perceptual, rich reconstructions of reality. Perceptual representations are taken to be the *initial* inputs to processes of inference, and are generally thought to only partially explain complex cognitive capacities (such as directed action) by themselves. For perceptual representations to have an impact on cognitive capacities beyond perception itself, it is typically assumed that they will need to be *transformed* into mediating non-perceptual representations by inferential procedures, which function as mental *reconstructions* of the external environment. That is, the initial products of sensory transduction need *transforming* before they feed into most higher cognitive capacities. Andy Clark describes this assumption perspicuously by contrasting it with a non-reconstructive alternative view,

> "…behavioural success is not the outcome of reasoning defined over a kind of inner replica of the external world…They do not use sensing, moment-by-

> moment, to build an inner model that recapitulates the structure and richness of the real world…" (2017: 732)

Call the idea that cognition primarily operates in this richly reconstructive manner the *strong representationalist* assumption.

Secondly, as we see in Shapiro's description above, traditional cognitive science takes these posited cognitive processes to occur strictly *after* sensory transduction (the process of transforming light in the case of vision, vibrations in the case of audition, and so on, into electrical signals that impinge upon the brain) and prior to the engagement of motor systems. In short, traditional cognitive science proceeds on the assumption that cognitive activity occurs in the brain (or, more liberally, in some areas of the central or perhaps even peripheral nervous system), since this is the location of the activity that follows transduction and precedes action. The subject matter of cognitive science resides firmly behind the veil of skin and skull[2]. Call this the *internalist* assumption.

Several of the key challenges to traditional cognitive science take issue with the assumptions of strong representationalism and internalism. I shall present these challenges in section 1.4. For now however, I shall articulate one other key feature of traditional cognitive science; its silence on matters of phenomenal experience.

1.2.2. Phenomenal Irrelevance

Explanation in traditional cognitive science, it is widely agreed (Drayson 2009; Bechtel 2009; Gervais & De Jong 2012), has as its target an individual's cognitive capacities or competencies. That is, traditional cognitive science aims to explain what individuals can *do*, cognitively speaking. For instance, a cognitive scientist may seek to explain facial recognition (Bruce & Young 1986), or speech production (Levelt 1989).

It is not as widely agreed exactly what the explanatory strategy of traditional cognitive science amounts to (or ought to amount to) beyond this. For many years,

[2] That is not to say that traditional cognitive scientists are committed to the processes of mind being *identical* to the processes of the brain, but rather that human mental processes are *in fact* the result of nervous system activity only. This is consistent with, for instance, various kinds of mental state/process functionalism.

the central idea was that cognitive science aims at *functional explanations* of cognitive capacities (Gervais & De Jong 2012; Lycan 1987). This is a style of explanation where one begins with a functional description of the cognitive capacity of interest (e.g. speech production) and proceeds by deconstructing that capacity into sub-capacities thought to compose it (e.g. conceptualization, formulation, and articulation), explaining the relationships between those capacities and, where necessary, further decomposing those sub-capacities (Drayson 2012; Gervais and De Jong 2012; Levelt 1989).

More recently there has been a great deal of interest in *mechanistic explanations* as the preferred style of explanation in cognitive science (Bechtel 2009; Gervais & De Jong 2012). These are closely related to, but importantly distinct from functional explanations. They begin, as do functional explanations, "with decomposing the capacity or function to be explained into different sub-functions…" (Gervais & De Jong 2012: 154). But instead of remaining silent on what realizes these functions, mechanistic explanations also specify "what entities perform those functions…" (154).

Regardless of which form of explanation one thinks is or should be more common in traditional cognitive science, they each focus on functionally described cognitive capacities. That is, they concentrate on explaining what humans, or other organisms, can *do*, cognitively speaking. They do not, however, pay much attention to *what it is like* to exercise those capacities, and still less to explaining *why* those capacities have a particular phenomenology. They typically lack phenomenological detail (they are *thin*) and often run a number of phenomenologically distinct categories together (they are *imprecise*). As Zoe Drayson notes, traditional cognitive science,

> …focuses on the functional details of the cognitive processing itself and has traditionally had little or nothing to say on 'phenomenology', the experiential quality of conscious mental life. Computational models allow for thought processes to be considered scientifically, but it has traditionally been difficult to find a place for the subjective first-person perspective in the orthodox science of cognition. (2009: 330)

One specific effect of this lack of focus on phenomenology has been little or no phenomenological constraint on explanation or interpretation of raw data. That is, traditional cognitive scientific explanations are not typically evaluated with respect to how well they explain what exercising a particular cognitive capacity is like. They are, instead, only required to explain the results of exercising that capacity. Call this the *phenomenal irrelevance* assumption.

I have provided an overview of critical assumptions of traditional cognitive science that will be brought into question in this thesis. Next I wish to highlight the role such assumptions have played in explanations of depression specifically. This will help the reader to get a grasp on the nature of the alternative explanations that this thesis explores.

1.3. Traditional Depression Research

1.3.1. Strong Representationalism in Depression Research

Information-processing theories that account for depressive symptoms in terms of abnormal representations, or else abnormal patterns of inference over representational states, are incredibly common in the psychiatric research literature. For instance, Kuhl & Helle propose that,

> the chronicity of depressive mood states is maintained by…degenerated (unfulfillable) intentions that claim working memory capacity needed to enact new (fulfillable intentions). (1986: 247)

This is explained in terms of the information encoded in (read: representational content of) intentions, degenerated or otherwise. The basic idea is that depression involves the persistent accumulation of intentions which lack the content necessary for them to be fulfilled (e.g. precisely how or when an action is to be performed). This leads to cognitive incapacity including working memory deficits, due to an overabundance of accumulated, and non-dischargeable representational states. These degenerated intentions are also the target of depressive rumination, since one persistently tries and fails to satisfy them. One central problem for sufferers of depression, then (on this account), is a pathological absence of certain kinds of representational content.

Another example is found in Ingram (1984), where it is proposed that depression is caused and maintained by the activation of a theoretical construct referred to as the 'depression node' due to life-events activating connected memory constructs that have cognitive and affective contents associated with loss, that are consistently 'recycled' through conscious awareness (452-7). There are a lot of concepts flying around here (some that are significantly clearer and more helpful than others), but the central takeaway here is that Ingram leans heavily on explaining depression with reference to specific loss-focused representational content, and aberrant ways of processing that content ('recycling').

This is not just a phenomenon to be found in the depression research of the mid-80's. Dozois & Dobson (2001) suggest that a central feature of depression is the presence of "an interconnected negative self-representational system and [the] lack [of] a well-organised positive template of self" (236). This is posited to be the result, fundamentally, of the way in which negative content (specifically that which is self-directed) "is activated, processed, and/or organized" (243). Roepke & Seligman (2015) propose that the negative "mental representation of possible futures" is "the core causal element of depression" (23). These are just a few representative examples of individual studies.

Moreover, Beck's cognitive theory of depression, which is immensely influential on psychiatric practice as well as being theoretically popular, centres on maladaptive, false beliefs (about oneself, others, and the surrounding world – the so-called *cognitive triad*) as being the ground for other affective and behavioural symptoms, which are considered secondary effects (1979). This theory spawned Cognitive Behavioural Therapy, an intervention designed to correct these negatively biased beliefs and so, it is hoped, provide relief from the upstream behavioural and affective depressive symptoms as well (Beck et al 1979).

Common to all of these theories, and to the interpretation of all the empirical data they are based on, is the search for or positing of either disordered representational structures, biased inferential processes off the back of such structures, or both. The cognitive science underlying much mainstream thought regarding depression is, in this sense, highly representational in character.

1.3.2. Internalism in Depression Research

> *There can be no such thing as a psychiatry which is too biological.* (Guze 1989: 316)

It is common to find that contemporary researchers (both philosophical and psychiatric) ultimately understand depression, along with many if not most other mental illnesses, in terms of disturbances to the ordinary function of neural circuitry (Hutto 2016; for philosophical examples of this tendency see Murphy 2006 & Gerrans 2014). This kind of neurocentrism is particularly apparent in the kind of research promoted by the NIMH's recent Research Domain Criteria project (Tabb 2017; Lilienfeld et al 2015; Parnas 2014). As Parnas puts it,

> The RDoC's theoretical underpinning appears to be a neurocentric "type-type" reductionism: specific chunks (types) of mental life (e.g. hallucination, anhedonia) are identical with, or nothing else than, certain specific chunks (types) of neural activity (say, a certain configuration of interactions between dysfunctional neural networks). (2014: 46)

And further,

> The RDoC's target constructs, [are] believed to reflect simple, natural-kind like behavioral functions and [to be] instantiated in circumscribed neural networks… (2014: 47)

The key feature of this way of understanding mental illness, including depression, is that it posits that mental illness' physical realisation is to be found in the activities of the brain or, perhaps, some activity of the wider nervous system. This reflects the internalist presumption of much cognitive science. It is generally thought that interactions between brain, body, and world characteristic of depression, to the extent that they are acknowledged at all, are secondary effects of abnormal neural activity, rather than having any independent explanatory weight of their own.

In particular, the natural and social environment are generally either ignored, cursorily mentioned, or (at best) given a simple causal role in the generation of depression and other forms of psychopathology, for instance via negative life-events, which nevertheless are said to *persist* due to their effects on disordered

internal representations and inferential processes, grounded in neural circuitry (Colombetti 2013; Drayson 2009). Much rarer is any investigation of the possibility that the environment plays an active role as an enabler or even partial realiser of certain capacities that are diminished or otherwise changed in cases of depression. This limited way of conceptualising the role of the environment masks certain theoretical possibilities from researchers, including the possibility that depressive symptoms may occur absent many (or, potentially, any) faulty inner processes. Diminished capacity may instead be a direct consequence of a failure of external support that is ordinarily required to achieve it, with the nervous system mediating the influences of this unsupportive environment in a perfectly ordinary, functional manner.

1.3.3. Phenomenology in Traditional Depression Research

The assumption of phenomenal irrelevance has infected traditional understandings of depression (that have their contemporary basis in traditional cognitive science) in at least two ways. Firstly, the diagnosis of a major depressive episode or, if recurrent, major depressive disorder typically relies on the identification in an individual of the thinly described construct "depressed mood" (DSM-5: 163). Secondly, when seeking to understand depressed mood and other symptoms traditional cognitive science has focused, not entirely unreasonably but certainly incompletely, on identifying the functional or mechanistic dimensions responsible for causing and maintaining them. I shall explicate each of these in turn.

In an obvious sense, identifying depression in part by identifying 'depressed mood' seems a little circular, or at least unhelpful. And while there are, of course, attempts to characterise depressed mood in a manner that clearly distinguishes it from at least some other emotional experiences, even the most comprehensive attempts to articulate it admit that the experience varies wildly. For instance, Oyebode writes,

> The mood varies from indifference and apathy to profound despondency, dejection, despondency and despair. (2015: 270)

Even on the basis of these descriptions of the character of depressed mood (which are still relatively thin), it is unclear what theoretical reasons we have to consider all

of the possibilities along this spectrum of feeling, from indifference to despair, to be unified. Perhaps we have pragmatic reasons, in that there might be good reason to think that all of them can be a central dimension (perhaps *the* central emotional dimension) of the same disorder, perhaps on the basis of the co-occurrence of other symptoms. I shall not attempt to evaluate this suggestion here. The central point here though is that a central diagnostic construct of major depression is, phenomenologically speaking, both *imprecise* and *thin*; imprecise in that the experience being described is more naturally understood as a range of at best faintly related subjective experiences, and thin in that the structure and content of these subjective experiences is not explained in any detail. Clearly, as far as diagnosis goes (rightly or wrongly), psychiatrists do not feel as if they require a proper description of a central depressive experience to do their jobs.

Further, traditional cognitive science pays little attention to phenomenology in its explanations of depression, or depressive symptoms. Studies will typically attempt instead to identify (or intervene on) mechanisms or sub-functions responsible for producing or maintaining certain depressive symptoms with complex subjective dimensions, such as anhedonia or depressed mood. In doing so, they will utilise a concept of these symptoms that either does not contain any significant subjective dimension as an *explanans*, or else only a thin and/or imprecise one (see, e.g. Nolen-Hoeksema & Morrow 1993; Ehret, Joormann & Berking 2018 on depressed mood; Pizzagilli 2014; Cooper, Arulpragasam & Treadway 2018 on anhedonia). For instance, depressed mood might be taken to be a construct which we can clearly identify both the presence and diminution of by using certain kinds of survey techniques, and hence one might think that studies seeking to intervene on it need not concern themselves with the details of what depressed mood is actually like (e.g. Ehret, Joormann & Berking 2018).

As another example, studies might run together distinct dimensions of anhedonia (for instance, the diminishments in pleasure, motivation, and interest), thus rendering the notion they are working with less phenomenologically precise than it might otherwise have been (Cooper, Arulpragasam & Treadway 2018). Not paying due attention to these distinctions reflects the phenomenal irrelevance assumption in

at least two ways; firstly because things like feeling pleasure, motivation, and interest are *prima facie* distinct (and so running them together is a failure of phenomenological precision); and secondly because the models generated on the basis of such a wide understanding have (and can have) no clear phenomenological *explanans*. Given that these dimensions of anhehdonia are phenomenologically distinct, they are at best features that can be collectively explained *functionally* or *mechanistically*.

That concludes my discussion of traditional assumptions in the study and understanding of depression. Next, I will move on to different ways in which one might challenge the kinds of traditional assumptions discussed above, and how this thesis will use these ideas to investigate depression in a novel way.

1.4. Challenges to Tradition

In this section I will give an overview of some central concepts from various non-traditional paradigms in cognitive science that will come up in the following papers. I will also give an indication of which papers will be approaching their topics through each particular non-traditional lens.

1.4.1. Challenges to Strong Representationalism and Internalism

Challenges to strong representationalism and internalism have often, though not always[3] come together in the literature (see, for instance, Hutto & Myin 2017; 2012; Chemero 2009) and, moreover, have typically emerged from *embodied*, *enactive*, and *embedded/situated* cognition research programmes. Some have gone so far as to reject the notion of mental representation altogether (e.g. Hutto & Myin 2017; 2012), whereas others have argued that its explanatory importance is not always as great as has traditionally been assumed (e.g. Clark & Toribio 1994). While the following papers (particularly chapters 2 and 4) will make use of explanatory frameworks that challenge the centrality of representation (especially non-perceptual representation) in cognitive scientific explanation, the focus will not be on giving explicitly anti-representationalist accounts of depressive symptoms. Rather, explanations of

[3] See, e.g. Rowlands 2009 for an account that accepts the explanatory indispensability of representation, but rejects internalism.

depressive symptoms that do not explicitly make use of (non-perceptual) representations will be evaluated and tested alongside their more heavily representationalist competitors (see chapter 4, on Agential Pathology). Moreover, the notion of representation will not be invoked where I do not believe it to be necessary, leaving representationalists and anti-representationalists free to dispute the significance of the ensuing results for their own projects (see chapter 2, on Anhedonia). Without being explicitly *anti-representationalist* then, parts of this thesis will make room for anti-representationalist views to gain a foothold when it comes to explaining and understanding (some) depressive symptoms. Moreover, chapter 2 will explicitly argue that anhedonia, as currently conceptualised by psychologists, involves a suppressed, but nonetheless significant, rejection of internalism. And chapter 4 will suggest that features of embodiment play a significant, though incomplete, role in explaining agential pathology, meaning that the phenomenon cannot be properly understood from a strictly internalist perspective.

One central notion that will be deployed in what follows is that of cognitive or affective *offloading* by way of external *scaffolds*. The basic idea is simple enough – there exist artefacts, other persons, and practices which, through our interactions with them, can enhance or even create novel cognitive or affective capacities (Sterelny 2010 ; Griffiths & Scarantino 2009; Colombetti & Krueger 2014). When this occurs, it is helpful to think of the cognitive resources required to realise some capacity as being partially provided by features of the external environment. For instance, binoculars enhance our visual capacities (Sterelny 2010), spatial organisation of items can enhance our capacity to solve complex memory-involving tasks with them (Clark 2008), and music enhances our capacity for emotional regulation (Colombetti & Krueger 2014). This capacity enhancement is often achieved by using interaction with features of the environment to reduce the demands that certain tasks place on limited internal resources. As Shapiro writes,

> Using the environment, the organism will be able to reduce or simplify abstract and difficult cognitive tasks to more primitive tasks involving perception and action. (2011: 63)

Scaffolds further help us to understand the notion of a cognitive or affective *niche*. Niches, in this context, refer to collections of artefacts and practices that are actively developed and retained by organisms across multiple time-scales for the purpose of reliably scaffolding our capacities (Colombetti & Krueger 2014). Clark offers an example,

> … the novice bartender inherits an array of differently shaped glassware and cocktail furniture and a practice of serving different drinks in different kinds of glasses. As a result, expert bartenders (see Beach 1988) learn to line up differently shaped glasses in a spatial sequence corresponding to the temporal sequence of drink orders. The problem of remembering which drink to prepare next is thus transformed…The bartender, by creating persisting spatially arrayed stand-ins for the drink orders, actively structures the local environment to press more utility from basic modes of visually cued action and recall. (2008: 62)

All of this suggests that scaffolded capacities cannot be explained without making reference to processes that extend beyond the boundaries of the brain. Moreover, if capacities that are diminished in depression are amongst those that are significantly externally scaffolded, then this diminishment cannot be understood, in the general case, by solely internal factors either. This is the basic form of the argument put forward in chapter 2 regarding anhedonia. I shall suggest that the basic explanations of anhedonia put forward by psychological researchers already contain an implicit rejection of internalism, which will cause difficulties with integrating this research programme with a similar one being pursued by neuroscientists.

Furthermore, since it is not natural to think of interactions with scaffolds as constituting representations, this view of the mind raises a challenge to the centrality of representations in cognitive science as well. Though I do not explore this idea fully in chapter 2, the argument presented there invites further work on the extent of the role (if any) that mental representation plays in psychological explanations of anhedonia.

Another notion that will prove important in what follows is that of *embodiment*. Various authors argue that the body, bodily feelings, and actions make indispensable contributions to cognition in ways that belie the internalist assumption of traditional cognitive science. Depending somewhat on one's further assumptions, in particular regarding bodily processes' ability to genuinely *represent* states of affairs, embodiment usually also challenges the centrality of representation in cognitive science (contrast, for instance, Wilson 2004 and Wheeler 2005). The key idea here is that our cognitive capacities are critically dependent in various ways on the state and ongoing processes of not just the brain, but also the whole, active, body (Wheeler 2005; Shapiro 2011). This idea plays out in this dissertation primarily in chapter 4, where I introduce a proposed explanation of agential pathology in depression that puts disturbances to patterns of bodily feeling, rather than abstract mental states (such as beliefs, desires, or intentions) at the centre. Though I end up being sceptical of the idea that such a hypothesis can explain agential pathology by itself, I conclude that such ideas are an indispensable *part* of any viable explanation of agential pathology.

One final important idea to mention here is that of various kinds of 'active' perceptual representation. Much of cognitive science thinks of perception and perceptual representations more generally as merely *informative* (rather than *guiding*) and passive (rather than being directly involved in action). The basic idea here is that perception itself tells us only how the world is, rather than how to act in it, and plays only a relatively minor role in action more generally (see e.g. Marr 1982; Kitcher 1988). It does not, for instance, directly generate action, or even immediately inform us about the action-relevant properties of our environment (but rather just shape and colour, from which the presence of these properties is subsequently inferred). Both of these things are thought to be the sole preserve of abstract, post-perceptual representational states such as beliefs and intentions (which are of course generated, in part, from informative perceptual representations). This underwrites the importance of non-perceptual representation in traditional cognitive science (i.e. strong representationalism); if perceptual representations are of the kind that

traditional cognitive science suggests, they have rather limited explanatory power by themselves.

An alternative view is that perception somehow directly guides action, or even generates it (Siegel 2014). That is, that perceptual representations can directly play the role traditionally ascribed to more abstract, detached, and inferred representation. There are various ways this thought might go. One might think that perception continues to be merely informative, but that it is directly informative of features of the world relevant to action. For instance, one might think that perception itself at least sometimes informs us of the presence of action properties, for instance, that a ladder is climbable, or that a ball is throwable or catchable (Nanay 2012). Or one might go further, thinking that perceptual representations can directly invite or demand actions of us (Siegel 2014; Watzl 2014), or even that they can sometimes contain content that *urges* us to act in particular ways (Siegel 2014). The significance of this range of possibilities will be explored in chapter 4 in relation to what I will term *perceptual explanations* of agential pathology.

This concludes my discussion of challenges to strong representationalism and internalism that are relevant to understanding this thesis. I shall now move on to attempts to challenge the assumption of phenomenal irrelevance.

1.4.2. The Renewed Significance of Phenomenology

There are two ways in which the assumption of phenomenal irrelevance is challenged in the approaches taken by the papers in this thesis. The first is *direct*, in the sense that chapter 5 engages with and evaluates a proposal made about depression on the basis of an explicitly phenomenological study (Ratcliffe 2014). The second is *indirect*, in the sense that phenomenological features of depression highlighted in the aforementioned study (and others) are used as constraints in the more explicitly explanatory project of chapter 4. I shall now give an overview of each way in which phenomenological approaches to cognitive science will be used in the service of this thesis.

First, the direct challenge. Detailed descriptions of what it is like to deploy certain cognitive capacities, as well as puzzles surrounding such descriptions, are now

increasingly being used to support novel theories of cognitive capacities including vision (O'Regan & Noe 2001; Ward 2012) and social cognition (Gallagher 2012; De Jaegher, Di Paolo & Gallagher 2010). Further, Matthew Ratcliffe has recently produced an extended book-length study of depression (2015), replete with insight, that relies upon constructing and analysing rich phenomenological descriptions of various aspects of the disorder. In chapter 5 I take a closer look at one claim that emerges from Ratcliffe's approach; that, unlike other kinds of experiential similarity, depression does not enhance one's capacity to empathise with other depressed people. While I ultimately reject this claim, I take it seriously, and use it to continue an investigation into the nature and cognitive basis of social dislocation in depression, which affirms the significance of Ratcliffe's analysis.

Secondly, the indirect challenge. As Michael Roberts has recently pointed out particularly well (2017), there has been a consistent current of discontent regarding the lack of phenomenological constraint on hypothesis formation and evaluation in traditional cognitive science. This is particularly true of cognitive scientists who are, in one way or another, of a non-traditional stripe (see e.g. Ward 2012: 734; Thompson & Cosmelli 2011: 165). Roberts argue that this discontent has led various advocates of non-traditional cognitive science to endorse the idea that explanations of cognitive phenomena ought not to be "phenomenologically off key", to borrow a term from McDowell (1994: 191), or more positively that processes invoked to explain the exercise of some capacity ought to mirror the phenomenology in some way (Roberts 2017: 386). More specifically, he believes that they tacitly endorse something like the following explanatory constraint,

> (SRC) An explanation of some conscious mental process X…should reveal a strong *structural resemblance* between (a) the combined constituent parts and relations that it invokes and (b) X as it is best characterised phenomenologically. (Roberts 2017: 380)

He suggests that this constraint provides several theoretical benefits for explanations that respect it, crucially including increasing "the intelligibility of making identity claims…" (Roberts 2017: 381). While I make no specific attempt in chapter 4 to *identify* agential pathology with any particular set of processes, or even an absence of

such processes, I do endorse the basic idea that good explanations of a phenomenon must bear a strong structural resemblance to our best phenomenological characterisation of it. I thus propose three specific phenomenological constraints on explanations of agential pathology, and use them to evaluate the success of both extant and novel explanatory strategies in the literature.

That concludes discussion of the challenges to phenomenal irrelevance that provide important background for this thesis. Before providing an overview of the details of the papers to follow, I will briefly introduce two further ideas. Firstly, I will defend the view that a Predictive Processing framework for understanding cognition offers a helpful way to explain (mechanically) how some of the alternatives to traditional cognitive science discussed above might operate. Secondly, I shall argue for the importance of the discussions in this chapter so far for understanding the appropriate standards for ethical psychiatric practice.

1.4.3. Predictive Processing

In the above discussion, we came to the conclusion that the assumptions of internalism and strong representationalism that pervade traditional cognitive science have been widely and influentially challenged. Nevertheless, we did not investigate how those challenges might exactly be cashed out, in the following sense; since the brain clearly has a crucial role to play in the production of cognition, what sort of processes could it be engaged in such that its operation produces cognition that is externalist and somehow minimally (or at least non-traditionally) representational?

I believe that Predictive Processing, specifically of the kind championed by Andy Clark (2016; 2013) offers us an answer to this question. According to this view, the basic function of the brain is twofold; 1) predict what the incoming pattern of exteroceptive and interoceptive sensory information will be, and 2) hypothesise an explanation of the distal cause of this particular pattern of sensory information. The idea is that the brain produces a model of the state of the external environment (its 'best guess', as it were) and uses this to predict what the incoming sensory information will be. When its predictions are wrong, an error signal (prediction error) is produced that either drives an update to the model to bring it more in line with external reality (passive inference), or drives action to bring external reality

more in line with the existing model (active inference) (Clark 2013; Friston 2011). In either case, the result, when all is going well, is less prediction error.

It is notable that a natural way of reading all this talk of models, predictions and update is in a heavily representational way. And it is admittedly not obvious how this theory might get away from the internalism discussed above. Indeed, Jakob Hohwy has explicitly argued that Predictive Processing as a research program necessitates a return to a fairly strong internalism in Cognitive Science. He writes that Predictive Processing,

> tells us how neurocentric we should be: the mind begins where sensory input is delivered…and ends where proprioceptive predictions are delivered… (Hohwy 2016: 18)

Since the delivery of proprioceptive predictions is the mechanism by which action is produced, according to Predictive Processing, Hohwy's claim is intended as simply a more precise variant of Shapiro's characterisation of the traditional internalist assumption; cognition is once again to be found "…nestled between the peripheral shells of sensory organs and motor systems…(2011: 28). Moreover, Hohwy further comments that Predictive Processing suggests the indispensability of mental reconstructions of the distal causes of sensory stimulation (i.e. post-perceptual mental representations). He writes, as an example, that,

> An agent can grasp and use her phone only because she has a more or less precise and accurate *internal representation* of the phone, the things in her drawer that occlude it, and the causal interactions between her fingers, eyes, voice and the states of the phone. (2016: 11, *emphasis mine*)

So, in the sense that it preserves internalism and strong representationalism at least, Hohwy thinks that Predictive Processing stands with tradition, not against it. I will not rehearse his arguments for this view here, since that would take us too far off-track, but I will note that it coincides with what I think is most people's naïve understanding of Predictive Processing on first description.

It would be fair to say, however, that Predictive Processing's consequences for these elements of traditional cognitive science are in serious dispute. For instance, Clark writes that Predictive Processing,

> ….does not imply the richly reconstructive model of perception according to which our actions are selected by processes of reasoning defined over the contents of rich inner models whose role is to *replace* the external world with a kind of inner simulacrum. (2017: 736)

Moreover, he reasons,

> …inference, as it functions in the PP/PEM [Predictive Processing] story, is not necessarily defined over internal states that bear richly reconstructive, or symbolic, or propositional contents…Hohwy frequently speaks of neuronal systems as seeking out the *hypotheses* that best explain the sensory information. But it would be more accurate to describe prediction error minimization as a process that finds the multilevel set of neuronal states that best *accommodate*…the current sensory barrage…[This] may take many forms, some of which involve low-cost methods of selecting actions that re-shape the sensory signal or maintain it…Accommodating the incoming signal thus need not…imply settling upon an action-neutral description of the external situation, nor need it imply finding a proposition…that best describes or predicts that incoming signals [*sic*] (734)

The crucial point here is that the kind of error minimisation that is at the heart of Predictive Processing is readily interpretable as involving at most lightweight, action-oriented, perceptual representations, of the kind that traditional cognitive science tends to avoid. Moreover, the error minimising strategies employed by the brain according to the Predictive Processing story may depend (or, to use a term of art, *be scaffolded by*) the presence of a richly and precisely structured external environment. That is, accommodating prediction error can be a matter of exploiting the pre-existing structure of the external environment, and generating actions to subtly or greatly manipulate it, in ways that do not require detailed inner models of that environment. Instead,

> PP/PEM strongly suggests that brains like ours will, wherever possible, exploit simple strategies that rely heavily on world-engaging action, delivering new sensory stimulations just-in-time to support behavioural success. (Clark 2017: 736).

Thus, to my mind, there is a body of evidence, which is admittedly inconclusive, which suggests that Predictive Processing constitutes a challenge to tradition in cognitive science. Thus it can legitimately be used as a framework for exploring the kinds of alternative explanations that are the focus of this thesis.

In this thesis, Predictive Processing's role will, however, be somewhat limited. In particular, it will be deployed in Chapter 3 to get clear on how motivational mental states perform their characteristic functions, improving on a classical understanding of such states in the process. This is relevant to the thesis in that motivation is a construct commonly invoked to account for certain symptoms of depression, and clarification of its functions and character will play a crucial supportive role in the conclusions of Chapter 4.

1.4.4. Psychiatric Ethics

When people present to mental health services, and over the whole course of receiving care, they will typically interact with people who endorse a more-or-less traditional view of mental illness, as described above. This tends to come packaged with a *technical* view of the appropriate interventions – the important features of interventions (pharmacological, or therapeutic) are their technical details, or the particular neurological or cognitive systems they are thought to intervene upon. SSRI antidepressants, for instance, are thought to be effective (when they are) because they increase the concentration of the neurotransmitter serotonin in the synapse (Nutt et al 1999). Cognitive Behavioural Therapy is thought to be effective (when it is) because it disrupts the influence of false and negative beliefs about oneself, others, and the wider world on subsequent feeling and action (Beck et al 1979). Accurate diagnosis then becomes a central concern, since particular mental illnesses are thought to necessitate different technical interventions. As a result, the service user need, in theory, play little more than a facilitative role in the whole process of diagnosis and treatment, accurately reporting their symptoms to the medical

professionals in charge of their care, who may then prescribe appropriate treatment. As long as the service user complies with such treatment, it is generally thought that they should experience some improvement (which is identified with the remission of symptoms). In some important sense, a mental health service user is treated as an object of inquiry for the medical professional; they are not an equal participant in the discussion.

If the traditional view of depression as a relatively straightforward kind of brain disease were correct, there would be a case to make that such a process of diagnosis and technical intervention is wholly appropriate. Despite the connotations of objectification, it is not obvious that I need always object to being an object of an expert's inquiry. If I have a chest infection, I have little inclination to participate in the search for diagnosis and treatment as an equal partner with the medical professional beyond their acknowledging the symptoms I describe. I am perfectly content with them using their expertise to work out what kind of infection I have and prescribing an antibiotic that will make it go away. My input is of secondary importance at best.

But if a non-traditional view is correct; if mental illnesses are the result of complex interactions between brain operation, bodies, and social and natural environments, with no clear basis either in neurology or individual thought, then one might think that this approach is misguided, because the scope of the proposed interventions appears far too narrow, in exactly the way that the traditional view of cognitive science is too narrow. And if the precise phenomenology of these conditions is critical to understanding them, then the lack of engagement with service users' personal descriptions and perspectives on their experiences is a serious oversight at best, and actively callous at worst.

Indeed, one might think that if a non-traditional view on depression is correct, then service users' side-lining amounts not to justifiable medical practice, but to a particular kind of injustice – *epistemic injustice*. Epistemic injustice occurs when somebody is unjustly harmed in a manner closely connected to their capacity and/or status as a knower (Fricker 2007). This can include having their credibility deemed unfairly low on some topic (*testimonial injustice*), being deprived of epistemic

resources for understanding their lives (*hermeneutical injustice*) (Fricker 2007), having the epistemic resources they have developed for understanding their own lives rejected by others (*contributory injustice*) (Dotson 2012), or having perverse social incentives to refrain from sharing information about their experiences (*testimonial smothering*) (Dotson 2011), amongst other ways. While some work has analysed the pervasiveness and harm of testimonial injustice in formal Psychiatry (Crichton, Carel & Kidd 2016; Kurs & Grinshpoon 2017), the others have largely been overlooked. In Chapter 6 I try to go some way to correct for this oversight, by tracing these three kinds of epistemic injustice through service user accounts of treatment in formal psychiatric services, and proposing how they might be minimised. I conclude that not only are these forms of injustice likely to be rife in psychiatric services, but that many of the reasons such injustices persist is in part due to the overly narrow, mechanical view of mental distress that much of the rest of this thesis is concerned with interrogating.

1.5. Paper Summary

This thesis will proceed as follows. In Chapter 2 I will tackle the topic of anhedonia. I will argue that both existing psychological theories of the phenomenon (which connect it to the diminishment of two different psychological capacities) cannot be thought of as high-level descriptions of neurobiological processes alone. Rather, whichever theory you adopt, anhedonia is better thought of as a kind of breakdown in an individual's affective niche; a collection of environmental objects, features, and opportunities that support various kinds of affective regulation for that person. In Chapter 3 I will focus on developing a predictive processing account of motivation; while this is not directly tied to depression *per se*, it will help us get clear on the nature of motivation, which will be important in Chapter 4. In that chapter, I will investigate the nature of Agential Pathology; disturbances to the ability to act that are characteristic of depressive illness. While often thought of as a disorder of motivation emerging from a single or relatively small number of sources, I will suggest that such an explanation is incomplete; Agential Pathology is a highly complex, multiply realised phenomenon that does not only involve disturbances to motivation, properly understood. In Chapter 5 I will evaluate a claim of Matthew

Ratcliffe's to the effect that people with depression are not better able to empathise with other depressed people than the general population. I evaluate his argument for this claim, based on a non-traditional, phenomenologically informed view of empathy. Though I find the evidence Ratcliffe provides for this claim wanting, I suggest that a more traditional view of empathy can be leveraged to provide some *prima facie* support for it. This investigation reveals significant complexities in the notion of an empathic deficit in depression; there are many senses in which empathy can fail, and correlatively many ways depression could interfere with it. Finally in Chapter 6 I will reflect on the ethical implications of classical assumptions on psychiatric care. In particular I will argue that overly-individualised and pathologized understandings of mental illness, including depression, are the root of psychiatric service users' vulnerability to a variety of epistemic injustices.

These papers, together, constitute a systematic attempt to understand the significance of non-classical cognitive science for understanding a variety of features of depression and mental illness more generally. Though the specific conclusions are mixed, they do vindicate the significance of further research into the psychiatric applications of non-classical cognitive science, both in terms of psychological explanation and the ethics of treatment.

1.6. References

Bechtel, W. (2009). "Explanation: Mechanism, modularity, and Situated Cognition". In Robbins, P. & Aydede, M. (eds.). *Cambridge Handbook of Situated Cognition*. Cambridge: CUP

Beck, A.T. (1979). "Cognitive Therapy and the emotional disorders". New York: Penguin

Beck, A.T. Rush, A.J. Shaw, B.F. Emery, G. (1979). "Cognitive Therapy of depression". New York: The Guilford Press

Bracken P, Thomas P, Timimi S et al. (2012), "Psychiatry beyond the current paradigm", *The British Journal of Psychiatry*. 201(6): 430-434.

Bruce, V. & Young, A. (1986). "Understanding face recognition". *British Journal of Psychology 77*(3): 305-327

Chemero, A. (2009). "Radical embodied cognitive science". Cambridge MA: MIT Press

Clark, A. (2008). "Supersizing the mind: Embodiment, action, and cognitive extension". Oxford: OUP

Clark, A. (2013). "Whatever next? Predictive brains, situated agents, and the future of cognitive science", *Behavioral and Brain Sciences 36*(3): 181-204

Clark, A. (2017). "Busting out: Predictive brains, embodied minds, and the puzzle of the evidentiary veil". *Nous 51*(4): 727-753

Clark, A. & Toribio, J. (1994). "Doing without representing?". *Synthese 101*(3): 401-431

Colombetti, G. (2013). "The feeling body: Affective science meets the enactive mind". Cambridge MA: MIT Press

Colombetti, G. & Krueger, J. (2015), "Scaffoldings of the affective mind", *Philosophical Psychology* (**28: 8**), pp.1157-1176

Cooper, J.A. Arulpragasam, A.R. & Treadway, M.T. (2018). "Anhedonia in depression: Biological mechanisms and computational models". *Current Opinion in Behavioral Sciences 22*: 128-135

de Jaegher, H. di Paolo, E. & Gallagher, S. (2010). "Can social interaction constitute social cognition?", *Trends in Cognitive Sciences 14*(10): 441-447

Crichton P, Carel H, Kidd I.J. (2017), "Epistemic injustice in psychiatry", *Psychiatric Bulletin*. 41(2): 65-70.

Dotson K. (2011), "Tracking epistemic violence, tracking practices of silencing", *Hypatia*. 26(2): 236-257

Dotson K. (2012), "A cautionary tale: on limiting epistemic oppression", *Frontiers: A journal of women's studies 33*(1): 24-47.

Dozois, D.J.A. & Dobson, K.S. (2001). "Information processing and cognitive organization in unipolar depression: Specificity and comorbidity issues". *Journal of Abnormal Psychology 110*(2): 236-246

Drayson, Z. (2009). "Embodied cognitive science and its implications for psychopathology". *Philosophy, Psychiatry, & Psychology 16*(4): 329-339

Drayson, Z. (2012). "The uses and abuses of the personal/subpersonal distinction". *Philosophical Perspectives 26*: 1-18

Ehret, A.M. Joormann, J. & Berking, M. (2018). "Self-compassion is more effective than acceptance and reappraisal in decreasing depressed mood in currently and formerly depressed individuals". *Journal of Affective Disorders 226*: 220-226

Fricker M. (2007), *Epistemic injustice: power and the ethics of knowing*. Oxford: OUP

Friston, K.J. (2011). "What is optimal about motor control?". *Neuron 72*(3): 488-498

Gallagher, S. (2012). "In defense of phenomenological approaches to Social Cognition: Interacting with the critics". *Review of Philosophy and Psychology 3*(2): 187-212

Gerrans, P. (2014). "The measure of madness". Cambridge MA: MIT Press

Gervais, R. & de Jong, H.L. (2012). "The status of functional explanation in Psychology: Reduction and mechanistic explanation". *Theory and Psychology 23*(2): 145-163

Griffiths, P. & Scarantino, A. (2009), "Emotions in the wild: The situated perspective on emotion", in Robbins, P. & Aydede, M. (eds.) *The Cambridge Handbook of Situated Cognition*, CUP: Cambridge

Guze, S.B. (1989). "Biological psychiatry: Is there any other kind?". *Psychological Medicine 19*(2): 315-323

Harvard Health Publishing (2017). "What causes depression?". *Harvard Medical School*. available at: https://www.health.harvard.edu/mind-and-mood/what-causes-depression [accessed 13/12/18]

Hohwy, J. (2016). "The self-evidencing brain". *Nous 50*(2): 259-285

Hutto, D.D. (2016). "A reconciliation for the future of psychiatry: Both Folk Psychology and Cognitive Science". *Frontiers in Psychiatry 7*: 25-36

Hutto, D.D. & Myin, E. (2012). "Radicalizing Enactivism". Cambridge MA: MIT Press

Hutto, D.D. & Myin, E. (2017). "Evolving Enactivism: Basic minds meet content". Cambridge MA: MIT Press

Ingram, R.E. (1984). "Toward an information-processing analysis of depression". *Cognitive Therapy and Research 8*(5): 443-478

Kitcher, P. (1988). "Marr's computational theory of vision". *Philosophy of Science 55*(1): 1-24

Kuhl, J. & Helle, P. (1986). "Motivational and volitional determinants of depression: The Degenerated-Intention hypothesis". *Journal of Abnormal Psychology 95*(3): 247-251

Kurs, R. & Grinshpoon, A. (2017), "Vulnerability of individuals with mental disorders to epistemic injustice in both clinical and social domains. *Ethics & Behaviour*. [Preprint]. available from: doi:10.1080/10508422.2017.1365302.

Levelt, W.J.M. (1989). "Speaking: From intention to articulation". Cambridge MA: MIT Press

Lilienfeld, S.O. Schwartz, S. Meca, A. Sauvigné, K. & Satel, S. (2015). "Neurocentrism: Implications for psychotherapy practice and research". *The Behavior Therapist* (October): 173-181

Marr, D. (1982). "Vision". San Francisco CA: WH Freeman

McDowell, J. (1994). "The content of perceptual experience". *The Philosophical Quarterly 44*(175): 190-205

McManus, S. Bebbington, P. Jenkins, R. Brugha, T. (eds.) (2016). Mental health and wellbeing in England: Adult psychiatric morbidity survey 2014. Leeds: NHS Digital. available at: https://webarchive.nationalarchives.gov.uk/20180328140249/http://digital.nhs.uk/catalogue/PUB21748

Murphy, D. (2006). "Psychiatry in the scientific image". Cambridge MA: MIT Press

National Health Service (2016). "Clinical depression". available at: https://www.nhs.uk/conditions/clinical-depression/ [accessed 15/12/18]

Nanay, B. (2012). "Action-oriented perception". *European Journal of Philosophy 20*(3): 430-446

Nolen-Hoeksema, S. & Morrow, J. (1993). "Effects of rumination and distraction on naturally occurring depressed mood". *Cognition & Emotion 7*(6): 561-570

Nutt, D.J. Forshall, S. Bell, C. Rich, A. Sandford, J. Nash, J. Argyropoulos, S. (1999). "Mechanisms of action of selective serotonin reuptake inhibitors in the treatment of psychiatric disorders". *European Neuropsychopharmacology 9*(3): S81-S86

O'Regan, J.K. & Noe, A. (2001). "A sensorimotor account of vision and visual consciousness". *Behavioral and Brain Sciences 24*: 939-973

Oyebode, F. (2015), *Sims' symptoms in the mind: Textbook of descriptive psychopathology (5th Edition)*, Saunders Elsevier: Oxford

Parekh, R. (2017). "What is depression?". *American Psychiatric Association*. available at: https://www.psychiatry.org/about-apa/vision-mission-values-goals [accessed 10/12/18]

Parnas, J. (2014). "Commentary: The RDoC program: Psychiatry without psyche". *World Psychiatry 13*(1): 46-47

Pizzagalli, D.A. (2014). "Depression, stress, and anhedonia: Toward a synthesis and integrated model". *Annual Review of Clinical Psychology 10*: 393-423

Ratcliffe, M. (2015). "Experiences of depression: A study in phenomenology". Oxford: OUP

Roberts, M. (2017). "Phenomenological constraints: A problem for Radical Enactivism". *Phenomenology and the Cognitive Sciences 17*(2): 375-399

Roepke, A.M. & Seligman, M.E.P. (2015). "Depression and prospection". *British Journal of Clinical Psychology 55*(1): 23-48

Rowlands, M. (2009)

Shapiro, L. (2011), *Embodied Cognition*, Routledge: Oxford

Siegel, S. (2014). "Affordances and the contents of perception". in Brogaard, B. "Does perception have content?". Oxford: OUP. 51-75

Sterelny, K. (2010), "Minds: extended or scaffolded", *Phenomenology and the Cognitive Sciences* (**9: 4**), pp.465-481

Tabb, K. (2017). "Philosophy of Psychiatry after diagnostic kinds". *Synthese* (online first). doi: https://doi.org/10.1007/s11229-017-1659-6

Thagard, P. (2014). "Cognitive Science". in Zalta, E.N. (ed.) "The Stanford Encyclopedia of Philosophy (Winter 2018 Edition)". Available at: https://plato.stanford.edu/archives/win2018/entries/cognitive-science/

Thompson, E. & Cosmelli, D. (2011). "Brain in a vat or body in a world? Brainbound versus enactive views of experience". *Philosophical Topics 39*(1): 163-180

Ward, D. (2012). "Enjoying the spread: Conscious externalism reconsidered". *Mind 121* (483): 731-751

Watzl, S. (2014). "Perceptual Guidance". *Ratio 27*(4): 414-438

Wheeler, M. (2005). "Reconstructing the cognitive world". Cambridge MA: MIT Press

Wilson, R.A. (2004). "Boundaries of the mind: The individual in the fragile sciences: Cognition". Cambridge: CUP

Chapter 2: Anhedonia and the Affectively Scaffolded Mind

2.0. Abstract

Anhedonia, roughly defined as the diminishment or absence of the capacity to experience pleasure or joy in the performance of daily activities, is a core symptom of *Major Depressive Disorder*, as well as other psychiatric illnesses. I argue that the two major psychological theories of anhedonia are committed to the view that anhedonia cannot, in the general case, be explained with reference only to neurobiological states and processes. This is despite the overwhelming explanatory focus on neurobiological factors in the existing literature. Instead, it is to be understood as the breakdown in the function of what Colombetti & Krueger (2015) term a subject's *affective niche*. Since affective niches are composed of elements of a person's natural and social environments, including artefacts, activities, and other people, anhedonia turns out to be a phenomenon deeply integrated into a subject's environment, inscrutable within the boundaries of skin and skull. I discuss and refute some objections to this view.

2.1. Introduction

Anhedonia is a core symptom of Major Depressive Disorder (MDD), amongst other psychiatric conditions (Oyebode 2015). It is broadly defined as "the loss of pleasure or interest in previously rewarding stimuli" (Heller et al 2009). It is further divided into *consummatory*, *motivational*, *anticipatory*, and *decisional* sub-types. The first refers to the loss of the subjective experience of pleasure (or positively-valenced affect) from partaking in previously enjoyable activities, the second to a loss of interest or motivation in pursuing such activities, the third to a diminished ability to anticipate future pleasure on the basis of past experiences, and the fourth to difficulties in choosing different courses of action due to indifference regarding their consequences (Mallorquí, Padrao & Rodriguez-Fornells 2014; Treadway & Zald 2011). There is significant debate over whether these four symptoms should fall under the same label (Treadway & Zald 2011), as they seem to be conceptually and diagnostically separable phenomena (Oyebode, 2015, 262 & 296). There is, however, some evidence that abnormal activity in certain brain regions[4] is associated strongly with all of them (Heller et al 2009), and that MDD patients are vulnerable to each of them (Treadway et al 2012). This notwithstanding, whenever I refer to anhedonia in what follows, I will be speaking about consummatory anhedonia. This is a wholly pragmatic decision, taken because this is the variant that is the primary target of the major psychological theories that I will be examining.

To give a clear example of what I take to be a paradigm case of anhedonia, let me introduce you to Ashley. Ashley has been a very keen marathon-runner for a large part of their adult life. After a hard day at work, Ashley usually likes nothing more than to go for a run. Not only do they enjoy the process of running, but they inevitably feel happier for an extended period after they have returned from a jog.

Recently, Ashley has been diagnosed with MDD. One of the key data points used in this diagnosis was that they no longer enjoy running at all. They no longer have any motivation to go for a run as they once did, and (crucially) even if they do force themselves to go outside for a quick jog, they do not enjoy the process and feel no

[4] Specifically, the nucleus accumbens and frontostriatal network.

better afterwards. As a result of this change and other similar ones, their doctor decided that Ashley exhibited anhedonia.

2.1.1. Biomedical Materialism

It is widely held by psychiatric researchers that MDD is fundamentally a neurobiological disorder, and that its symptoms ultimately have their physical basis in the brain, and perhaps some parts of the wider nervous system (see e.g. Lima-Ojeda et al, 2017; Wohleb et al 2016; Köhler et al 2016). As Stephan Köhler and colleagues write,

> Present theories of MDD… focus on hyperactivity of the neuroendocrine stress axis…, impaired cellular plasticity in general and impaired hippocampal neurogenesis in particular…, perturbations in neurotrophin signalling…, and neuroinflammation... (2016: 14)

If this were true, since anhedonia is a core symptom of MDD, it would imply that anhedonia has its physical basis in the brain – again a common position in the literature (see e.g. Rizvi et al 2016; Der-Avakian & Markou 2012; Treadway & Zald 2011). As Sakina Rizvi and colleagues write,

> …we assume that anhedonia can arise from impairments in various facets of reward processing… Evidence suggests that there are specific neuroanatomical areas that underlie various facets of reward processing, including the prefrontal cortex (orbitofrontal cortex, ventromedial prefrontal cortex and anterior cingulate cortex), dorsal striatum (caudate and putamen), nucleus accumbens and amygdala (2016: 23)

And further,

> …anhedonia mechanisms may be a promising area for biomarker research in MDD, *since they map onto specific and partially dissociable neurocircuitry and signalling pathways*… (2016: 32, *emphasis mine*)

This is simply a representative case of a much wider trend in psychiatric science, which Will Davies (2016), following Peter Zachar (2000) labels *biomedical materialism*. According to Davies,

> On this view, the nature of mental illness is exhausted by neurological properties and events, and all facts about such conditions are scrutable from neural facts. (2016: 291)

In addition to this neurobiological perspective, there are also *psychological* explanations of anhedonia; that is, explanations that appeal to constructs at the level of psychological, rather than neurobiological, activity. Such explanations of anhedonia generally fall into one of two broad camps. The first of these theorises that anhedonia is an outcome of diminished *hedonic capacity*, that is, the sufferer is severely limited in the maximum quantity or degree of pleasure they can experience in response to (at least some) stimuli (Meehl 1975; Fiorito & Simons 1994). Call this the *capacity theory*. More recently, it has been suggested that anhedonia may instead involve diminished ability to *sustain positive affective responses* over time (Tomarken & Keener, 1998; Heller et al 2009). Call this the *sustainability theory*.[5]

If one were to grant the truth of biomedical materialism (at least in the case of anhedonia), then one might think that when *psychological* explanations of anhedonia are offered, these hypotheses are intermediary explanatory levels on the way to the ultimate, neurobiological underpinnings of anhedonia. That is, though they may prove useful for research and clinical psychologists, accurate psychological theories of anhedonia ultimately reflect only neurobiological facts, properties or dynamics of some sort. On this sort of view, psychological theories of anhedonia are not false – they simply reflect a *fundamentally neural* process at a relatively high level of abstraction.

It is this thought that I will resist in this paper[6]. Whichever psychological account of anhedonia one chooses, the psychological abnormality posited is best understood as a breakdown of functionality in what Colombetti & Krueger (2015) call an individual's *affective niche*. Just what kind of breakdown I have in mind, and what an affective niche is will become clear in what follows. It will suffice for now to note

---

[5] Both of these theories seem to have their origin in Myerson (1922).

[6] For an investigation and sustained critique of this kind of neuro-centrism in Psychiatry more generally, see Huber & Kutschenko (2009).

that, if I am right, then both the capacity and sustainability theory suggest that anhedonia's ultimate physical basis a) is likely to be different from case-to-case and b) will in *the general case* include elements of an individual's material and social environment. Entirely neurobiologically grounded cases of anhedonia will turn out to be comparatively rare limit cases of a disorder that typically exhibits dependence on interactions between agents and their environments.

My argument will proceed in two main stages. In the first, I shall argue that the psychological capabilities hypothesised to be diminished in anhedonia by the capacity and sustainability theories are best understood as variable across contexts and significantly dependent on environmental resources available to the agent. In the second I shall argue that this suggests that the ultimate physical basis of anhedonia is neither uniform, nor generally to be found in the brain. These arguments will be made in sections 4 and 5 respectively. Neither of these claims, it should be emphasised, imply that anhedonia's physical basis is not generally *partially* to be found in the brain, nor that this partial contribution is unnecessary. Indeed, I believe that the brain plays a necessary role in grounding anhedonia, as it does all psychological phenomena. But it is not the whole story.

Before I argue for my key claims in sections 4 and 5, however, I must introduce and explain several key concepts that the arguments will depend on. In section 2 I shall introduce and explain the capacity and sustainability theories, focusing on the psychological capability each hypothesises to be diminished or otherwise disrupted in anhedonia. And in section 3 I shall introduce and explain two key ideas from the situated cognition literature – *affective scaffolds* (Griffiths & Scarantino 2009; Colombetti & Krueger 2015; Colombetti 2017) and the construction of a *cognitive/affective niche* (Sterelny 2010; Colombetti & Krueger 2015).

2.2. Two theories of anhedonia

In this section I shall explain the two main competing psychological theories of anhedonia and briefly review some of the main empirical evidence available for each. The crucial take away from this section will be that each posits a diminishment or loss of a distinct psychological capability. In the case of the capacity theory, this is the capability of experiencing a particular maximum amount of positive affect. In the

case of the sustainability theory, this is the ability to sustain positively valenced affective responses (henceforth 'positive affect') over time. These capabilities will be the targets of analysis in the rest of the paper. An easy (though rough) way to think about this distinction is that the capacity theory holds that anhedonia is about diminished affective *intensity*, and the sustainability theory that it is about affective *duration*.

2.2.1. Capacity Theory

For many years the dominant psychological theory of anhedonia, which still has many defenders, was very simple; people with anhedonia suffer from a diminished hedonic capacity (Meehl 1975; Pizzagalli et al 2008). Hedonic capacity is supposed to refer to the maximum amount of positive affect a person is capable of experiencing in their day-to-day life. According to the capacity theory, anhedonia is a matter of the cap on this maximum amount of pleasure being significantly lowered. An anhedonic person is simply unable to ever experience the amount of pleasure that others can, and consequently fails to experience the same amount of pleasure from certain activities as they used to.

Imagine a pint glass filled with water. The maximum capacity of the glass is the maximum amount of pleasure a person can experience from a given activity. The basic claim of the capacity theorists is that an anhedonic person's glass has been replaced with a thimble. Recall Ashley's case; here the capacity theory of anhedonia amounts to a claim that they are no longer able to experience the same amount of pleasure from running as they once did; the maximum cap on the pleasure they are able to experience from running (and probably, though not necessarily, other activities) has been significantly lowered.

Some data seem to support this theory. It has been found that anhedonic people report that they find positive images less positive, and that both positive and negative images are less impactful on their emotional state, than those without anhedonia (Fitzgibbons & Simon 1992; Fiorito & Simons 1994). Visceral physiological (heart rate change) and overt behavioural measures (facial expression) of emotional arousal have also been found to be diminished in anhedonic people (Fiorito & Simons 1994). That said, these data have not proven to be easily reproducible

(Germans & Kring 2000; Kaviani et al 2004). Given this, some researchers have sought support for an alternative hypothesis.

2.2.2. Sustainability Theory

Several studies suggest that the issue in anhedonia may not be one of hedonic capacity, or at least not solely (see Heller et al, 2009). The sustainability theory suggests that anhedonic individuals are capable of experiencing the same peaks of enjoyment as the rest of us. The problem is that the ability of an anhedonic individual to sustain a pleasurable or joyful response to certain stimuli is impaired (Heller et al, 2009; Tomarken & Keener, 1998). The claim here is that pleasurable reactions to stimuli are typically significantly temporally extended, but in anhedonia the length of time positive affect continues is diminished.

Imagine the glass again. According to sustainability theorists, the problem in anhedonia is not the size of the glass, but its ability to hold the water that is poured in. A healthy person's glass will hold all or nearly all of the water it can for an appropriate length of time. An anhedonic person's glass, on the other hand, is tremendously leaky. Returning to Ashley, the sustainability theorists suggest that what is going on is not as simple as it might have initially appeared. Although Ashley's ability to receive pleasure from the initial act of running may be relatively undiminished, their ability to feel good throughout the course of the run and afterwards is significantly impaired. As a result, Ashley does not experience the run as being enjoyable in the way they once did, as the experience is no longer sufficiently temporally extended.

An FMRI study by Heller and colleagues (2009) support this idea. Areas of the brain associated with reward processing and regulation of positive affect "showed a specific decrease in activation [in response to positive stimuli]… across time, while control subjects maintained their level of activation" (Heller et al 2009, 22448). Moreover "the amount of decrease in…activity across time predicted overall self-reported affect" (22448). Further, Liu and colleagues (2011) have demonstrated that anhedonia in MDD reflects an inability to sustain behaviour directed towards salient incentives over time. They take this to (broadly) support the sustainability theory,

assuming ongoing positive affect's role in sustaining reward-related behaviour over time.

So we have identified diminishments in two different psychological capabilities, which are hypothesised to characterise anhedonia – the in/ability to experience particular 'highs' of positive affect, or the in/ability to sustain positive affective responses over relatively extended time periods. These are currently, to my knowledge the main contenders. Some may think that they each specify a real, though distinct, symptom that are conflated to the single concept of (consummatory) anhedonia. Others may think that (consummatory) anhedonia is a unified phenomenon, but is instead characterised to different degrees by *both* of the diminished capabilities hypothesised. These nuances are not important in what follows. The goal in section 4 will (roughly) simply be to argue that, whichever diminished psychological capability (or capabilities) one takes to characterise anhedonia, it turns out to depend significantly on particular kinds of environmental resources.

Of course, it is possible that a new theory will emerge that specifies that a novel psychological capability, call it *c*, characterises anhedonia instead. If *c* turned out *not* to be interestingly dependent on environmental resources, then this theory would be a refuge for those who wanted to insist on a neuro-centric view of anhedonia, though of course properly 'joining up' the neurobiological and psychological levels of explanation in a suitably comprehensive manner would still involve numerous problems. That said, I do not think that it is beholden on me to argue on behalf of imaginary theories. Readers concerned by this possibility may think of my eventual claim (in section 5), that anhedonia is not entirely neurobiologically grounded, as being implicitly conditionalised – it holds if either (or both) of the currently popular psychological theories of anhedonia are onto something substantial.

2.3. Situating Affect & Affective Niche Construction

One of the unifying claims of the situated emotion research literature is that natural, social, and cultural resources in agents' environments (whether intentionally made available by the agent themselves or otherwise) is indispensable for the development and realisation of particular affective capabilities. Call resources that exhibit this

quality *affective scaffolds* (Griffiths & Scarantino 2009; Colombetti & Krueger 2015). Another common claim is that there is often a relationship of *reciprocal* influence between agents' affective capabilities and affective scaffolds. That is, a particular resource may enhance an agent's (or many agents') affective capabilities in such a way that it leads to the development or availability of new scaffolds, which in turn may further enhance some aspect of an agent's affective life. Call contextually or universally available collections of affective scaffolds to which we become accustomed (and that which may often emerge in their entirety as a result of the aforementioned kind of reciprocal influence) *affective niches* (Sterelny 2010; Colombetti 2017). It is to fleshing out these concepts that I turn in this section.

2.3.1. Functional Gain and Reciprocity

We can intuitively distinguish between those conditions or resources that provide necessary background to the execution of some function, and those that are actively recruited into the process of executing that function; those that merely enable some capacity and those that genuinely *enhance* its realisation in some way. The presence of oxygen falls into the first camp with respect to most cognitive capacities. There is no obvious sense in which the kind of potential impediments to living that cognition exists to overcome need ever involve recruiting oxygen as a resource directly, though certainly no cognitive problems can be usefully resolved if the brain is starved of oxygen. Another example; I cannot win a game of chess if the sun explodes shortly before I start playing (my opponent's and my being dead, along with the lack of non-atomised chess boards being great impediments). Nevertheless, we rightly do not speak of a non-exploded as being a resource that I recruited, or made use of, in order to win. It is not this kind of trivial dependence of cognitive capabilities on external resources that I am interested in for the purposes of this paper.

I am interested in cases where the execution of certain cognitive functions seem to actively make use of resources, information, and structure external to the organism. For instance, skilled bag packers in grocery stores in the USA often arrange items spatially by category (heavy, light, fragile, etc) as they come off the conveyor. This is to subsequently facilitate optimal packing arrangements, while not placing an unbearable load on working memory. One need not remember how heavy each

individual object was if all the heavy and light objects have been placed in distinct and mentally labelled regions of space (Kirsh, 1995, cited in Robbins & Aydede, 2009: 6). Here, we see external spatial resources being used to lessen the need for substantial internal memory resources to be used in the execution of the task. Moreover, in the case of there being some very large number of items, it is sensible to suppose that optimal bag-packing could not be accomplished *without* recruiting the relevant spatial resources (or some functional analogue that goes beyond bio-memory). The fact that the space exists, and humans are able to interact with it in a particular way (and do), produces a functionally significant improvement in a particular cognitive ability. That is, interaction with the environment in particular ways both *enhances* and *extends* cognitive capacity.

Theorists in the situated cognition camp argue that these kinds of cases are the rule rather than the exceptions; most important cognitive tasks are resolved at least in part thanks to the recruitment of pre-existing environmental structure, or the active construction of it (both are present in the bag-packing case). This is similar to what Lawrence Shapiro labels the *Principle of Ecological Assembly*:

> [p]roblem solving… is a function of the resources an organism has available to it [or can make available to it] in its surrounding environ. (2011, 63, [additions mine])

It is crucial to note however that, on the situated picture of cognition, not only problem solving, but the capacities that enable it, are functions "of the resources an organism has available to it [or can make available to it]" (2011, 63, [additions mine]). The capacity of the bag-packers to perform their task is enhanced by the availability and exploitation of surrounding spatial resources. And many of our most core cognitive capacities are enhanced by such externally located resources (much of the time ones that we have developed or arranged ourselves for precisely that purpose). We enhance our perceptual capabilities with binoculars, our mathematical reasoning capabilities with calculators, our navigational capabilities with maps, and so forth (Sterelny 2010). To give the phenomenon a name, some external resources *scaffold functional gain* in our cognitive abilities – they grant us particular cognitive abilities that we would not otherwise possess, or deliver significant enhancement, such that

our cognitive capacities would be seriously impoverished without them. Call such items *cognitive scaffolds*. Some cognitive scaffolds, like in the examples I have given, scaffold our cognitive capacities in particular instances, over the relatively short-term. Others, such as written language, plausibly enhance our cognitive abilities across the course of our development and indeed across generations (Wilson & Clark 2009). Call the latter kind *diachronic cognitive scaffolds* and the former *synchronic cognitive scaffolds*[7]. Moreover, while some cognitive scaffolds are common to us all (or almost all of us at least), some are highly specific to individuals. And we actively shape our environments and the resources within it so as to have particular kinds of cognitive scaffolds available when they are needed, both generic and highly specific. Many of us carry binoculars with us when we know we may have to see great distances and mark maps with lines and scribbles that serve our own, unique navigational needs.

External resources scaffold functional gain in our affective abilities as well (in the discussion that follows I draw predominantly from Griffiths & Scarantino 2009 and Colombetti & Krueger 2015). Maintaining a reasonable mood is a capability many of us enjoy, but our strategies for achieving it are often dependent on the presence of certain kinds of external resources. Examples include cigarettes, music players, gentle lighting, scented candles, other people, or vacations. Different people exhibit different patterns of individuality, reliance, and trust with respect to the affective scaffolds they make use of. Our music players for instance tend to be highly individualised artefacts, representing our particular taste in music, with particular playlists often deliberately designed to achieve different affective ends *for us* (music that 'pumps us up' for exercise, music that relaxes us, music that makes us happy, music that helps us cry, and so forth). We may not *rely* on a yearly holiday for a positive temperament nearly as much as we do a weekly drink with our friends; the loss of the latter may put us entirely out of sorts whereas we can more-or-less cope

[7] Here, I shall make my case via examples of synchronic scaffolding. This is not because I think affective capabilities are in general only synchronically scaffolded, or that anhedonia has nothing to do with diachronic scaffolding. Both of these claims are, in my view, almost certainly false. But such discussion will complicate matters, and thus impede rather than enhance understanding of the basic idea being presented.

without the former. And we may trust our cigarettes to perk us up considerably more than a nice healthy walk in a park (or vice versa). The upshot is that the presence of any of these things may enhance our ability to exhibit, regulate or maintain particular affective states, and their absence may significantly impede these abilities. This enhancement may occur over many timescales; social norms regarding suitable emotional displays at funerals or birthday parties may significantly affect the development of our situation-sensitive emotional repertoires, and constrain immediate emotional range where necessary. And the specific collection of emotional concepts and words that we possess – due to both individual developmentally significant experiences and cultural influence - may affect our actual emotional experiences in-the-moment, by affecting how we view, categorise, and interpret them, as well as how they are used to guide behaviour (Barrett 2012; 2014).

Crucially, affective scaffolds cannot be reduced to *elicitors* of emotion or emotional *stimuli*. The first reason is discussed above. Certain of our affective abilities may be impaired without momentary or sustained support from an affective scaffold. They are *active contributors* to the qualities of our emotional life, rather than simply triggers of it. We would not possess many of the affective abilities we do were it not for the presence of a suitable collection of scaffolds at appropriate times. The second reason is that the contribution is not unidirectional; we shape and create affective scaffolds just as much as affective scaffolds shape and create our affective capacities, and often *in virtue* of their previous contributions. Some of us deliberately construct the contents of our music players to reflect a range of affective needs we may have, and carry them with us regularly enough that they are reliably available when those needs arrive. Others of us design our ideal homes, and decorate particular rooms so that they have relaxing or energising qualities. Sometimes the contribution of one affective scaffold may help us to recruit the contribution of another; for instance, I may listen to calming music when I am very anxious about an upcoming exam, precisely so that I am able to call my friend and discuss the problem with them. Had I not listened to the music first, the friendly discussion route may well not have been open to me, because my high levels of anxiety could have prevented me from

communicating sufficiently clearly, or plucking up the courage to even pick up the phone.

2.3.2. Constructing and Deconstructing Affective Niches

The upshot of all of the above is that people are constantly engaged in an active ongoing relationship with the world, and the individual, and often unique, contributions of their brains, bodies, and environment to this relationship *collectively* result in certain affective capabilities that would be unachievable without one or another of the components. Given humans' propensity to actively shape and choose (to some degree) the worldly dimension of this distributed affective network, we may say that humans occupy partially generic and partially individualised affective niches for much of their lives. That is to say, we build and select affectively relevant artefacts, people, and locations *for the sake of* their abilities to make helpful contributions to our affective lives. These niches are generic insofar as certain social norms, cultural understandings, and architectural designs, etc (all of which may count as affective scaffolds) are shared amongst all, or nearly all, people who occupy a particular country, community, or city. But they are also individualised to the degree that individuals make choices and take actions regarding the kinds of affective scaffolds they make use of, or carry with them, on a day-to-day basis. Naturally, we occupy many niches throughout our lives, and some scaffolds are more-or-less permanent fixtures in all of them, for a variety of reasons. And when we do occupy different niches, which enable correspondingly different affective abilities, everything from our comportment, to our manner of speaking, to our capacities for emotional regulation may change (Colombetti & Krueger 2015: 1169).

These variances can help to maintain, or destabilise, particular affective niches, or our connections with particular affective scaffolds. For instance, our sense of humour, distinctive as it is of particular kinds of social setting, may help to maintain our friends and colleagues as affective scaffolds. Yet if you go to a gathering of strangers and the close friend you expected to be there does not show up, not only might your emotional regulatory capabilities be initially diminished, such that you exhibit high degrees of anxiety and nervousness, so might this propensity further impoverish the affective niche in which you find yourself (if the others around you

are uncaring, or misinterpret your anxiety as rudeness or indifference, for example). The absence or unavailability of even one affective scaffold may precipitate further diminishment in one's affective capabilities until they are *severely* impoverished.

The main points to take away from this section are threefold. Firstly, our affective capabilities, in general, depend on various different kinds of affective scaffolds that exist across multiple timescales. Secondly, various factors may contribute or detract from particular affective scaffolds' availability to us in any particular situation, such that our affective capacities are correspondingly enhanced or diminished. And thirdly, there is often a relationship of reciprocal dependency between the affective capabilities dependent on certain affective scaffolds (or combinations of them) and the affective scaffolds themselves, which may result in both upward and downward spiralling of affective capabilities when otherwise stable affective niches are perturbed or otherwise altered in some way.

None of this is to say that the brain does not make significant, indeed invaluable and to some degree irreplaceable contributions to our affective capabilities. But it does suggest that our affective life in general – the particular affective capabilities we possess and exhibit – does not have its physical basis solely in brain activity. To the extent that elements of our own unique affective niches make both long and short-term contributions to the extent and existence of particular affective capabilities (such that if they were removed, such capabilities would either diminish in their potency, or cease to be exhibited altogether), the physical basis of affect is not simply locatable within the boundaries of skin and skull. In the next section, I shall argue that this is true of the two *specific* affective capabilities hypothesised to be at the core of anhedonia.

2.4. Situating Hedonic Capabilities

In this section, I shall argue for the following claim:

> [**Niche-Dependent Capacities**] Whether you endorse a capacity or sustainability theory of anhedonia, the relevant capabilities (of which anhedonia is taken to be a diminishment or loss) are, typically, the partial

outcome of functional gain reliant on external affective scaffolds forming an adaptive affective niche.

Let's unpack this claim (**NDC**). The capacity and sustainability theories each posit that a *different* affective capability is characteristically diminished in cases of anhedonia. The capacity theory posits that this capability is that of being able to experience relevantly high quantities of pleasure. When people experience anhedonia, according to the capacity theory, they are experiencing diminished overall *hedonic capacity* – they just can't experience *as much* positive affect as they used to, or should be able to, or some relevant comparison population, or is required for normal function given their life challenges (depending on your preferred notion of pathology[8]). The sustainability theory posits the same thing regarding the ability to maintain highs of positive affect over time after coming into contact with positive stimuli – call this *hedonic sustainability*. My goal in this section is to argue that both hedonic capacity and hedonic sustainability are, in general, partially dependent on the characteristics of the affective niches (with a view to both long and short timescales) occupied by the relevant individual; in short, to argue that they are *situated* psychological phenomena.

2.4.1 Situating Hedonic Capacity

I begin with hedonic capacity. Recall that hedonic capacity is defined as the degree of positive affect a person is capable of experiencing. Our question is simply this; what determines an individual's hedonic capacity?

It might be tempting to think that the answer is simply genetics (or the genome's phenotypic expression in brain and nervous system structure), and indeed hedonic capacity does appear to be partially so determined (Meehl 1975; Dworkin & Saczynski 1984, 623-24). But there is pretty clear evidence from twin comparison studies that the developmental environment plays a large role as well (Dworkin & Saczynski 1984, 623-24). That is simply to say that an individual's hedonic capacity is co-determined by genetic and developmental factors. Unfortunately, however, this is

[8] See Law & Widdows (2008), Bircher (2005), and Kovacs (1998), for some opinionated overviews of positions on the nature of health and illness.

not on its own a convincing consideration in favour of the idea that hedonic capacity is scaffolded by functional gain dependent on external affective resources.

The reason for this is simple – it could be (see Sonuga-Barke 2017) that the developmental environment has an impact on brain development that *in turn* partially determines overall hedonic capacity. If it turned out that the phenotypic expression of the genome and developmental impact on the brain completely co-determined an individual's hedonic capacity (for the record, I doubt this), then an opponent to my view could claim that hedonic capacity had its physical basis in the brain; it would just be that this physical basis arose from a variety of sources. And if, once established, this physical basis (and indeed one's hedonic capacity) was relatively invariant, their case would be even stronger – there would appear to be no indispensable need to reference the wider environment in accounting for an individual's hedonic capacity.

Fortunately there are other reasons to believe in strict partial dependence of hedonic capacity on affective scaffolds. Imagine you spend much of any given day plugged into a mp3 player, or that you get together with your friends at the weekend to play board games.

Take the first case. Imagine you go for a walk one day and forget your mp3 player. The effect is likely not to simply be frustration, you may find throughout the day that you are just ever so slightly (or indeed significantly) less happy than you would have been had you had the mp3 player. If this is not the case it is likely because you have found some other activity or resource in the environment that is able to play the same overall *affective functional role* (you have adapted your affective niche to manage the change). That is, the mp3 player is plausibly thought of as typically enhancing your moment-to-moment hedonic capacity, and its absence as diminishing it. Your ability to enjoy the rest of the day's activity *as much* as you might otherwise have done is significantly impoverished. It is only by adapting one's affective niche that one can restore hedonic capacity to what it was. This suggestion is far from merely speculative – Skånland (2013) provides qualitative evidence of mp3 players' utility, and sometimes utter indispensability, in intensifying individuals' day-to-day positive affective experiences.

Take the other case. Imagine your friends all cancel on you at short notice for one weekend's board game extravaganza. You may not be particularly angry or upset (they all have very good reasons not to attend, and you react to them appropriately). Nevertheless, you may find that the rest of the day just isn't quite as bright as it might otherwise have been. Again, rectifying that will involve finding some other resource in the external environment capable of making the same affective contribution as playing board games would have done. Without something along those lines, you are just not quite as able to experience the heights of positive affect as before. It's not just the enjoyment of the games themselves you miss out on, but the general enhancement to your *capacity* to enjoy yourself that the whole experience facilitates.

Neither of these cases are ones where we can plausibly link diminished hedonic capacity to structural changes in the brain. The changes are simply too quick and immediately associated with alterations or perturbations to an affective niche. It is much more plausible to account for the change directly in those terms. It's important to note that nothing about the above examples necessitates that the diminishment in hedonic capacity be particularly severe. It may not even be particularly obvious to the person experiencing it unless they reflect on it. The point is only that small alterations in affective niches can quickly precipitate diminished hedonic capacity. This is enough to suggest that elements of our affective niche typically scaffold functional gain in our hedonic capacity. That is, we achieve enhancements in our hedonic capacity that would not be possible without suitable environmental resources.

One might wonder why an internalist could not object to this on the following basis; the unavailablity of environmental artefacts here is simply causing a change in *mood* (assumed, perhaps fairly, perhaps not, to supervene on certain pattern of brain or nervous system activity), which is in turn responsible for the diminished hedonic capacity that the anhedonic person experiences. Thus the physical basis of hedonic capacity would, it seems, be found to be in the brain after all.

This is not quite right however. Even if this description is correct, the possibility of entering particular moods depends on particular environmental resources; we may

be unable to enter into particular moods (and thus enjoy their hedonic-capacity enhancing effects) *precisely because* of the unavailability of certain environmental supports. So this response, even if it describes the onset of diminished hedonic capacity accurately, does not refute the suggestion that the physical basis of hedonic capacity is not solely to be found within humans.

Further, the general idea that the environmental resources one has access to at a given time affects one's hedonic capacity has empirical support. Stressful environments (typically ones in which one is deprived of many normal affective scaffolds in one way or another – including military service and college exam periods) have been found to diminish individuals' hedonic capacity across a number of scales of measurement (Pizzagalli et al 2007; Berenbaum & Connelly 1993). And people seem to use established interpersonal relationships to enhance their hedonic capacity when in social groups (Gable & Reis 2010). The key takeaway here, and in this section as a whole, is that one's hedonic capacity is not stable, or fixed, but significantly dependent on material and social circumstance. A given individual's hedonic capacity is achieved in significant part through functional gain scaffolded by elements of the affective niches that they inhabit.

2.4.2 Situating Hedonic Sustainability

Let us move on to hedonic sustainability. A useful way of conceiving of this capability is as a form of emotional regulation. Koole defines emotional regulation "as the set of processes whereby people seek to redirect the spontaneous flow of their emotions." (2009: 6). Given this way of thinking, hedonic sustainability is about the processes used to 'redirect the flow' of affect towards a positive valence, in the sense that it is about preventing positive affective experiences moving towards more neutral (or even negative) ones. It is about how people 'keep themselves going' once they are already undergoing a positive affective experience. The goal of this section is to argue that many of these regulatory processes involve components distributed across individual's environments, and thus that an individual's capability for hedonic sustainability is typically partly dependent on affective resources located in the environment. This is just to say that affective scaffolds enhance an individual's capability for hedonic sustainability.

Theorists have argued persuasively that emotion regulation in general is scaffolded by both material and social environmental resources (e.g. Koole & Veenstra, 2015; Colombetti & Krueger, 2015). They have significant empirical support too. Actions (both conscious and unconscious) of caregivers, and subsequent interaction, enormously enhance the emotional regulatory capacities of infants (Varga 2016; Manian & Bornstein 2009; Stern et al 1985), and there is significant evidence that these sorts of processes continue between people and their friends and family throughout development and adulthood (Zaki & Williams 2013). Linguistic expression seems to enormously enhance emotional regulatory capabilities, as a result of enabling both emotional communication (Burleson 1985), and simply emotional articulation (Samur et al 2013; Lieberman 2011). People's ability to redirect their emotional experiences is also tremendously enhanced by a variety of material resources and activities including hot showers for alleviating loneliness (Bargh & Shalev 2012), cuddling soft toys or seeking interpersonal touch for down-regulating fear (Koole et al 2014), and seeking out and utilising cleaning products to down-regulate disgust (Koole et al 2015; Vogt et al 2010).

So it is fair to say that the evidence in favour of emotional regulation being a significantly environmentally scaffolded capability is strong. This does not, however, necessarily imply that the ability to sustain positive affect is specifically scaffolded in this way. Fortunately, we can make a case for that as well.

The first example will involve music again. Many people use it during workouts, mostly via portable music players, to enhance their ability to complete an otherwise gruelling regimen of physical exercise. It is natural to think that one significant task involved in motivating oneself to continue a workout is to maintain a positive affective state as consistently as possible throughout the regimen. Studies have shown not only that listening to music enhances positive affect throughout periods of exercise (Elliott et al 2004), but that it helps to maintain it during the workout (Lim et al 2009), and lengthens the time post-workout during which people continue to experience elevated affect (Elliott et al 2005). To the extent that we take such research seriously, it would appear that music players significantly enhance people's degree of hedonic sustainability during exercise. Moreover, for many people, other kinds of

stressors other than exercise are also made more manageable, and thus hedonic sustainability enhanced, through the consistent availability of music throughout the day (Skånland 2013).

There is also evidence that these sorts of enhancements are not limited to the domain of exercise. For instance, maintaining the right kind of positive affect over the course of a day is often achieved through selection (entirely deliberate or otherwise) of the right clothing or accessories at the beginning of it. Everything from the colour, to the texture, to the overall appearance of clothing seems to play a role in helping us to maintain a positive mood in the face of daily stressors (Moody et al 2010; Valdez & Mehrabian 1994; Kwon 1991). This is also clearly reflected in our language – we describe colours as 'calming' and entire outfits as 'powerful', amongst other things. So hedonic sustainability is also seemingly enhanced by clothing selection.

These are just a few examples of items of material culture scaffolding hedonic sustainability. Here is one more. On a night out, one's perceived energy levels and sense of continued enjoyment may be sustained through the use of cigarettes or alcohol (or indeed other, more interesting, substances). Moreover, this effect cannot easily be reduced to the chemical composition of these substances – simply the ritual of indulging in them, in a particular environment, with a particular group of people may help to sustain positive affect through exhaustion. Many of us are familiar with the phenomenon of becoming 'drunk on atmosphere'. This is simply to say that the scaffolding effects of these substances in certain contexts have a marked intersubjective quality.

So it seems as if material items and their associated cultural rituals are often recruited into the processes invoked to sustain positive affective experiences over time. There are of course numerous other examples beyond what I suggest above. Nevertheless, those examples are suitably representative to make my point.

Hedonic sustainability also exhibits a similar relation of dependence upon interpersonal (or social) scaffolds. It is a common enough experience that after a while of not interacting with other people, even individuals who are otherwise doing well begin to experience a dip in mood and general demeanour. Isolation is a

powerful force for dysregulating positive affective responses, seemingly regardless of intervening stressors.

Going beyond intuitions, studies indicate that interpersonal interaction is a major mechanism of both general positive affect maintenance and the prevention of negative emotional spirals in the presence of stressors (Zaki & Williams 2013), as well as being a proposed (and well-supported) mechanism by which social support acts as a protective mechanism against the onset of depression (Marroquin 2011). This is achieved in at least two ways; through the simple recognition that the affective resources of another person are available to react to a particular stressor (Zaki & Williams, 2013: 806) and/or via simple recognition that some experience is understood and shared by another person (Zaki et al 2011). Both of these methods permit a degree of stability in one's positive affective experiences that would otherwise be difficult or impossible to obtain, as evidenced by the strong predictive relationship between absent or maladaptive social support and mood disorders in general (Hofmann 2014: 490-491). This idea is further supported by an argument provided by Somogy Varga and Joel Krueger, to the effect that pairs of humans in close relationships (child-caregiver as well as adult-adult) exhibit synchronization of bodily movements, which acts as a mechanism of distributed emotion regulation (2013: 287). Studies of the regulatory dynamics of romantic partners (e.g. Saxbe & Repetti 2010), as well as emotional dysregulation following the death of a close partner (e.g. Sbarra & Hazan 2008), seem to provide confirming evidence for their picture. Thus, it seems likely that interpersonal regulation scaffolds functional gain in hedonic sustainability.

The upshot of the argumentation and evidence presented in this section, is that people actively contribute to the construction of affective niches, by creating, carrying with them, and consistently engaging with a variety of material and social scaffolds of affective capabilities. Moreover, from the composite elements of these affective niches spring significant determiners of people's moment-to-moment hedonic capacity and hedonic sustainability. Thus, *whichever* affective capability one believes to be diminished in anhedonia, that capability is, in the general case, enhanced or even made possible through engagement with environmental resources.

That is to say that the given capability will typically have part of its physical basis in the environment, and part in the activity of a person's brain and nervous system, or (perhaps better) that it will have its physical basis in the ongoing, active relationship between these internal and external elements.

So **NDC** is established. All that is now left to do is demonstrate that such a conclusion extends to the *impairment* of this capability as well.

2.5. Situating Anhedonia

Let us assume that the evidence above has convinced you that you should grant me **NDC** (at least for the sake of argument). We can now stop distinguishing between the two options for the characteristic affective capability involved in anhedonia. Take whichever position you prefer, or remain neutral if you wish. Call the capability *actually* characteristic of anhedonia 'the hedonic capability'. We know two important facts about the hedonic capability, whatever it may turn out to be. Firstly, its diminishment or loss is the defining psychological feature of anhedonia. Secondly, it is generally partially dependent on affective scaffolds (more precisely, an appropriately supportive affective niche) for its existence, or at least the degree to which a certain agent is able to exercise it. The extent to which this is the case will vary, but the hedonic capability is, in general, significantly enhanced by the presence of suitable affective scaffolds (to the degree that it would be unrecognisable in their absence). This suggests that the physical basis of the hedonic capability is not in general to be found solely in the brain. So, is this sufficient to make the further claim that the physical basis of anhedonia (i.e the loss or diminishment of the hedonic capability) is, consequently, not generally to be found solely in the brain either?

Certainly, the main claim of section 4 seems to speak strongly in favour of it. If the physical basis of some capability is generally, in part, to be found in the environment, then it seems likely that the physical basis of the diminishment of that capability will also generally, in part, be found in the environment. That is, if some capability is significantly enhanced by the presence of certain material and social resources in the surrounding environment, then it stands to reason that its diminishment will sometimes be explicable only with reference to the loss or diminishment of those same resources.

But we should not proceed so hastily. Firstly, it seems strange to speak of the physical basis of a diminishment in a capability. It sounds curiously close to speaking of the physical basis of an *absence* (which, in turn, sounds like nonsense). Secondly, is it not possible that, as it turns out, anhedonia is always, or nearly always the outcome of an issue in the neurological dimensions of the hedonic capability's physical basis? Although I have perhaps secured the conclusion that the physical basis of anhedonia is *possibly* not solely biological, I have failed to demonstrate that this is ever so (or so often enough to be interesting).

On the first point, the worry seems to rest on a mild ambiguity in the way I have been speaking so far. When I speak of the physical basis of a phenomenon, I am referring to the physical (in a broad sense of the term that includes the biological) entities and processes that interact to produce the phenomenon in question. So the claim that the physical basis of anhedonia is not to be found solely in the brain is to be read as something like 'the entities and processes that interact to produce anhedonia are not to be found solely in the brain'. My claim is that these entities and processes may include elements of an affective niche that has been impoverished in such a way that it is less capable, or even completely *incapable*, of supporting the hedonic capability in the manner it otherwise could. The operations and function of the affective niche of an anhedonic person have, in the general case, been disrupted so as to limit or eliminate its ability to scaffold the hedonic capability.

It seems that I must concede the second point to some degree. The reason why I have made reference throughout this paper to 'the general case' when giving my explanation has been that I seemingly must accept the possibility of limit cases where the hedonic capability is suitably degraded so as for the case to count as one of anhedonia, but where this is due solely to disruption in the brain activity that (partially) underpins it. While the affective scaffolds remain untouched in such a case, the neurological activity required to support the hedonic capability is disturbed. Thus we would have anhedonia *without* disruption to the affective niche.

I can't deny that, for all I can definitively demonstrate, such a situation is possible. That said, I have two responses that should diminish the force of this objection. Firstly, it seems suspiciously unmotivated to claim that, as it happens, all or even the

majority of cases of significant diminishment of the hedonic capability (read: anhedonia) are to be explained only with reference to the *brain-based* enablers and enhancers of the hedonic capability. This also seems simply unlikely. The only obvious reason for doing so is that one already endorses biomedical materialism, which is simply to beg the question at issue. The possibility I have raised here at the *very least* entails an obligation for researchers to take seriously the idea that many cases of anhedonia are partially based in disturbances to the individual's wider affective niche, and not simply their head. Further empirical study may, naturally, prove me wrong. Maybe affective niches do not scaffold the hedonic capability to a sufficient degree to account for an interesting range of cases. But (and this is key) to establish this, empirical researchers must first take the alternative possibility seriously in a way that they have not yet done. The possibility I have identified is grounded in well-evidenced empirical claims about the pervasively situated nature of human affective capabilities; it is hardly so speculative as to warrant summary dismissal. Nor do I think this question is *likely* to be resolved in the internalist's favour; the enormous degree to which affective capabilities require complex suites of environmental scaffolds in general speaks to the likelihood that many cases of anhedonia are the result of disturbances to these scaffolds themselves, rather than their neurological 'partners'.

Secondly, the reciprocally interactive relationship between humans and their affective scaffolds complicates a picture on which anhedonia is produced, as it happens, by disturbances to only the neurological dimensions of the hedonic capability's physical basis. It is hard to envisage how a diminished hedonic capability that is perhaps initially based solely in the brain could fail for long to disturb an individual's connection to their occupied affective niche. Lower hedonic capacity and/or hedonic sustainability regularly causes individuals to give up on previously enjoyable activities and results in them pushing other people away (as well as other people simply avoiding engaging with them). When that occurs, the diminished capability will not only be reflected in, but also *reinforced* or *worsened* by

the state of the affective niche that it has 'helped' to impoverish[9]. The upshot here is that even the tricky limit cases mentioned above are unlikely to remain fully describable as such for long – the relationships involved between brain and world are simply too interdependent. That is, the physical basis of an even somewhat enduring case of anhedonia is *particularly unlikely* to be found only in the neurological enablers of the hedonic capability, even if this was once the case.

A related worry presents itself. Here is an example adapted from Bechtel (2009). One might be concerned that I have bought my conclusion too cheaply. Imagine a typical car. One of its capabilities (that is, one of the things it can 'do') is moving forward. Call this 'the moving capability'. As in the case of the hedonic capability, this capability is enhanced, indeed made possible, by a variety of resources external to the car that scaffold it. The moving capability requires, amongst other things, relatively smooth, well maintained roads, petrol in the tank, a driver, perhaps (occasionally, one hopes) mechanics, and so on. It also requires an engine. Now, it makes sense to think that, while such external resources are indeed necessary for the existence and characteristic quality of the moving capability, there is clearly a sense in which the engine is *more central* to its realisation. The engine is a necessary element of the moving capability, and could only be replaced by something that is *very* similar to it. But the external requirements are multiply realisable – various different kinds of relatively even roads would do, as would many different kinds of petrol, and a wide variety of drivers.

On the basis of this, it might seem plausible to say the following. In the case of the moving capability, the engine is the locus of control; it is of primary importance in the realisation of the capability. Further, this suggests that the mechanism that underpins the moving capability is centred on the engine (and perhaps the wider car). Certainly, this mechanism can only realise the moving capability in the right environmental conditions (sodden, muddy grass on a steep hill, for instance, will stop the car in its tracks), but this does not imply that the mechanism itself is in any

[9] Note that this observation cuts both ways – disturbances in affective niches will not go unreflected in neural dynamics for long, further degrading the affective capabilities that were once scaffolded.

sense to be found in the wider environment. Hence, one might say that all that is *really important* in the analysis of the moving capability is exhausted by looking at the car itself – the environment is a relatively dull added extra. And, so the objection will go, everything I have said about the car and the moving capability applies also to the brain and the hedonic capability. Everything important in the realisation of the hedonic capability is going on in, or is mediated by, the brain.

The issue with this concern is that it misconstrues what I need to secure my main claim. I can happily grant everything that the above analogy might suggest. One may think that claiming that the hedonic capability is environmentally scaffolded is to imply that the mechanism of its realisation is to be (partially) found in the wider environment. This may well be true; it depends upon *how* exactly affective scaffolding operates, and how we ought to demarcate cognitive systems. If so, we should either reinterpret or reject the car analogy. But I need not commit to anything like that picture here.

Let me grant for the sake of argument that the car analogy is entirely accurate. The *mechanism* responsible for the realisation of the hedonic capability is localised in the brain, though it is dependent on suitable worldly circumstances to operate properly. Now extend the original analogy. Imagine that the car fails to move forward, or begins to do so slowly and with great strain. Independent of any further information, it would be patently illegitimate of me to assume that the engine is broken. Perhaps the surface is poor, or the driver is misusing the clutch. Likewise, if we envisage or perceive a case where the hedonic capability is not being realised, or is being realised only weakly, it is illegitimate of us to assume that the problem lies in the brain, absent any further information. Perhaps the individual in question has been excluded from a significant social network, or have (for some reason) grown affectively insensitive to some other activity they used to partake in to maintain or enable the hedonic capability (or, more likely, a multitude of such activities). Perhaps they are simply unable to engage with previously supportive elements of their affective niche because they have no time, or lack other required resources.

Even if we grant that the brain is the locus of control of the hedonic capability (and, I should emphasise again, this is not obvious), it still requires a whole collection of

external resources to effectively realise it. And since anhedonia *just is* (according to both of the dominant psychological theories) the lack or diminishment of the hedonic capability, the source of that diminishment is irrelevant to identifying cases of anhedonia. Since we can demonstrate partial dependence of that capability on an individual's affective niche in general, we must allow suitable disturbance or disruption to that niche as potential (partial) bases of anhedonia.

Let me make the point another way. Ruth Millikan (2013) distinguishes between a mechanism or system *malfunctioning* and it simply *failing to perform its function*. In the former case, the mechanism fails to execute its function despite 'normal conditions[10]' obtaining, due to "abnormalities in the *constitution of the device itself*" (Millikan 2013: 40). In the latter case, it is prevented from performing its function due to abnormal circumstances obtaining. This corresponds intuitively to the difference between an otherwise healthy 30 year-old person in a completely healthy environment suddenly having a heart attack, and the heart failing to pump blood after it has been removed from a living organism. In the first case the heart is naturally thought to have malfunctioned, in the second everything is fine with the heart itself; it simply fails to perform its function because it is not embedded in remotely convivial circumstances.

Even *if* the brain genuinely *is* the locus of control for the hedonic capability, and even if the *mechanism* responsible for the capability's realisation is properly specified internally, my main claim is still secure. If I am right that the hedonic capability is typically dependent for its existence and extent on an individual's particular affective niche, then the realisation of anhedonia does not generally depend on the brain, or a particular neurological subsystem malfunctioning, in Millikan's sense (though this may sometimes occur initially). Genuine cases of anhedonia may equally well arise in a situation where the relevant mechanism simply fails to perform its function, due to an unsuitable affective niche. The hedonic capability is diminished or lost *in both cases*. And, certainly, the course of anhedonia will tend to reflect both neural and environmental dynamics. To the degree that you endorse

[10] Take this notion intuitively here; what exactly Millikan intends by the term 'normal conditions' is complex, subtle, and not important for our purposes.

either of the contemporary theoretical perspectives on anhedonia, there is no good reason to think that only neurologically realised anhedonia (if it even exists at the actual world) is 'the real thing'.

2.6. References

Bargh, J.A. & Shaley, I. (2012), "The substitutability of physical and social warmth in daily life", *Emotion* (**12: 1**), pp.154-162

Barrett, L.F. (2012), "Emotions are real", *Emotion* (**12: 3**), pp.413-429

Barrett, L.F. (2014), "The conceptual act theory: A precis", *Emotion Review* (**6: 4**), pp.292-297

Bechtel, W. (2009). "Explanation: Modularity, mechanism, and Situated Cognition". in Robbins, P. & Aydede, M. (eds.), *The Cambridge Handbook of Situated Cognition*. Cambridge: CUP

Berenbaum, H. & Connelly, J. (1993), "The effect of stress on hedonic capacity", *Journal of Abnormal Psychology* (**102: 3**), pp.474-481

Bircher, J. (2005), "Towards a dynamic definition of health and disease", *Medicine, Healthcare and Philosophy* (**8: 3**), pp.335-341

Burleson, B.R. (1985), "The production of comforting messages: Social-cognitive foundations", *Journal of Language and Social Psychology* (**4: 3&4**), pp.253-273

Colombetti, G. (2012), "Psychopathology and the Enactive Mind", in Fulford, K.W.M., Davies, M., Gipps, R., Graham, G., Sadler, J. *et al* (eds.) *The Oxford Handbook of Philosophy of Psychiatry*, OUP: Oxford

Colombetti, G. (2017), "The embodied and situated nature of moods", *Philosophia* (**45: 4**), pp.1437-1451

Colombetti, G. & Krueger, J. (2015), "Scaffoldings of the affective mind", *Philosophical Psychology* (**28: 8**), pp.1157-1176

Davies, W. (2016), "Externalist Psychiatry", *Analysis* (**76: 3**), pp.290-296

Der-Avakian, A. & Markou, A. (2012), "The neurobiology of anhedonia and other reward-related deficits", *Trends in Neuroscience* (**35: 1**), pp.68-77

Dworkin, R.H. & Saczynski, K. (1984), "Individual differences in hedonic capacity", *Journal of Personality Assessment* (**48: 6**), pp.620-626

Elliott, D., Carr, S. & Savage, D. (2004), "Effects of motivational music on work output and affective responses during sub-maximal cycling of a standardized perceived intensity", *Journal of Sport Behavior* (**27: 2**), pp.134-147

Elliott, D., CArr, S. & Orme, D. (2005), The effect of motivational music on sub-maximal exercise", *European Journal of Sport Science* (**5: 2**), pp.97-106

Fiorito, E.R. & Simons, R.F. (1994), "Emotional imagery and physical anhedonia", *Psychophysiology* (**31: 5**), pp.513-521

Gable, S.L. & Reis, H.T. (2010), "Good news! Capitalizing on positive events in an interpersonal context", in Zanna, M.P. (ed.) *Advances in Experimental Social Psychology* (**42**), Elsevier: London

Germans, M.K. & Kring, A.M. (2000), "Hedonic deficit in anhedonia: Support for the role of approach motivation", *Personality and Individual Differences* (**28**), pp.659-672

Griffiths, P. & Scarantino, A. (2009), "Emotions in the wild: The situated perspective on emotion", in Robbins, P. & Aydede, M. (eds.) *The Cambridge Handbook of Situated Cognition*, CUP: Cambridge

Heller, A.S., Johnstone, T., Shackman, A.J., Light, S.N. *et al* (2009), "Reduced capacity to sustain positive emotion in major depression reflects diminished maintenance of fronto-striatal brain activation", *PNAS* (**106: 52**), pp.22445-22450

Hofmann, S.G. (2014), "Interpersonal emotion regulation model of mood and anxiety disorders", *Cognitive Therapy and Research* (**38: 5**), pp.483-492

Huber, L. & Kutschenko, L.K. (2009), "Medicine in a neurocentric world: About the explanatory power of neuroscientific models in medical research and practice", *Medicine Studies* (**1: 4**), pp.307-313

Kaviani, H., Gray, J.A., Checkley, S.A., Raven, P.W. *et al* (2004), "Affective modulation of the startle response in depression: Influence of the severity of depression, anhedonia, and anxiety", *Journal of Affective Disorders* (**83: 1**), pp.21-31

Köhler, S., Cierpinsky, K., Kronenberg, G. & Adli, M. (2016), "The serotinergic system in the neurobiology of depression: Relevance for novel antidepressants", *Journal of Psychopharmacology* (**30: 1**), pp.13-22

Koole, S.L. (2009), "The psychology of emotion regulation: An integrative review", *Cognition and Emotion* (**23: 1**), pp.4-41

Koole, S.L., Webb, T.L. & Sheeran, P.L. (2015), "Implicit emotion regulation: Feeling better without knowing why", *Current opinion in Psychology* (**3**), pp.6-10

Koole, S.L. & Veenstra, L. (2015), "Does emotion regulation occur only inside people's heads? Toward a situated cognition analysis of emotion-regulatory dynamics", *Psychological Inquiry* (**26**), pp.61-68

Kovacs, J. (1998), "The concept of health and disease", *Medicine, Healthcare and Philosophy* (**1**), pp.31-39

Kwon, Y. (1991), "The influence of the perception of mood and self-consciousness on the selection of clothing", *Clothing and Textiles Research Journal* (**9: 4**), pp.41-46

Law, I. & Widdows, H. (2008), "Conceptualising health: Insights from the capability approach", *Health Care Analysis* (**16: 4**), pp.303-314

Lieberman, M.D. (2011), "Why symbolic processing of affect can disrupt negative affect: Social cognitive and affective neuroscience investigations", in Todorov, A., Fiske, S.T. & Prentice, D.A. (eds.) *Social Neuroscience: Toward understanding the underpinnings of the social mind*, OUP: Oxford

Lim, H.B.T., Atkinson, G., Karageorghis, C.I. & Eubank, M.M. (2009), "Effects of differentiated music on cycling time trial", *International Journal of Sports Medicine* (**30: 6**), pp.435-442

Lima-Ojeda, J.M., Rupprecht, R. & Baghai, T.C. (2017), "Neurobiology of depression: A neurodevelopmental approach", *The World Journal of Biological Psychiatry*, available at: https://doi.org/10.1080/15622975.2017.1289240

Liu, W., Chan, R.C.K., Wang, L., Huang, J. *et al* (2011), "Deficits in sustaining reward responses in subsyndromal and syndromal major depression", *Progress in Neuro-Psychopharmacology and Biological Psychiatry* (**35: 4**), pp.1045-1052

Mallorquí, A., Padrao, G. & Rodriguez-Fornells, A. (2014), "Electrophysiological signatures of reward processing in anhedonia", in Ritsner, M.S. (ed.) *Anhedonia: A comprehensive handbook Volume I: Conceptual issues and neurobiological advances*, Springer: Dordrecht

Manian, N.& Bornstein, M.H. (2009), "Dynamics of emotion regulation in infants of clinically depressed and nondepressed mothers", *The Journal of Child Psychology and Psychiatry* (**50: 11**), pp.1410-1418

Marroquin, B. (2011), "Interpersonal emotion regulation as a mechanism of social support in depression", *Clinical Psychology Review* (**31**), pp.1276-1290

Meehl, P.E (1975), "Hedonic capacity: some conjectures", *Bulleting of the Menninger Clinic* **(39: 4),** pp.295-307

Millikan, R. (2013), "Reply to Neander", in Ryder, D., Kingsbury, J. and Williford, K. (eds.) *Millikan and her critics*, Wiley-Blackwell: Oxford

Moody, W., Kinderman, P. & Sinha, P. (2010), "An exploratory study: Relationships between trying on clothing, mood, emotion, personality and clothing preference", *Journal of Fashion Marketing and Management: An International Journal* (**14: 1**), pp.161-179

Oyebode, F. (2015), *Sims' symptoms in the mind: Textbook of descriptive psychopathology (5th Edition)*, Saunders Elsevier: Oxford

Pizzagalli, D.A., Bogdan, R., Ratner, K.G. & Jahn, A.L. (2007), "Increased perceived stress is associated with blunted hedonic capacity: Potential implications for depression research", *Behaviour Research and Therapy* (**45: 11**), pp.2742-2753

Pizzagilli, D.A., Iosifescu, D., Hallett, L.A., Ratner, K.G. *et al* (2008), "Reduced hedonic capacity in major depressive disorder: evidence from a probabilistic reward task", *Journal of Psychiatric Research* (**43**), pp.76-87

Rizvi, S.J., Pizzagalli, D.A., Sproule, B.A. & Kennedy, S.H. (2016), "Assessing anhedonia in depression: Potentials and pitfalls", *Neuroscience and Biobehavioral Reviews* (**65**), pp.21-35

Robbins, P. & Aydede, M. (2009), "A short primer on situated cognition", in Robbins, P. & Aydede, M. (eds.) *The Cambridge Handbook of Situated Cognition*, CUP: Cambridge

Samur, D.,Tops, M., Schlinkert, C., Quirin, M. *et al* (2013), "Four decades of research on alexithymia: Moving toward clinical applications", *Frontiers of Psychology* (**4**), p.861

Saxbe, D. & Repetti, R.L. (2010), "For better or worse? Coregulation of couples' cortisol levels and mood states", *Journal of Personality and Social Psychology* (**98: 1**), pp.92-103

Sbarra, D.A. & Hazan, C. (2008), "Coregulation, dysregulation, self-regulation: an integrative analysis and empirical agenda for understanding adult attachment,

separation, loss and recovery", *Personality and Social Psychology Review* (**12**), pp.141-167

Shapiro, L. (2011), *Embodied Cognition*, Routledge: Oxford

Skånland, M. (2013), "Everyday music listening and affect regulation: The role of MP3 players", *International Journal of Qualitative Studies on Health and Well-being*, (**8: 1**), 20595

Sonuga-Barke, E.J.S. (2017), "Commentary: Extraordinary environments, extreme neuroplasticity and mental disorder – reflections on pathways from adversity to mental disorder prompted by McCrory, Gerin, and Viding (2017)", *The Journal of Child Psychology and Psychiatry* (**58: 4**), pp.358-360

Sterelny, K. (2010), "Minds: extended or scaffolded", *Phenomenology and the Cognitive Sciences* (**9: 4**), pp.465-481

Stern, D.N., Hofer, L., Haft, W. & Dore, J. (1985), "Affect attunement: The sharing of feeling states between mother and infant by means of inter-modal fluency", in Field, T.M. & Fox, N.A. (eds.) *Social Perception in Infants*, Ablex: New Jersey

Tomarken, A.J. & Keener, A.D. (1998) "Frontal brain asymmetry and depression: A self-regulatory perspective", *Cognition and Emotion* (**12**), pp.387-420

Treadway, M.T., Bossaller, N.A., Shelton, R.C. & Zald, D.H. (2012), "Effort-based decision-making in major depressive disorder: A translational model of motivational annhedonia", *Journal of Abnormal Psychology* (**121: 3**), pp.553-558

Treadway, M.T. & Zald, D.H. (2011), "Reconsidering anhedonia in depression: lessons from translational neuroscience", *Neuroscience and Biobehavioral Reviews* (**35: 3**), pp.537-555

Valdez, P. & Mehrabian, A. (1994), "Effects of color on emotions", *Journal of Experimental Psychology: General* (**123: 4**), pp. 394-409

Varga, S. (2016), "Interaction and extended cognition", *Synthese* (**193: 8**), pp.2469-2496

Varga, S. & Krueger, J. (2013), "Background emotions, proximity and distributed emotion regulation", *Review of Philosophy and Psychology* (**4**), pp.271-292

Vogt, J., Lozo, L., Koster, E.H.W. & De Houwer, J. (2010), "On the role of goal relevance in emotional attention: disgust evokes early attention to cleanliness", *Cognition and Emotion* (**25**), pp.466-477

Wilson, R.A. & Clark, A. (2009), "How to situate cognition: Letting nature take its course", in Robbins, P. & Aydede, M. (eds.) *The Cambridge Handbook of Situated Cognition*, CUP: Cambridge

Wohleb, E.S., Franklin, T., Iwata, M. & Duman, R.S. (2016), "Integrating neuroimmune systems in the neurobiology of depression", *Nature Reviews Neuroscience* (**17**), pp.497-511

Zachar, P. (2000), *Psychological Concepts and Biological Psychiatry*, John Benjamins: Amsterdam/Philadelphia

Zaki, J., Schirmer, J. & Mitchell, J. (2011), "Social influence modulates the neural computation of value", *Psychological Science* (**22: 7**), pp.894-900

Zaki, J. & Williams, W.C. (2013), "Interpersonal emotion regulation", *Emotion* (**13: 5**), pp.803-810

Chapter 3: On Motivation

3.0. Abstract

This paper puts forward two central theses. Firstly, I argue that a significant extant account of motivational mechanisms (i.e. how motivational mental states perform their characteristic function), due to Wayne Wu (2011), fails to adequately explain two of the three central functional roles of motivational mental states; action-initiation and action-guidance. Nevertheless, I think it provides helpful insight into the third; action-control. Secondly, I argue that a predictive processing framework gives us the explanatory power required to account for all three, and thus that the predictive theory of motivation should be preferred to Wu's.

I argue that Wu's account fails to explain how motivational mental states actually initiate action. Instead, I suggest, it merely offers an account of how motivational mental states could simplify the process of initiating action. Following that, I argue that Wu's story implies an explanation of action-guidance that is both computationally expensive and biologically unrealistic, to the point of being implausible. Nevertheless, I note that his account explains motivational mental states' role in action control well, and suggest that other theories should preserve this insight.

I go on to argue that a predictive processing account of motivational mental states can provide an alternative to Wu's story, providing plausible explanations of all three of motivation's central functional roles. I conclude that the predictive theory is to be preferred to Wu's as a result.

3.1. Introduction

The main goal of this paper is to develop and defend a theory of the mechanisms underlying motivation. I aim for it to provide a plausible explanation of the key functions of motivation that have previously been identified in the literature – namely motivation's role in the **initiation**, **guidance**, and **control** of action[11] (Shepherd 2017; Pacherie 2012: 96; Elliott 2006). I shall begin by clarifying the notion of motivation I am working with, before moving on to clarifying the question this paper will concern itself with.

When philosophers and psychologists speak of motivation, they typically speak of *motivational mental states*. These are typically glossed as any mental state, or state-complex, with the power to generate action (I will clarify this shortly). The paradigm of this class of mental states is an *intention*, but the class more generally is widely thought to include states such as emotions (Scarantino & Nielsen 2015; Tappolet 2010; Frijda 1986), pains (Corns 2014; Bain 2011), desires (Smith 1994) or action-desires (Mele 1992), and perhaps even certain kinds of belief[12] (McDowell 1979; Dancy 1993).

Further, many action theorists also distinguish between *proximal* and *distal* intentions[13] (Pacherie, 2012; Mele, 1992; Bratman, 1987; Searle, 1983). Proximal intentions are those that have the power to initiate, guide, and control action (Shepherd 2017; Pacherie, 2012: 96). Roughly, that is, they are those mental states that (at least according to a *causal* theory of action, which I shall assume throughout[14]) are

---

[11] Not all elements of the literature identify these functions in quite the same way – Elliott (2006) for instance speaks of motivation as that which 'energises' and 'directs' behaviour. This does not reflect a significantly different opinion regarding motivation's core functions – it is simply a less precise characterisation of the same basic functions mentioned above.

[12] I use 'belief' here as a convenient shorthand; what McDowell and Dancy think is *actually* motivating is some kind of 'moral conception/perception of the situation', which may or may not count as a belief *per se*.

[13] Terminology here is far from uniform – Searle, for instance, distinguishes between *intentions-in-action* and *prior intentions*, while having basically the same distinction in mind (1983). Here, I use Mele's (1992) terminology.

[14] Typically, these theories of action contrast with *agent causal theories*, which posit that only entire agents, rather than any individual state (or state-complex) of those agents, can cause bodily behaviours in such a way that they count as actions (see O'Connor 2002 for an overview). I will not consider such theories here, preferring to address my remarks towards those broadly located in the traditional causal camp.

able to cause bodily behaviours in such a way that they count as actions rather than mere reflexes. If I pick up a glass of water, then that bodily behaviour counts as an action only if it was caused (non-deviantly[15]) by an intention of mine to pick up the glass. The way to think about such intentions is as standing states (Wu 2011) that are disposed to act as the immediate mental antecedents (Nanay 2013) of particular bodily actions, when suitable circumstances arise or, at least, are perceived or judged to have arisen.

A distal intention is one that is associated with the functions intentions have been suggested to play in practical reasoning (see Pacherie 2012: 95-96; Bratman 1987). That is, they are those kinds of intentions that act as prompters and terminators of different forms of practical reasoning ('means-end' and 'what-to-do' reasoning respectively), as well as coordinating agents' activities over time and with other agents.

The labels of 'proximal' and 'distal' are thus roughly intended to capture a distinction between the motor and deliberative functions of intention. Some also think that this distinction does some significant metaphysical work in action theory proper, for instance by opening up a path to a solution of the problem of minimal actions[16] (Pacherie 2012: 96; Searle 1983).

My focus in this paper will be on motivational mental states, understood as mental states with the power to play the roles ascribed above to proximal intentions; action initiation, guidance, and control (I will unpack these notions in the next section). Being able to cause action, rather than being able to play a particular kind of deliberative role is, in my view, the basic power that distinguishes motivational mental states from motivationally inert ones.

---

[15] What non-deviance consists in is an open question that I will not try to answer here – see Wu (2016) and Shepherd (forthcoming) for some recent suggestions.

[16] Minimal actions are those which are performed routinely, unthinkingly, and pre-attentively. Examples include shifting gear while driving or pacing anxiously across a room (Searle 1983). The problem is that while such bodily behaviours seem much too complex and controlled to count as mere reflexes, they are apparently not preceded by a clear, specific intention of which the agent is aware. So how can a causal theory class them as actions (Bach 1978)? The standard strategy is for a causal theorist to claim that while such actions are not preceded by *distal* intentions (which are presumed to be the kind of which an agent are largely aware) they *are* nevertheless preceded by some kind of *proximal* intention. The details and difficulties of various specific responses of this kind need not concern us here.

One reason to think this is that it is widely agreed that emotions are motivational mental states. But emotions are not always (perhaps even often) thought of as being the initiators or terminal points of practical reasoning, but rather as being directly elicited by environmental triggers or stimuli, such as threats or losses. What they *can* intuitively do, however, is cause actions; the fear I experience when I see a spider can intuitively cause me to do things that are far more than simple reflexes – I might call for my housemate if I think she is in, or cautiously retreat from the room and seek out a mug and a postcard. Absent other criteria, it thus makes sense to characterise both emotions and intentions as motivational mental states in virtue of their action-causing power and function, since this is what they have in common, at least on the surface.

Secondly, it is clearly this direct *action-causing* power that is at stake in philosophical debates over what sort of mental states are able to count as motivational (for instance between Humeans and Anti-Humeans). Thus, while Michael Smith (1994) thinks that beliefs are never *by themselves* motivational (always requiring the presence of some attendant desire, or 'pro-attitude' to actually generate action), Jonathan Dancy (1993) disagrees. Properly understood, according to him, our moral beliefs alone are sufficient to move us to act. What is at stake in their exchange is precisely whether (some) moral beliefs are able to cause action without an associated desire, in the way Dancy claims and Smith denies. Thus, we see that moral belief's status as a motivational mental state-type hangs on its purported capacity to directly cause action. And the functions outlined above are simply a breakdown of the functional roles mental states are thought to need to play in order to count as *causing action* – they must (at least) initiate, guide, and control bodily behaviour.

More exactly, my focus in this paper is the mechanism by which motivational mental states play their action-initiating, guiding, and controlling functions. Why do I focus only on action initiation, guidance, and control? Primarily, the choice is a pragmatic one, informed by the facts that, 1) these are the functions most often identified in the wider literature as under the purview of motivational mental states, and 2) they are thought to be the key components of motivational mental states' overarching action-causing function (Shepherd 2017; Pacherie 2012). Since identifying *all* relevant

functional roles of motivational mental states in action appears to be a significant ongoing empirical and conceptual task[17], we must begin somewhere. Widely agreed upon functions of motivational mental states are at least as good starting points as any others.

Assuming that theorists are correct about these three distinguishing functional roles of motivational mental states, it is clearly a necessary feature of any theory of the mechanisms of motivation that it be able to explain how they are performed. It may well be that there are other critical desiderata; perhaps the ability to explain other critical functional roles yet to be widely identified, for instance. But it is important to recognise that if a purported theory of motivational mechanisms does not clearly specify how motivational mental states can initiate, guide, and control behaviour, then it is failing to explain the functional roles that distinguish motivational states from motivationally inert states. Thus, it cannot qualify as an adequate theory of motivation. Naturally, it might count as a *partial* theory of motivation, if it successfully accounts for some but not all of these central functional roles, or takes incomplete steps toward accounting for them. Nevertheless, a minimal adequacy condition on a *complete* theory of motivation is that it account for all three of these central functional roles.

This paper puts forward two central theses. Firstly, that a significant extant account of motivational mechanisms, due to Wayne Wu (2011), fails to adequately explain two of these three central functional roles of motivational states. Secondly, that a predictive processing framework gives us the explanatory power required to account for all three, and thus that the predictive theory of motivation should be preferred to Wu's.

The following three sections focus on establishing the first thesis. In section 3.2., I unpack the notions of initiation, guidance, and control in more detail, providing a clear explanatory target to guide the rest of the paper. In section 3.3., I present Wu's account, focusing on how it purports to explain initiation, guidance, and control of

[17] Recently, for instance, Wulf et al (2010) provide evidence that motivational mental states play a crucial role in processes of complex motor skill learning.

actions. In section 3.4., I argue that it fails to account for initiation, and is deeply unparsimonious and biologically unrealistic in its explanation of guidance. I do, however, note that its account of control, when augmented by Wu's more recent work (2016) is a good one, that ought to be preserved by any alternative theory.

Next, I turn to establishing the second thesis. In section 5 I introduce the predictive processing framework. Finally, in section 6, I shall use this framework to present an alternative theory of the mechanism by which motivational mental states fulfil their three key functions, and explain how it does a better job than Wu's. My basic claim is that motivational mental states play their role of initiating, guiding and controlling action by causing the prediction of, and selective redeployment of attention towards, action-relevant proprioceptive and exteroceptive sensory signals. The devil, naturally, is in the details, which I shall explain closely in section 6.

3.2. Three Functional Roles for Motivation

How exactly should we understand the three key functional roles of motivational mental states identified above? Let's start with action initiation. This is the most basic of the functions ascribed to motivational mental states by causal theories of action. On any such theory, motivational mental states are "responsible for triggering or initiating the intended action" (Pacherie 2012: 96). No causal theory of action would be worthy of the name if it did not claim that motivational mental states of some kind or another possess this function, simply because whatever else intentions do for processes of action, if they do not trigger them, they cannot be sensibly said to *cause* them. Notice that it would be insufficient here for a theory to suggest that motivational mental states somehow *simplify* the process of initiating action, or that they merely *enable* action, or that they provide some *background conditions* necessary for the action to be initiated. None of these things by themselves amount to anything like *causing* the relevant action. Simplifying, enabling, or setting up some necessary background conditions for action to be initiated may turn out to be necessary functions of motivational mental states in causing actions, but they are not by themselves sufficient. For a theory of motivation to account for the initiating function, it needs to posit a mechanism by which motivational mental states play a *direct*, rather than merely *supportive* role in causing actions.

To help distinguish these notions, consider the following. If I successfully climb a ladder, it is plausible that many mental states will have supported my doing so – my belief that there is a ladder to be found somewhere in the environment, visual states that inform me of the actual presence of a ladder, its location, shape, and so on, proprioceptive states that inform me of the relative position of my hands and feet, etc. It is possible that without any of these mental states/processes I would fail to successfully climb the ladder. But when I do successfully climb a ladder, it is unnatural to say that they initiated the action. Rather, they are enablers of my climbing. On the other hand, it is considerably more natural to speak of my *intention* to climb the ladder as the thing that initiated my climbing behaviour (even though it is necessarily supported by a range of other states). Even though I will likely not climb a ladder (even if I intend to) if I cannot find one, it seems wrong, or at least unnatural, to answer a question of 'Why did you climb that ladder?' with 'Because I saw it'. After all, I regularly see ladders and do not climb them. Intuitively, the difference-maker in the cases where I *do* climb ladders is my intention (or whatever other motivational state) to climb them. It is important that theories have a principled explanation of this intuitive distinction – they must explain why it is right to think of motivational states (rather than any other kind of state) as the key difference makers in cases where we do act, as contrasted with cases where we do not; glibly, how they actually make actions happen, rather than how they make it merely possible for actions to happen.

Next, I will clarify the notion of action guidance. Motivational mental states are generally thought to support the unfolding course of an action through to completion, by specifying both an agent's goal, and how it will be arrived at – an *action plan* (Pacherie 2012: 96). More precisely, a motivational mental state is thought to accomplish its guidance function just in case it (partially) determines the overall bodily movement that the agent exhibits and directs it towards the satisfaction of the goal the mental state represents.

I say 'partially' determines simply because it is typically held that motivational mental states specify a coarse-grained action plan (i.e. to kick the ball) rather than a fine-grained one (i.e. to bend the knee to some precise degree, move the foot forward

at just such a speed, and make contact with the ball at some specific angle) (see Wu, 2011: 60). That is, motivational mental states do not specify particular, individual motor trajectories, but abstract plans satisfiable by many different concrete movements. Thus any particular concrete bodily action is likely to be strictly underdetermined by the motivational mental state that gave rise to it. Nevertheless, it is thought that these plans play *some* role in specifying the final concrete movement. One noteworthy feature of the predictive account of motivation that I will present and defend in section 3.6 is that it appears that motivational mental states, on such a theory *do* specify particular trajectories rather than merely abstract plans. That is, on such a view, the state in fact *does* fully determine the precise trajectory of the relevant action, at least in a sense. I explain why this is so, and why we should not see it as a bug but rather as a feature of the predictive theory, in section 3.6.2.

Further, by 'directing a movement toward the satisfaction of a goal', I mean to suggest a process of ensuring that the movement executed is appropriate given the agent's represented goal. So, if the agent intends to kick a ball, the intention satisfies its guidance role only if it causes the *leg* to *kick* towards the *ball*, rather than, say, causing an *arm* to *punch* towards a nearby *defender* (regardless of how much easier it might be to subsequently kick the ball).

Finally, I will clarify my working definition of control. I think Josh Shepherd (2014: 400) has it right when he argues that an agent exhibits control over some action-type to the degree that they exhibit *flexible repeatability* in their performance of such actions. The action must be repeatable, because merely possessing the brute causal power to do something does not imply that one has any significant degree of control over that ability (Shepherd 2014: 400). Since I have the ability to kick a ball a short distance in front of me, I (strictly speaking) have the ability to kick a ball a small distance in front of me such that it goes into a football net. Nevertheless, I lack the ability to kick a ball in any particular direction with any degree of repeatability, and so I lack the ability to reliably score penalties. Intuitively, what I lack is sufficient *control* over my kicking.

Moreover, repeatability in identical sets of circumstances is very easy to come by. Imagine I can repeatedly score a penalty when I am perpendicular to the goal, but not when I change my angle with respect to the goal by a few degrees. Once again, it appears that I do not, in such a scenario, exhibit any significant degree of control over my kicking behaviour. My ability to perform the action must be repeatable in circumstances that differ in "some theoretically interesting way" (Shepherd 2017: 266). We need not investigate here exactly what 'theoretically interesting differences' amount to. Suffice it to say that the greater the degree of repeatability of some action, and the wider the range of circumstances in which that repeatability is exhibited, the greater the degree of control an agent exhibits over that action-type.

Note that the mere fact that an agent has only performed an action once does not preclude the agent exhibiting control on this account. The relevant repeatability and flexibility may be specified with respect to a well-selected collection of counterfactual circumstances (Shepherd 2014: 400-403) – i.e. an agent may be said to exhibit a reasonable degree of control over an action if they *would* be able to repeat performance of the action in a reasonable range of counterfactual circumstances.

In the sense that I will use it here, a motivational mental state's control *function* is satisfied when the state ensures a significant degree of flexible repeatability in the performance of the action it specifies. Generally, this is thought to be achieved by the motivational mental state monitoring the progress of some bodily movement, and correcting this trajectory where it deviates from the action plan (Pacherie 2012: 96). In this sense, control, *qua* function of a motivational mental state, is thought to be achieved by various processes of monitoring and adjustment.

3.3. Wu's Account

In his 2011, Wayne Wu proposes an account of how motivational mental states play their causal role in generating action. This is part of a broader project of identifying necessary and sufficient conditions for intentional bodily agency and control. The following is a paraphrase of Wu's view of the mechanism by which motivational states perform their role in generating action.

> An agent's (A) motivational state, M, plays a causal role in the generation of an action, ϕ, by structurally causing the selection required in implementing a solution to the non-deliberative Many-Many Problem appropriate to A ϕ-ing in the given context. (2011: 68)

Naturally, this requires some unpacking. Let us start with the non-deliberative Many-Many problem. This is a problem necessarily faced by any creature that is able to exhibit bodily agency. In any given situation, an agent will confront innumerable perceptual inputs, and a similarly staggering number of possible bodily behaviours. To act, an agent must, on the one hand, select a target from amongst the many identified by their perceptual inputs (that is, they must *attend* to some aspect of their perceptual field). On the other hand, agents must also select one from the many possible bodily behaviours they could perform (Wu, 2011: 56).

A solution to the Many-Many problem, then, is a selection of both a target for action and a behaviour to perform on that target (roughly, e.g., paying attention to a hammer, and choosing to pick it up). The deliberative variant of this problem arises in the context of planning and practical reasoning/reflection. The non-deliberative problem (the important one for our purposes), on the other hand, concerns *implementation*. An easy way to think about the distinction is as connected to the distinction between distal and proximal intentions discussed in section 1. A solution to the deliberative problem involves, roughly, the formation of a distal intention, whereas a solution to the non-deliberative problem involves a proximal intention (or relevantly similar motivational mental state) playing its characteristic functional roles in action. In Wu's terms the non-deliberative problem "is solved in part by the agent's exercising perceptual and motor capacities" (2011: 56). Since the three key functional roles we identified in section 1 are generally thought to constitute a relatively fine-grained analysis of motivational mental states' action-generating power, even though Wu does not himself structure his discussion around these central functions, it is fair to think that Wu intends his theory to account for their exercise. That is, the perceptual and motor capacities involved in solving the non-deliberative Many-Many problem are supposed to underpin the initiation, guidance, and control of action. Though I agree that Wu has a good story to tell about the latter

function, I shall raise problems below for his theory's explanatory adequacy regarding the first two.

Next, we need to better understand the notion of a structural (or structuring) cause. Wu's understanding is a slight expansion on that of Dretske (1993), who coined the term – for Wu, a structuring cause is an event or standing state "that produces certain enabling conditions whereby one event… can cause another..." (2011: 58). For Wu, motivational mental states are standing states that are also "structuring causes enabling specific links between attention and movement that amount to a solution to the non-deliberative Many-Many Problem" (2011: 58). That is, they *enable* the selection of perceptual inputs and behavioural outputs appropriate to the execution of the action that they specify, given the context in which the agent finds themselves.

3.3.1. Wu on Control

In his 2016, Wu argues that causing the task-relevant selection of particular objects relevant to the performance of an action is a significant way in which motivational mental states play their controlling function. I basically agree with Wu on this point[18], and will return to the main idea shortly. The first point worth noting however, is that Wu's theory of motivation involves positing that motivational mental states cause the deployment of attention towards task-relevant objects. The upshot, if we agree with the basic thrust of his 2016 paper, is that his 2011 paper can be read charitably as having offered an explanation of how motivational mental states perform their control function.

Wu (2016) begins by noting that the causal theory of action has classically been assailed by problems of causal deviance. Roughly, these are cases where motivational mental states cause bodily behaviour, but in the wrong way for that behaviour to count as intentional action, As Wu puts it, in such cases "behavior and its effects, even if intended, seem to happen accidentally" (2016: 103). For example, imagine a scenario where I am driving, and form an intention to kill my father. Realising that I have formed such an intention, I become seriously unnerved, so

[18] I disagree with him on a few philosophically unimportant details regarding the most plausible neurobiological mechanisms of attention.

much so that I drive dangerously fast. This results in me knocking down and killing a pedestrian, who turns out to be my father (this sort of example is due to Roderick Chisholm (1966)). Plausibly my intention caused me to kill my father, but seemingly not in the right way. It does not seem I exercised agentive control over the behaviour that ended up with the death of my father – this sequence of events certainly would not mean that I exhibit flexible repeatability with respect to the killing of my father. So this is not a case where I *acted* so as to kill my father, because my intention did not cause the killing behaviour 'in the right way'. More generally, cases of deviance are ones where "some control-undermining state or event occurs between the agent's reason states [read: motivational mental states] and an event produced by that agent." (Schlosser 2007: 188). That is, the intuition underlying the thought that cases of causal deviance are not genuine actions is that such cases are ones where the caused behaviours are uncontrolled (or insufficiently controlled). Thus, if we can identify a mechanism that operates only in cases of non-deviantly caused action (and plausibly explains such actions' lack of deviance), then we have identified a good candidate for a contributor to action control.

What then is 'the right way' for a motivational mental state to cause an action? Wu claims, roughly, that cases where control is undermined are ones where the motivational mental state does not cause the deployment of attention to task-relevant elements of the environment. The consequence is that a primary source of information relevant to the monitoring and correction of the agent's movements is not available – effectively monitoring and correcting bodily movements depends on having information about the intended external targets of those movements. This position seems plausible when we reflect on paradigmatic cases of causal deviance. Where I am simply unnerved by my intention to kill my father, the intention does not cause me to attend to those aspects of the environment relevant to monitoring my movements' progress towards the goal-state, nor do I engage in any such monitoring relevant to achieving the outcome. Indeed, it is plausibly due to my *lack* of attention to the 'target' of my intention (but rather, perhaps, to my own nervousness) that I end up hitting my father. More generally, my failure to attend to task-relevant features of the environment will ensure that in a suitably selected

collection of counterfactual circumstances, I will *fail* to kill my father. My intention does not perform its role of redirecting attention appropriately, thus I fail to exhibit flexible repeatability in my killing of my father and so, definitionally as well as intuitively, my control over my behaviour has been undermined.

Consider a variation of this case, the structure of which is due to Wu (2016: 115-116). I intend to kill my father, and drive at high speed towards his home. My attention is suddenly captured by a person wearing a brightly coloured coat, and I inadvertently steer towards them, resulting in me hitting and killing them. The killing is, once again, unintended. That is, it does not exemplify agentive control. But this time it is a case of distraction rather than deviant causation (after all, assuming the person is not my father, I did not thereby satisfy the content of my intention). Once again, it seems like what went 'wrong' is that the target of my attention did not enable monitoring and correction of my movements towards the satisfaction of my intention, but rather caused me to exemplify an inappropriate set of movements. While attention certainly influenced my motor behaviour, it did not do so in a way that exemplifies control, because it, in turn, was not deployed in such a way that was relevant to the satisfaction of my intention (my attention was captured, rather than guided by my intention).

These considerations count as some evidence that intentions contribute to action control by causing attentional deployment towards the target of the action (as represented in the intention's content). Paradigmatic cases where we fail to exhibit control (or at least fail to exhibit *much* control) are plausibly ones where our intentions do not cause appropriate task-related attentional deployment.

Moreover, there is independent empirical evidence for this claim. For instance, Michael Mrazek and colleagues (2013) found that, compared to controls, participants who partook in two weeks of 'focused-attention meditation' training, significantly improved their performance in a verbal reasoning and working memory task, and reported fewer and less severe instances of mind-wandering while they were performing the task. The suggestion here is that honing the process of intention-caused attentional deployment improved agents' control over the actions necessary to complete certain tasks. While these results are instructive, we must of course note

the complication that the actions primarily involved in completing the aforementioned tasks were mental rather than bodily. Slightly more embodied examples (where participants are instructed to respond physically to, or withhold response from, certain cues appearing on a screen), also reveal a similar effect. Disengagement of attention from task-relevant aspects of the external environment, as might be expected, leads to impaired task performance (McVay & Kane, 2009; Smallwood, Beach, Schooler & Handy, 2008). Assuming that impaired task performance is a good measure of a lack of agentive control over the task (which seems plausible, and is done elsewhere, see Shepherd (2017)), these results indicate that attention makes a crucial contribution to action control, making its selective deployment a plausible mechanism by which motivational mental states play this key functional role.

3.3.2. Wu on Guidance and Initiation

Such is Wu's suggestion when it comes to motivational mental states' control function. How does he account for their role in action guidance? According to Wu, motivational mental states aid the computation of a suitable concrete movement by significantly constraining the initially enormous space of motor possibilities from which that concrete movement is to be selected. This, in turn, makes it tractable for a distinct subsystem (2011: 60) to compute, for the constrained set of concrete movements, those that minimise the value of some cost function, which typically amounts to computing which movement minimises the value of some (or several) relevant motor parameter(s), such as distance travelled to target, or jerk (2011: 60).

This work of constraining the computational space is purportedly achieved thanks to the representational content of the intention – intending to perform a movement *with your right arm*, for instance, greatly reduces the number of possibilities from amongst which movement must be specified. That is, selection is limited to contraction or relaxation of the muscles in the right arm, and to those muscular activations in the right arm consistent with the general character of the arm movement intended (for instance, the contraction of the bicep in the case of a 'curling' motion). This constraining role can be seen as the abstract specification of an action plan, which is transformed into a concrete movement by a further computational process.

Consequently, this picture is clearly a consistent and *prima facie* plausible explanation of motivational mental states' guiding function.

On the motor side of the Many-Many problem (as opposed to the perceptual side), Wu describes motivational mental states' role *only* in terms of constraining the set of motor possibilities, as described above. That said, it sometimes seems like Wu envisions the outcome of cost-function computation (ideally, the specification of a single concrete movement that minimises the value of the cost-function) as the initiation of action, though the details are never made clear. Wu writes,

> Once the lowest cost movement is identified, the agent's body moves and does so in accordance with his intention (2011: 60)

I think that any plausible interpretation of what Wu is suggesting here leaves him without an account of action initiation. I shall explain this thought in more detail in the following section, but it is worth noting at this point that identifying how Wu's story is supposed to account for action initiation is difficult at best. Given Wu's goals in his 2011 paper, this is an understandable omission. But it does, I shall argue, make it difficult to think of him as providing a complete theory of motivational mechanisms. What Wu has provided on the matter is insufficient to scratch that particular itch.

3.4. Problems for Wu

3.4.1. Problems with initiation

Problems emerge when considering Wu's theory's ability to account for motivational mental states' initiating function. The problem is that after motivational mental states have constrained the set of motor possibilities in the manner Wu describes, all subsequent work is attributed to the computation of cost functions. Either Wu must accept that this includes the work of initiating action, or he must accept that he hasn't explained how action gets initiated at all.

Why should Wu not accept the first disjunct? The quote I provided at the end of section 3.3 certainly made it seem as if he did. The problem is that it is far from clear that Wu can coherently hold this view and simultaneously maintain that the function of action-initiation *per se* is to be attributed to motivational mental states.

Let us assume that when Wu writes,

> Once the lowest cost movement is identified, the agent's body moves and does so in accordance with his intention (2011: 60)

he intends to suggest a causal relationship by the conjunction of 'identification of the lowest cost movement' and 'movement of the agent's body'. That is, let us attribute to Wu the suggestion that identification of the lowest cost movement is causally sufficient for action-initiation. Then it seems that, on Wu's view, intentions and other motivational mental states at most *enable* the initiation of action, rather than actually playing that role (remember: intentions, on Wu's picture, do not identify the lowest cost movement, but merely ease the computational strain of figuring it out).

Of course, this fits with Wu's basic picture of motivational mental states as structuring causes of action. But the lesson to draw from that is simply that Wu's picture is insufficient for a causal theory of action. Arguably the most central claim of such theories, we saw in section 2, was that intentions *actually initiate* action – they do not merely set up some background conditions whereby something else can do so. So, if this is the right way of interpreting Wu, then we must either give up on cashing out motivational mental states' functional roles in the manner of a causal theorist, or conclude that Wu's theory fails to successfully account for motivational mental states' initiating function. Since this paper openly assumes a causal theory of action, I shall not spend much time discussing why we should not give up on it. But it is worth pointing out that adopting the only obvious alternative, a theory on which agents themselves, rather than their motivational mental states, are the generators of action (an agent-causal theory), will not help Wu here either. This is because cost-function computation is typically (and in the work Wu cites) thought of as a sub-personal process of mind. Since such computation is where Wu's theory places the point of action-initiation, the theory is also incompatible with the claim that the initiator of the action is the agent themselves. Instead, the initiator turns out to be a sub-system of which the agent is largely or entirely unaware. Wu is stuck between failing to explain action-initiation, or explaining it in terms of a sub-personal homunculus. Neither is conducive to a convincing theory of human action.

One could object that I have made too substantive an assumption about what is to count as the point of action-initiation. To be clear, I take it that bodily action is initiated when the bodily movement starts. Since on Wu's picture the work of motivational mental states has ended well before movement begins, and while there is still necessary work to be done (since the agent's body moves only after the cost-function computation), it is thus difficult for me to see how motivational mental states could count as having initiated those actions. But perhaps my chosen point is arbitrary. Why could Wu not insist that action-initiation begins at the point where the narrowing down of the motor landscape begins. On this picture, the action-initiating function of motivational mental states would simply be folded into the mechanism by which they play their guidance function. It would no longer be unexplained.

There is a problem with making this kind of move, however. Imagine for a moment that Wu is correct that action is the outcome of a) a process of narrowing down the space of motor possibility before b) computing the value of a cost function over the movements that remain. It is possible that this second process does not always follow the first. Say I intend to pick up my coffee mug, and my intention plays its guidance role in the way Wu thinks it does – it significantly constrains the set of my possible motor responses. Imagine though, that after this process is completed, I change my mind (I realise that the mug is scaldingly hot, perhaps, and so lose my intention to pick it up). What are we to say has occurred in such a case?

I suggest that the obvious reply is that I cancelled the action before it began, or, rather, *prevented it from happening*. I had not begun the action necessary to pick up the coffee mug, before deciding that it was too hot for me to do so. The issue here is that if Wu adopts the view above, in order to dismiss my objection regarding his account's inability to explain how intentions initiate action, he cannot say any such thing. He must instead insist that the above is a case of my action being *interrupted*. Perhaps it is not absurd to say such a thing. But I cannot help but find it deeply counter-intuitive. If I have not yet moved, it should follow that I have not yet started

to perform a (bodily[19]) action. Denying this principle seems to be a significant cost that we should not want to pay.

3.4.2. Problems with guidance

So much for action-initiation. How does Wu's account of action-guidance fare? While he undoubtedly has *some* account of the mechanism underpinning the guidance function, it is worth noting that some of its key elements are increasingly disputed. In particular, the claim that motivational mental states' guidance role is achieved by enabling the efficient computation of cost functions should be treated with suspicion. Many working roboticists (Feldman, 2009 and Mohan & Morasso, 2011), as well as psychologists and philosophers (Friston, 2011; Clark, 2016: 132) have branded explicit cost-function-based solutions to the problem of efficient motor selection inflexible and biologically unrealistic. As Mohan & Marasso put the issue,

> …such engineering paradigms were designed for high bandwidth, inflexible, consistent systems with precision sensors. The difficulty lies in adapting these models to the typical biological situation, characterized by low bandwidth, high transmission delays, variable/flexible behavior, noisy sensors, and actuators. (2011: 3)

Consequently, we should be suspicious of whether, despite their explanatory success, models involving the explicit and resource-intensive process of cost-function calculation are likely to reflect what is actually going on in complex biological systems such as ourselves. There is, for instance considerable doubt in the literature regarding whether the impressive formal solutions to many *prima facie* problems faced by these models can actually be efficiently implemented through distributed neural networks[20] (Todorov, 2006; Mohan & Morasso, 2011: 3). One upshot of these concerns ought to be suspicion of the claim that motivational mental states play their guidance function by enabling the activity of these biologically dubious constructs. If our actions simply *are not* the outcome of cost-function computations, then there is

[19] Naturally, I do not think that movement need have begun before we can say a *mental* action has begun (no movement may be associated with such actions at all), but this is beside the point.

[20] The reasonable assumption being that such models are our best guess at how the brain 'computes', to the extent that it does at all.

no reason to think that our motivational mental states operate in such a way as to simplify them.

Let us say, however, that it could be demonstrated that these models were, as far as we can tell, biologically implementable. There is still a further problem. Wu's proposed solution presumes that the representational content of intentions is sufficiently detailed so as to narrow the space of motor possibilities to a degree that makes the computation of a cost function *tractable* for each possible residual motor behaviour. Maybe this is so, but there are good reasons to doubt it. For instance, the initial set of possible concrete motor behaviours numbers in the region of 2^{600} (Wolpert & Ghahramani 2000). Solving a cost-function for even a tiny portion of that space is phenomenally computationally expensive. Consequently, intentions would need to be having an *enormous* constraining effect on this behavioural space for our actions to proceed at the rate at which they typically do. It is not clear, given the characteristically coarse-grain at which we think of the content of our intentions (and other motivational mental states), that this is a safe assumption. At best it is a hostage to empirical fortune. At worst, it is downright implausible.

There is no knock-down argument against Wu's proposal here (or even in the vicinity), absent further empirical study of action-generation. But it is possible to make sensible theoretical decisions in the meantime. In particular, an alternative theory that did not posit the computation of cost functions as a key mechanism would have the potential to appear significantly more biologically realistic, and less of a hostage to empirical and computational fortune. By extension, a theory of motivational mental states' guidance function that did not place them at the centre of an overly computationally complex and biologically unrealistic process should, I think, currently be preferred over one that does, *ceteris paribus*.

In section 3.6, I argue that what I call the predictive theory of motivation satisfies this requirement, as well as offering an account of motivational mental states' role in action-initiation and preserving Wu's insights on the topic of control. Thus, I argue, it should be preferred to Wu's. First, however, I offer a brief sketch of the predictive processing framework that will give the necessary background to understanding the proposed theory.

3.5. Predictive Processing: An Overview

In this section, I outline the key features of the predictive processing (PP) research paradigm in Cognitive Science. I do not attempt a systematic defence of the framework. I take that it has been well enough articulated, philosophically defended (see Clark 2016; 2013a; 2013b; Hohwy 2013), and empirically supported (Talsma 2015; Barrett & Simmons 2015; Seth 2013; Adams et al 2013) to warrant at least being taken seriously. Instead, I explain the framework here, and make use of it in the articulation of the predictive theory of motivation in section 5.

The central concept that the theory of PP brings to the table is **prediction error**. This, roughly, denotes a signal hypothesised to be produced when the brain's predictions about the sensory signal (broadly construed to include exteroceptive, interoceptive, and, crucial for our purposes later on, proprioceptive signals) fail to match the actual sensory signal. The overall structuring principle of the human mind, according to PP, is the minimisation, over the long-term, of prediction error signals.

Let us call the brain's best current guess regarding the source of the incoming sensory signal (i.e. its causal origins) the **world model**. On the basis of the **world model**, the brain produces a **generative model**; "a flow of virtual or mock sensory signals that predicts the…sensory signals generated by external[21] causes" (Gładziejewski 2016: 562). The generative model is hierarchical in nature; the highest level of the generative model is generally thought to encode relatively abstract predictions about the expected state of the world, which unfold as the hierarchy descends into relatively concrete, precise predictions about the expected character and intensity of particular sensory signals. For instance, a high-level prediction that an elephant is very close in front of you might unfold into particular visual (e.g. grey, textured), auditory (e.g. trumpeting, stomping), and olfactory (e.g. strong, muddy) predictions at the lower levels of the hierarchy.

[21] In fact, this qualifier does not always apply to the predictions of the generative model; it is crucial to the proposal I will make later in the paper that the sensory signals it is predicting do not always have external (at least in the sense of 'outside the boundaries of skin and skull') causes.

Those aspects of the *incoming* sensory signal that are successfully predicted by the generative model are dismissed – that is, they have no further effect on any aspect of the neural economy of which they were a part. They have been, to use a term of art, successfully 'explained away'. Those aspects that are *not* explained away, on the other hand, generate a further upward-flowing signal. This, explained slightly more precisely this time, is the prediction error.

The function of prediction error signals is to drive mental processes that have the effect of minimising future prediction error. At any given time,

> …the cognitive system attempts to settle on a "hypothesis" about their [the sensory signals'] causal origins, namely that which has the highest posterior probability (i.e. the probability of being true in light of the data) among alternatives, given its likelihood…and prior probability… (Gładziejewski 2016: 561-562)

The result of this process of 'settling' on a hypothesis (that is adopting a hypothesis that generates minimal prediction error), according to PP, corresponds to what we perceive, believe, do, and so on. The reason why the process of settling can produce all of these features of human mentality is that there are, PP advocates are keen to point out (Clark 2013; Friston 2011), *two ways* for embodied, active neural systems such as ourselves to minimise prediction error. The first way is for generated prediction error to drive updates to the world model (and hence the generative model). That is, the brain engages in a process of (roughly) Bayesian inference to the best explanation of the prediction error signal, and adopts a new world model that (when all is going well) produces a generative model that results in less prediction error than before. Call this **passive inference**.

The second way operates in the other direction; instead of driving internal changes to the world model, generated prediction error drives changes to the source of the sensory signal (i.e. the world), in order to bring it more in line with the predictions of the generative model. That is, generated prediction error can cause us to *act* on the world in a way that alters the sensory signal so as to bring it more in line with what is predicted. Call this **active inference** (Clark, 2013a; Friston, 2010).

We should not, however, expect these processes to entirely eliminate prediction error, even if we suppose the world model is a perfect representation of the distal causal structure of the world the agent is currently occupying. This is because the typical sensory signal (generated as it is by imperfect sensory transducers) is *noisy* – that is, the signal will typically partially misrepresent the state of the world.

This poses a problem. Given that we typically get around in the world successfully, it seems like our brains are, for the most part (though certainly not always) able to distinguish between prediction error that it *needs to do something about* and that which it does not. What might allow the brain to distinguish between prediction error born of mismatch between the world model and reality, and that born of mismatch between the world model and sensory noise?

The answer involves postulating further predictive processes. As well as generating predictions regarding incoming sensory signals, the generative model also produces predictions regarding the expected quality of the incoming sensory signal (i.e. how noisy/clean it is) as well as its own predictions/mock sensory signal. This is known as the **expected precision** of the signal. Various contextual features, as well as past experience regarding certain kinds of sensory signals, are used to calculate a particular signal's expected precision (Clark 2013a; 2013b; 2016; Gładziejewski 2016: 562). The proposed details need not concern us here. The crucial point is that both active and passive inference are significantly more likely to be driven by prediction errors driven by *high-precision* sensory signals or predictions of the generative model. Prediction errors resulting from *low-precision* signals/predictions have significantly less power to drive update or action.

If a sensory signal has low expected precision, as might occur, for instance, in the case of visual signals on a very foggy day, it will not drive processes of perceptual or active inference. In the PP worldview, the process of modulating the expected precision of the sensory signal (and thus the impact of related prediction errors on the brain's inferential processes) in this way, corresponds (at least roughly) to the notion of *attention* as it is understood in psychology and neuroscience more widely.

Having explained the basics of predictive processing, in the next section I present the predictive account of motivation, and suggest that it is to be preferred to Wu's story, according to the standards previously identified.

3.6. The Predictive Account of Motivation

As we have seen, according to the predictive processing framework, action (or active inference) is just another manner by which the brain-body system minimises overall prediction error. One central upshot is that action is the outcome of basically the same processes as perception, or belief. But it is important not to jump from this observation to the claim that there is no significant difference between motivational mental states and motivationally inert ones. As Shea notes (2013) we still have to account for why certain prediction errors are minimised passively (i.e. by revising the predictions of the generative model), and why others are minimised actively (i.e. via environment-altering action). To put the point another way, the mere possibility of action is insufficient; we need the predictive processing framework to give us a plausible story of how action *actually* comes about, not merely to gesture in the general direction of why the occurrence of action is *compatible* with the overall story being told.

I take it that the complete story is yet to be told. But friends of the predictive processing story can certainly point out one kind of situation where the model predicts that agents will act, rather than revise their generative model. This is one in which one or more of the highest-level predictions of the generative model are assigned arbitrarily high precision in comparison to the incoming sense data with which they might conflict (Klein 2018). Any prediction errors generated by conflict between such predictions and incoming sense data *must be* resolved actively, because the degree of precision assigned to the prediction of the generative model renders it *effectively un-revisable*. Intuitively speaking, we can think of this as a situation in which the prior probability of the prediction is judged so high that no realistic amount of conflicting sense data can drive the highest levels of the cognitive hierarchy to revise it. In such a situation, passive inference is not an option for minimising prediction error. Since the overarching goal of the system to minimise

prediction error is still in place, the only remaining possibility is active inference (i.e. action) to make the relevant prediction true.

The central claim of the predictive processing account of motivation I will be putting forward here is that some kind of dependence relation[22] holds between these effectively un-revisable predictions of the generative model and our motivational mental states. Hence we have moved from the mere possibility of action, to actual action; motivational mental states actually cause action, and are hence distinguished from motivationally inert states, because they generate prediction errors that can *only* be minimised through active inference. I also hold that this theory of motivational mental states can more successfully account for the functional roles of motivation than Wu's story.

In order to obtain this result, however, I must add one more posit to the theory. We are currently working on the supposition that motivational mental states depend closely on the activity of the highest-level predictions of the generative model. We must say one more thing about these predictions; they decompose (at lower levels of the hierarchy) into at least two kinds of prediction; *high-precision* proprioceptive predictions and (perhaps surprisingly) *low-precision* exteroceptive predictions. The proprioceptive predictions can be thought of, intuitively, as predictions of the expected proprioceptive consequences of actually acting in the present situation so as to satisfy the content of the motivational mental state. So, for instance, if I presently intend to pick up my coffee cup, the associated high-level prediction that I *will* pick up my coffee cup gives rise (at a lower level of the cognitive hierarchy) to a high-precision prediction of the proprioceptive consequences of actually moving my arm in the relevant direction and performing a grasping motion. The exteroceptive predictions can be thought of, intuitively, as predictions of the location and (action-relevant) qualities of task-relevant features of the external environment. So, continuing with the same example, the associated high-level prediction gives rise (at

[22] I refrain from speculation regarding the exact nature of this dependence relation. It could turn out to be identity, determination, ground, supervenience, or perhaps something else. Nothing that I have to say in what follows requires me to take a position on this question.

a lower-level of the cognitive hierarchy) to a low-precision prediction of the ego-centric location and circumference of the coffee cup (amongst other features).

To summarise; with the following two posits, and the background assumption that the cognitive system functions so as to minimise overall prediction error, I believe I can offer a more satisfying account of the functional roles of motivational mental states than Wu can.

> 1) A dependence relation holds between motivational mental states and high-level, effectively un-revisable predictions of the generative model.
>
> 2) Effectively un-revisable predictions of the generative model decompose at lower levels of the hierarchy into a) *high-precision predictions of the proprioceptive consequences* of acting so as to satisfy the motivational mental state in the current situation and b) *low-precision predictions of the location and (action-relevant) qualities of task-relevant features* of the external environment.

I will now detail the proposed explanation of each of the central functional roles of motivation in turn.

3.6.1. Initiation

We saw in section 3.4 that Wu's account failed to offer any satisfactory account of the initiation function of motivational mental states. The central problem there was that, according to Wu's story, no bodily movement begins until well after the role of the motivational mental state has been completed, and the operation of a distinct subsystem has begun. Thus we cannot say that the motivational mental state *initiates* the action in question, and thus there is no explanation of this central functional role of motivational mental states.

According to the predictive theory, initiating an action is the consequence of an effectively un-revisable prediction that the change to the world associated with the action will actually occur. For instance, if I intend to pick up my coffee cup, then the initiation of my picking up the coffee cup is an unavoidable consequence of the generative model's effectively un-revisable prediction that the cup is (or will shortly be) in my hand.

Unlike Wu's story, wherein motivational mental states, at best, provide the necessary cognitive background against which an action *may* be performed, here motivational mental states determine, ground, or are identical to an effectively un-revisable prediction that *guarantees* active inference (i.e. action). Because the prediction is effectively un-revisable, and because the brain consistently functions so as to minimise overall prediction error, in such a situation bodily movement *must* occur (naturally, the bodily movement in question may nevertheless fail to actually satisfy the agent's goal). To put it another way, the stable presence of a motivational mental state is, in this context, sufficient to ensure that action is attempted, since there is no other way for the cognitive system to minimise prediction error generated by the following conflict; the prediction that the coffee cup is in my hand, conflicting with the incoming sensory information that it currently is not.

Naturally, the guarantee of action in this situation is not a matter of metaphysical necessity. It is, rather, something like a matter of *cognitive necessity*; given the general principles governing the operation of human cognition (according to the predictive processing view), the presence of an effectively un-revisable prediction is sufficient for a bodily movement to be initiated that will attempt to make the world such that it conforms to the predicted state of affairs.

Note that this account allows us to preserve the intuition that in the case where I initially intend to pick up my mug, but then cease to do so before any movement has begun because I recognise it is too hot to safely touch, I have *prevented* rather than *interrupted* an action. This is because the sensory information that the mug is hot can be expected to revise down the precision of the prediction that I will shortly have it in my hand. That is, the recognition that the mug is hot functions to render the prediction that the mug will be in my hand *revisable* (indeed, to force its revision). The persistence of an effectively unrevisable prediction guarantees active inference; such a prediction's failure to persist can curtail it.

3.6.2. Guidance

As we also saw in section 3.4, Wu's account offered an unsatisfactory account of motivational mental states' guidance function. The central problem there was that Wu's story depended on a computational process of simplifying the minimisation of

some complex cost function. This sort of process turned out to be biologically unrealistic, and computationally expensive, in a manner that made it look implausible as a hypothesis regarding the actual processes underpinning motivational mental states' guiding function.

According to the predictive theory, motivational mental states are closely tied to effectively un-revisable predictions of states of affairs that decompose into (amongst other things) high-precision predictions of the proprioceptive consequences of bringing those states of affairs about. As Clark writes,

> …motor control is, in a certain sense, *subjunctive*. It involves predicting the non-actual proprioceptive trajectories that would ensue were we performing some desired action. (2016: 121)

This is how motivational mental states satisfy their guidance function, or so I shall suggest. Instead of first restricting the enormous set of possible motor trajectories, and then computing a cost-function over this set to select the least costly concrete behaviour, the predictive theory proposes that the motor trajectory is already specified directly by the associated proprioceptive prediction. In other words, predicting the proprioceptive consequences of bringing about a certain state of affairs already amounts to having selected an executable action plan. This is because distinct trajectories are also proprioceptively distinct. A prediction of the proprioceptive consequences of a particular movement specifies a unique trajectory for such a movement (i.e. the trajectory that will result in exactly the flow of proprioceptive sensory information that is predicted). As Friston writes,

> …in active inference, these problems are resolved by prior beliefs about the trajectory (that may include minimal jerk) that uniquely determine the (intrinsic) consequences of (extrinsic) movements…. (2011: 496)

And as Clark boldly states,

> It is easy…to specify whole paths or trajectories using prior beliefs about (you guessed it) paths and trajectories! (2016: 130)

None of this, it should be pointed out, reduces the *overall* computational complexity of selecting from the intimidatingly large set of possible motor trajectories (Clark

2016: 130; Friston 2011: 492). It simply pushes the problem back to the acquisition of an effective model, which is good at predicting *actionable* proprioceptive consequences of bringing about certain states of affairs. This is certainly not a trivial task.

This solution is, however, a more realistic prospect than Wu's in at least one important way. On Wu's proposal, the problem must be solved in real-time, on each occasion the agent is motivated to act. On the predictive theory, on the other hand, this problem is being constantly solved, on each occasion the priors of the generative model for a given situation are updated in response to prediction error. That is, the time given over to solving this problem of serious computational complexity is considerably greater on the predictive theory than it is on Wu's. This should lead us to think that it will be significantly more tractable on the predictive theory than on Wu's. This, I suggest, is a significant advantage of the predictive theory, which is otherwise able to do all the work of Wu's account of guidance.

In fact, this story about guidance does considerably more than Wu's, in a way that one might initially find troubling. Recall that when defining the notion of action-guidance, I noted that the most common way of thinking supposed that motivational mental states did not specify entire trajectories, but rather abstract action plans, satisfiable by many different concrete movements. That is, the action-guiding role of motivational mental states does *not*, it is thought, involve solving the many-one mapping problem of motor control.

Clearly, on the predictive account, the proprioceptive predictions achieve precisely this. They *uniquely* specify a particular motor trajectory, by way of such a trajectory's proprioceptive consequences. Indeed, the predictive story explains by way of one process that which Wu is forced to explain by way of two; while he must appeal to both motivational mental states' restricting the class of concrete actions and a cost-function computation over the remaining options, both of these processes are taken care of by high-precision proprioceptive predictions on the predictive account.

Whether or not this is a problem, I think, largely depends on how one looks at it. If you think that one's explanation of motivational mental states' role in action-

guidance should not, in principle, imply that they also solve the many-one mapping problem, then this will strike you as a problem. But it seems likely that this restricted view on the limits of the action-guidance role was a product of precisely the kind of view that Wu advocates; one on which it was unclear *how* a motivational mental state could have a sufficiently fine-grained content to be able to solve the many-one mapping problem. Since it is now less mysterious how this could come about, we should not insist from the start that the two-step solution is the right one. Thus we should not accept the restricted role of motivational mental states in action-guidance that such a view suggests.

One may well think, of course, that motivational mental states still *do not* have such precise content even on the view I am advocating. Effectively un-revisable predictions of the generative model needn't themselves have these high-precision proprioceptive predictions as content. What *is* true is that (based on past experience of such situations) they cause a downward flow of increasingly precise predictions that 'unpack' the likely consequences of that prediction, including the specific expected proprioceptive signal. Whether you should see this as amounting to the motivational mental state itself (or at least the effectively un-revisable prediction to which it bears some kind of dependency relation) having the relevant proprioceptive content, or simply causing a lower-level state that does, is unclear. But whichever one opts for is, for our purposes here, somewhat inconsequential. The necessary work gets done either way.

3.6.3. Control

Recall that I endorse Wu's insights on how motivational mental states play their action-controlling function. The key idea there was that paradigmatic cases of action-control being undermined were ones in which there was a failure of attention; it was not appropriately directed towards task-relevant external stimuli. In this final substantive section, I shall argue that the predictive theory retains this insight, and thus gives at least as good an account of motivational mental states' role in action-control as Wu's.

The predictive theory retains this insight in the following way. Recall that on the predictive theory, effectively un-revisable high-level predictions of the generative

model decompose at lower levels of the cognitive hierarchy into (amongst other things) low-precision exteroceptive predictions of the location and (action-relevant) qualities of task-relevant features of the external environment. This, I suggest intuitively implies a recalibration of attention towards the relevant features of the external environment, in a manner that allows sensory information to drive updates in the world model. The reason is that since the predictions are of low-precision, conflicting sensory information is significantly more likely to be treated as data, not noise.

The central point here is that if the predictive theory is correct, then motivational mental states generate exteroceptive predictions about action-relevant features of the task environment that are especially prone to update in light of conflicting sensory information. Thus throughout the execution of the action, the agent's world model is especially susceptible to being updated in line with the evidence of their senses. To put it simply, motivational mental states perform their action-controlling function by opening up the generative model to being revised by the task-relevant features of the environment. This seems to be a predictive twist on Wu's suggestion that the deployment of attention towards task-relevant features of the environment allows the online monitoring and modification of bodily behaviour. Incoming sensory information is permitted to 'correct' mistaken predictions about an object's location or other action-relevant features (because these predictions are assigned low-precision), which will feed into the details of the associated proprioceptive predictions. Moreover, the fact that the low-precision exteroceptive predictions concern action-relevant features of the present environment means that *those properties* rather than any others are the ones being 'monitored'. Thus the predictive theory can retain Wu's insights into action control.

3.7. Conclusion

I have argued that any account of how motivational mental states cause action must account for how they play three specific functional roles; intitiation, guidance, and control. I have argued that a significant extant account of motivation, due to Wu (2011; 2016), offers reliable insight only into how the latter of these three roles is satisfied. I then proposed a different theory of motivation drawn from the predictive

processing paradigm (the predictive theory) and argued that it was able to explain how motivational mental states perform all three functional roles in quite some detail. Thus I conclude that it should currently be preferred over Wu's.

It remains to be seen whether there might emerge accounts, including perhaps some revised version of Wu's, that are able to offer more complete explanations of the functional roles identified as centrally important here. It is also an open question whether an adequate theory of motivational mental states must explain more than just these three functional roles. And, finally, any such minimally adequate theory must be subject to empirical test. That said, what I have proposed here is, I think, a significant step forward, as well as being more grist for the predictive processing mill.

3.8. References

Adams, R.A. Shipp, S. & Friston, K.J. (2013). "Predictions not commands: Active inference in the motor system". *Brain Structure and function 218*(3): 611-643

Bach, K. (1978). "A representational theory of action". *Philosophical Studies 34*(4): 361-379

Bain, D. (2011). "The imperative view of pain". *Journal of Consciousness Studies 18*(9-10): 164-185

Barrett, L. & Simmons, W.K. (2015). "Interoceptive predictions in the brain", *Nature Reviews Neuroscience 16*: 419-429

Bratman, M. (1987). "Intention, plans, and practical reason". Cambridge MA: Harvard University Press

Chisholm, R. (1966). "Freedom and action". in Lehrer, K. (ed.) "Freedom and Determinism". New York: Random House

Clark, A. (2013a). "Whatever next? Predictive brains, situated agents, and the future of cognitive science", *Behavioral and Brain Sciences 36*(3): 181-204

Clark, A. (2013b). "Expecting the world: Perception, prediction, and the origins of human knowledge". *The Journal of Philosophy 110*(9): 469-496

Clark, A. (2016). "Surfing uncertainty: Prediction, action and the embodied mind". Oxford: OUP

Corns, J. (2014). "The inadequacy of unitary characterizations of pain". *Philosophical Studies 169*(3): 355-378

Dancy, J. (1993). "Moral reasons". New Jersey: Blackwell

Dretske, F. (1993). "Mental events as structuring causes of behavior". in Heil, J. & Mele, A.R. (eds.) "Mental Causation". Oxford: OUP

Elliott, A.J. (2006). "The hierarchical model of approach-avoidance motivation". *Motivation and Emotion 30*(2): 111-116

Feldman, A.G. (2009). "New insights into action-perception coupling". *Experimental Brain Research 194*(1): 39-58

Frijda, N.H. (1986). "The emotions". Cambridge: CUP

Friston, K. (2010). "The free-energy principle: A unified brain theory"

Friston, K.J. (2011). "What is optimal about motor control?". *Neuron 72*(3): 488-498

Gładziejewski, P. (2016). "Predictive coding and representationalism". *Synthese 193*(2): 559-582

Hohwy, J. (2013). "The predictive mind". Oxford: OUP

Klein, C. (2018). "What do predictive coders want?". *Synthese 195*(6): 2541-2557

McDowell, J. (1979). "Virtue and reason". *The Monist 62*: 331-350

McVay, J.C. & Kane, M.J. (2009). "Conducting the train of thought: working memory capacity, goal neglect, and mind wandering in an executive-control task". *Journal of Experimental Psychology: Learning, Memory, and Cognition 35*(1): 196-204

Mele, A.R. (1992). "Recent work on intentional action". *American Philosophical Quarterly 29*(3): 199-217

Mohan, V. & Morasso, P. (2011). "Passive motion paradigm: An alternative to optimal control". *Frontiers in Neurorobotics 5*: 4

Mrazek, M.D. Franklin, M.S. Phillips, D.T. Baird, B. & Schooler, J.W. (2013). "Mindfulness training improves working memory capacity and GRE performance while reducing mind wandering". *Psychological Science 24*(5): 776-781

Nanay, B. (2013). "Between perception and action". New York: OUP

O'Connor, T. "Agent-causal theories of freedom". in Kane, R. (ed.) "Oxford Handbook of Free Will" (2nd ed.). Oxford: OUP

Pacherie, E. (2012). "Action". in Frankish, K. & Ramsey, W. "The Cambridge Handbook of Cognitive Science". Cambridge: CUP. 92-111

Scarantino, A. & Nielsen, M. (2015). "Voodoo dolls and angry lions: How emotions explain arational actions". *Philosophical Studies 172*(11): 2975-2998

Searle, J. (1983). "Intentionality: An essay in the Philosophy of Mind". Cambridge: CUP

Seth, A. (2013). "Interoceptive inference, emotion, and the embodied self". *Trends in Cognitive Sciences 17*(11): 565-573

Shea, N. (2013). "Perception versus action: the computations may be the same but the direction of fit differs". *Behavioral and Brain Sciences 36*(3): 228-229

Shepherd, J. (2014). "The contours of control". *Philosophical Studies 170*(3): 395-411

Shepherd, J. (2017). "Halfhearted action and control". *Ergo 4*(9): 259-277

Smallwood, J. Beach, E. Schooler, J.W. & Handy, T.C. (2008). "Going AWOL in the brain: Mind wandering reduces cortical analysis of external events". *Journal of Cognitive Neuroscience 20*(3): 458-469

Smith, M. (1994). "The moral problem". Oxford: Blackwell

Talsma, D. (2015). "Predictive Coding and multisensory integration: an attentional account of the multisensory mind", *Frontiers in Integrative Neuroscience 9*(19)

Tappolet, C. (2010). "Emotion, motivation and action: the case of fear". In Goldie, P. (ed.) "Oxford handbook of Philosophy of Emotion". Oxford: OUP

Todorov, E. (2006). "Optimal control theory", in Doya, K. Ishii, S. Pouget, A. & Rao, R.P.N. (eds.) "Bayesian brain: Probabilistic approaches to neural coding". Cambridge MA: MIT Press

Wolpert, D.M. & Ghahramani, Z. (2000). "Computational principles of movement neuroscience". *Nature Neuroscience 3*: 1212-1217

Wu, W. (2011). "Confronting many-many problems: Attention and agentive control". *Nous 45*(1): 50-76

Wu, W. (2016). "Experts and deviants: The story of agentive control". *Philosophy and Phenomenological Research 93*(1): 101-126

Chapter 4: Explaining Agential Pathology in Clinical Depression

4.0. Abstract

In this paper I introduce the phenomenon of agential pathology, a characteristic symptom of depressive illness, and evaluate two existing classes of theory that try to explain why it occurs (which I term mental state accounts and somatic accounts). I propose three phenomenological constraints on successful explanation of agential pathology, and find that neither of these two classes of theory, nor a further class of theory that I motivate independently (perceptual accounts), are able to satisfy all three of these constraints. I propose instead that we should adopt a hybrid theory of agential pathology, and sketch some proposals of this kind that are motivated by the foregoing explanatory concerns.

4.1. Introduction

People with depressive illnesses (that is, those in which depressed mood features as a central criterion for diagnosis, including Major Depressive Disorder and Bipolar Disorder) often report and exhibit difficulties in initiating and sustaining day-to-day action (Oyebode 2015: 294-298). This symptom can range in severity in particular cases from a mild or moderate increase in perceived effort when acting, to a degree of apathy, or lethargy, all the way through to near-total catatonia (Starkstein et al 1996).

Femi Oyebode distinguishes between diminution of *motivation*, characterised as a "hypothetical [action-]activating factor" (2015: 294) and *will*, thought to encompass the capacity to make choices or express preferences (Jeannerod, 2006, cited in Oyebode, 2015: 296). Both of these subtly different phenomena are typically thought to be exhibited in some cases of depression and not in others, though disturbances to motivation are more typical than disturbances to will (298). Oyebode suggests that it may be difficult to distinguish them, remarking that "[t]he observable end result [of both these impairments] is lack of action in the absence of any motor abnormality impairing action". I take it that by 'absence of any motor abnormality', Oyebode intends to exclude clearly organic causes of action-impairment, such as may result from peripheral or central nervous system damage, amongst other things. It is this characteristic kind of phenomenon that I refer to in the rest of this paper as *agential pathology*. That is, my working definition of agential pathology is somewhat negative; the impairment of action, not resulting from organic motor abnormality, characteristic of depressive disorders.

The main goal of this paper is to examine the different suggested explanations of agential pathology that have been put forward in the philosophical psychopathology literature. Specifically, it seeks to make moves towards answering the question of why people with Depression fail and/or struggle to act, by evaluating the explanatory power of several different hypothesised psychological impairments. The current research landscape on this question in philosophical psychology can broadly be divided into two competing camps. The first explains agential pathology by some hypothesised change in the folk mental states of the person with depression; their

beliefs, desires, intentions, and so forth (Kuhl & Helle 1986; Smith 1994; Law, 2009). Call these *mental state accounts*. The second explains it by a hypothesised disturbance in the experience of the bodily feelings of the person with depression (Smith 2013; Ratcliffe 2015). Call these *somatic accounts*.

Even though there is significant variation within these camps, I argue that neither of them has the explanatory power necessary to account for the relevant phenomena. That is, I believe there are important desiderata for an explanation of agential pathology that neither can satisfy alone. Nor, I claim, can a third class of theory, which explains agential pathology by a hypothesised change in a depressed person's perception (I motivate such a theory independently). My conclusion will be that we should not seek a single-factor explanation of agential pathology. Rather, we should endorse a broad and inclusive pluralism regarding the possible underlying causes of agential pathology. All minimally plausible theories of agential pathology require some combination and integration of the theories identified above. I shall also argue that all plausible versions of such a hybrid theory will identify dimensions of agential pathology that are, properly understood, disorders of neither motivation nor will. This runs counter to Oyebode's characterisation of the phenomenon (2015).

While some authors have seemed to assume that a single-factor explanation might sufficiently explain the phenomenon of agential pathology (e.g. Smith 1994; Law 2009), others have explicitly noted that they believe or suspect their suggestions to only partially explain the relevant data (Smith 2013). Consequently, the recognition of the need for some kind of explanatory pluralism is not novel to this paper. To my knowledge, however, nobody has systematically investigated a) what the minimal explanatory demands of a theory of agential pathology are or b) what the nature and scope of the relevant explanatory pluralism would need to be in order to satisfy these demands. My goal in this paper is to do both of those things. That said, I do not purport to be developing a complete theory here. I claim simply that the explanatory requirements that the theory presented here satisfies are ones that *any* theory of agential pathology must satisfy, and that it is at best unclear how a single-factor theory of any kind could possibly do so. I suspect, however, that these requirements will turn out to be insufficient.

A quick note regarding scope; in this paper I focus on those explanations that operate at a distinctively psychological level. I do not aim to assess the plausibility or explanatory power of neurobiological accounts of agential pathology. I suspect that such a survey would conclude by proposing that agential pathology is similarly multi-factorial at the neurobiological level, though I cannot rule out the possibility that the multiple psychological factors I highlight here may have a single kind of neurobiological ground. Nor do I intend to examine explanations of agential pathology that make ineliminable reference to social or interpersonal variables and factors. Again, I strongly suspect that doing so would reveal that the phenomenon involves multiple different kinds of factors at this level of explanation as well, though I cannot rule out that the range of psychological factors covered here may be the outcome of a single kind of social factor; for instance social defeat (e.g. Björkqvist 2001).

Next, I shall explain the desiderata that I shall be using to assess the explanatory power of different theories of agential pathology. After that, I will begin my survey of the existing suggestions in sections 4.2 and 4.3, before presenting my own suggestion in section 4.4. Finally, in section 4.5 I shall present my case for the claim that agential pathology involves numerous psychological factors.

4.1.1. Minimal desiderata

In this section, I shall describe and briefly justify some minimal adequacy conditions on any theory of agential pathology[23]. The motivation for these constraints is as follows. Research into the experience of agential pathology reveals that it has certain peculiar aspects to its phenomenology. That is, agential pathology in depression is not best thought of as being like *any old* struggle in initiating and sustaining action; the struggle has a broadly specifiable, if still somewhat variable, phenomenal character. The three constraints I impose here are intended to reflect the fact that this

[23] I believe that the requirements I describe here are sufficient to rule out any single-factor explanation currently advanced in the literature. Moreover, I think they are sufficient to evidence the need for pluralism in our theorising going forward. I do not, however, claim that a theory satisfying them would thereby be complete.

peculiar phenomenal character is in need of explanation, rather than merely the generic end result.

4.1.1.1. The Impossibility Constraint

Firstly, theories must accommodate the fact that those who experience agential pathology report that action *feels* difficult or impossible – not merely that they do not want to act or feel no satisfaction in completing their goals (Ratcliffe, 2015: 156). In Iain Law's terms,

> every job seems bigger and harder. Every setback strikes me not as something easy to work around or get over but as a huge obstacle…I perceive [my abilities] to be far more meagre than I did when I was not depressed. (2009: 355)

Call this the *impossibility constraint*. Any plausible theory of agential pathology must explain why agential pathology is accompanied by a feeling that tasks are considerably more difficult than they were for that person prior to the onset of agential pathology. Note that, although the basic idea here is one of a feeling, rather than an explicit judgment, that tasks are impossible, it is arguable that a judgment that tasks are impossible would naturally lead to a feeling that they are, and perhaps vice versa. Certainly, if I sincerely judge that tasks are impossible in my day-to-day life, consideration of those tasks seems to take on a distinct phenomenal character that could easily be described as a feeling of impossibility. I will therefore assume in what follows that if a theory can explain where a *judgment* of impossibility may emerge from, then it has done enough to satisfy the impossibility constraint.

4.1.1.2. The Lethargy Constraint

Secondly, theories must explain the sense of embodied effort, lethargy, and heaviness, associated with trying to act in cases of agential pathology. A number of participants in a qualitative study of depression describe their bodies as feeling "tired and lethargic", "heavy, unresponsive", or even "leaden" (#308, Q4; #21, Q4; #137, Q4, cited in Slaby Paskaleva & Stephan, 2013), and report that this renders action difficult; even painful (Slaby, Paskaleva & Stephan 2013). Any viable theory of

agential pathology must provide a plausible explanation for why agential pathology is strongly associated with this kind of bodily discomfort or weakness. Call this the *lethargy constraint*.

Naturally, it seems that any theory that satisfies the lethargy constraint is also likely to satisfy the impossibility constraint (as bodily leadenness presumably explains why everyday tasks are experienced as being difficult), though not vice versa. But this does not mean that the impossibility constraint is redundant. As we shall see in section 2, the fact that some theories may satisfy the impossibility constraint, but not the lethargy constraint, allows us to better diagnose their weaknesses, and thus make theoretical progress. By simply eliminating the impossibility constraint altogether, certain theoretical mistakes remain obscured.

4.1.1.3. The Practical Significance Constraint

Finally, theories must account for changes in the apparent action-relevance of object and features of the world typical of cases of agential pathology. This phenomenon is less well-documented than either the sense of impossibility or lethargy, but such changes are reported in first-personal descriptions of agential pathology. Typically, the relevant changes are related to the degree or manner in which objects and features of the world are experienced as useable, accessible, attractive, or otherwise available for action. For instance,

> When you're depressed it feels as though there is a huge distance between you and things, which are inert, unresponsive to your wishes. Now that I was feeling better, a pen would leap into my hand, soap seemed to cover me of its own accord… (Lewis 2006: 225).
>
> I want to reach out to the world, but it isn't there to reach out to… (Benson et al 2013: 65)

In section 4, I shall spell out in detail how I think the notion of 'distance', evident in each of these quotations, ought to be understood in these sorts of cases. For now, however, I shall simply highlight a few features of these reports deserving of our attention.

In these descriptions, the world and the objects in it are reported to be *unresponsive* to attempts to interact with them. The apparent inertness of objects is an impediment to making effective use of them, and contrasts with the experience of being well, where these objects can be used effortlessly. A natural interpretation of these passages is that the authors are reporting a change in certain ways in which objects appear to be related to the agent's practical concerns. When an agent is depressed, objects and the world they occupy somehow 'recede' from them, practically speaking. Despite *in fact* being relevant to the fulfilment of certain wishes the agent has, they are experienced passively, possessing no (or at least much less) *practical significance* to the agent's concerns. They do not seem accessible, available, or inviting for use in satisfying the agent's wishes (or at least they seem much less so than when the agent was well).

Benson and colleagues identify this experience in their work on the feeling of being suicidal (2013), where most participants were diagnosed with some form of depressive disorder. They argue on the basis of extensive in-depth interview data, that in such cases, an agent commonly experiences a near-total disappearance of the sense that she is "integrated with her environment in the relationship of reciprocal action that is associated with ordinary living." (2013: 65). Agents instead report a feeling of inefficacy, "of physical distance having been introduced between the self and the world" (65), of being "unable to make any impression on the world whatsoever" (65), and even of being stuck behind some kind of physical barrier, such as "a 'bubble', 'glass wall', 'glass box', or 'glass coffin'" (65-66). This is readily interpretable as a generalisation or extreme variant of the kinds of relatively limited experiences described above; *all* or nearly all objects and features of the world have ceased to have practical significance to the agent, and so they feel constrained, or even trapped, with no resources for acting on the world seeming available to them.

A similar experience is identified by Slaby and colleagues in their description and analysis of Agential Pathology, where they suggest that, for people with depression the world seems to be "highly inaccessible, from a practical, active point of view" (Slaby, Paskaleva & Stephan 2013: 44).

Note that the disturbances to practical significance encountered in Agential Pathology are, in an important sense *unmediated*. The world, and its objects and

features, simply seem to be lacking the availability or invitingness that they once had. This may persist despite an agent knowing or judging, in some sense, that the items *are* useable in this way. A depressed person who struggles to take a shower in the morning typically knows full well that their shampoo is available and useful to them in satisfying this goal, but nevertheless it somehow does not immediately *seem so* to them. This may be overcome with significant effort, but the point is precisely that it will *require effort*, whereas someone who is well will not have to overcome the apparent lack of significance in the first place. Depressed people are not *deluded* as to the basic utility of everyday objects, but something nevertheless seems to be 'off' about them, in a manner clearly associated with their practical utility.[24]

The critical takeaway is this; diminished practical significance that is indicative of agential pathology is not closely connected to considered judgments about objects' utility, nor a process of deliberative practical reasoning. While I will not rule out at this stage that consciously considered mental states such as beliefs may *lead* to diminished experiences of practical significance, the experience itself is not so much like *believing* or *thoughtfully considering* objects to be unavailable, uninviting, or otherwise useless, but of *immediately experiencing* them to be that way.

Call the necessity of a theory explaining such experiences the *practical significance constraint*. This is, admittedly, the vaguest of the three constraints, but will suffice for the analytical purposes it is put to in this paper. Importantly, I will flag occasions when how literally one takes descriptions of objects being 'inviting' will matter for whether or not particular theories are able to satisfy this constraint.

4.1.2. Summary: The Burden of Explanation

The above are core *phenomenological* constraints on our theorising about agential pathology. It is important when building theories of psychological, and especially

[24] It might be useful to think of this kind of experience as being 'perception-like' in the sense that it involves how the world appears, or is presented, to an agent, rather than necessarily how they explicitly judge it to be. I avoid this terminology in the main text because it might be seen as illegitimately stacking the deck in favour of Perceptual Theories of agential pathology (to be introduced in section 4.4.). Perhaps it does this despite my best efforts; but if so, this must be put down to the fact that, in certain ways, the *actual character* of agential pathology stacks the deck in this way. Ignoring this feature has certainly impoverished a number of accounts of agential pathology, as we shall see.

psychiatric phenomena that we account not just for some coarsely-specified functional outcome by invoking plausible psychological mechanisms and constructs, but for what the phenomenon is really like. In part, this is simply because the coarse-grained functional outcome alone (of failing or struggling to act), which is widely used to frame the debate in the literature, is likely to be susceptible to a *huge variety* of possible explanations. Respecting the phenomenology allows us to narrow the field of possibilities in a principled manner in advance of further empirical study; it permits us to make the outcome that our explanations should target much more precise.

This is why I treat the sense of impossibility, lethargy, and loss of practical significance as indispensable explanatory burdens for a theory of agential pathology, rather than items of explanatory interest that merely *accompany* agential pathology. To put it perhaps a little glibly, the phenomenon of agential pathology is not just any old struggle or failure to act, but one which is systematically inflected by the phenomenological characteristics detailed above. Thus, our explanations must not target any old failure to act, but rather this phenomenologically particular *kind* of failure to act.

4.2. Mental State Theories

The first class of theories I will examine attribute onset and maintenance of agential pathology to the absence, interference, or degeneracy (in a sense to be clarified) of folk mental states involved in the planning and/or execution of action. Many of these theories owe a lot to philosophical theories of motivation, and some emerged specifically in the context of debates about the nature of (moral) motivation.

4.2.1. Desire Theories

For instance, both Michael Stocker (1979) and Michael Smith (1994) make use of the character of 'the depressive' to offer an argument in favour of Humeanism about motivation. This is the view that there is a sharp ontological divide between beliefs (which lack intrinsic motivational force) and desires, willings and similar pro-attitudes (which are the ultimate source of all motivational force). According to Humeans, "cognitive states are wholly lacking in motivational power" (Law, 2009).

Anti-Humeans judge that at least some beliefs (perhaps those regarding what is good or right) can motivate an agent in the absence of a suitable desire-like state.

Both Stocker (1979: 744) and Smith claim that the character of the depressive is one whose "evaluative outlook [is] intact" (Smith, 1994: 120) but who is nevertheless altogether lacking in motivation. The salient point here for Smith is that if this description is correct, then the depressive has exactly the same set of beliefs as an imagined virtuous agent, and yet fails to be motivated to act in the same way as they are (or, indeed, at all). This, he claims, is decisive evidence against the Anti-Humeans, since they cannot explain how such a situation (the apparent *total* separation of motivation from belief) comes to be. The salient point for us is that Smith has implicitly offered a theory of agential pathology; in such cases as the one he describes, the agent is suffering from an absence of a number of the desires, or other pro-attitudes, required to motivate her to act. Her evaluative outlook (that is, her belief set) may remain unchanged, but her desires drain away. When she fails to act, it is because she lacks the desire to do so[25].

An absence of desires *could* potentially satisfy the practical significance constraint if we were to assume, plausibly, that the appropriate modulation of such practical significance is one of the characteristic functional roles of desires (and/or other desire-like mental states). This assumption is plausible, because it seems undoubtedly true that as my desires change, so does my experience of the practical significance of objects. If I am thirsty, the glass of water on my desk seems particularly important, perhaps even inviting. If I instead desire a break from essay-writing, the water is experienced entirely passively, and recedes into the background, while my smart-phone acquires greater salience. Since desire seems to at least partially determine practical significance, a widespread absence of desire

[25] Iain Law (2009) points out that one *could* interpret Michael Smith as not talking about *actual* people with depression, but rather some possible kind of person that he just so happens to evocatively name 'depressives'.If so then it would be unfair of me to treat this as a serious theory of agential pathology. Nevertheless, he also points out that it would be uncharitable to do so, since unless Smith has reason to believe the people he describes actually exist, he is merely asserting the possibility of exactly the situation that the Anti-Humean denies is possible. This would barely amount to an argument, let alone a dialectically decisive one, against the Anti-Humean position.

could be imagined to produce a widespread diminution of practical significance. There are two major difficulties for this proposal, however.

Firstly, the unmediated character of the loss of practical significance characteristic of Agential Pathology speaks against the idea that this is all to be explained by a lack of the relevant desires. One of the peculiar aspects of Agential Pathology is that objects and features lose their practical significance, even when one wishes to make use of them to satisfy particular goals. These (common) cases cannot be explained away by the suggestion that we lack the relevant desires. Their peculiarity lies precisely in the fact that an agent's desire-profile is dissociated from their experience of practical significance.

Secondly, Smith and Stocker's idea entirely fails to satisfy our first two constraints. Take the impossibility constraint first. If I do not desire to do something that I once did, that fact is entirely distinct from my judging that I cannot do so, or that it would be particularly difficult. Of course, such a judgment of impossibility might lead to my losing the desire to pursue that course of action, but that is beside the point. Even a widespread loss of previously held desires would not be expected to produce the relevant sense of impossibility. A re-evaluation of my life's work and making a decision to ditch Philosophy because being in the academy is bad for my health might well provoke a loss of all sorts of desires, and quite possibly offer me no replacements. Such an upheaval *might* even cause me to become (more) depressed. But the loss of my 'philosophy-related' desires by itself is insufficient for those previously-desired actions to seem impossible for me to do. So, Smith & Stocker's theory cannot satisfy the impossibility constraint; it does not explain why agential pathology involves an aspect of impossibility.

A similar fate befalls a related theory, on which, as Matthew Ratcliffe succinctly puts it,

> depressed people fail to act because of an inability to satisfy their desires, due to loss of positive affect. The depressed person contemplates doing *q*, imagines that doing so will bring no relief from negative feeling, and therefore does not do *q*. (Roberts, 2001, cited in Ratcliffe, 2015: 156)

Once again, whatever its other merits, Roberts' theory fails to satisfy the impossibility constraint, because lack of anticipation of future satisfaction from performing some action, is quite distinct from feeling in some way that the action in question is impossible. Moreover, there is no obvious story to tell which could take you from the former to the latter. Why should a person who judges that doing *q* will provide them with no relief from negative feeling thereby come to feel, or believe, that *q* is impossible or unusually difficult? This is certainly not *generally* true. Rapping my knuckles against my desk is about as unsatisfying an action as I can imagine, but I do not feel it to be impossible or unusually difficult. This is still the case if I form the desire to rap my knuckles against the desk, perhaps in order to prove a point. I anticipate significantly less (indeed practically no) satisfaction from doing that than from satisfying any number of my other desires, for instance making myself a coffee. But this does not produce any sense that this action is somehow harder than all the rest, or even unusually difficult at all. I can imagine easily enough how the relationship might work in the other direction; actions that are judged to be unusually difficult might be anticipated to be unsatisfying to perform (due to the large amount of predicted effort that would be required to do so). Nevertheless, there does not seem to be any reason to suppose that the inverse is also true; imagining the performance of an action to be unsatisfying does not very often (if ever) seem to lead to the feeling that it is or would be unusually difficult to perform. Thus we would need a special reason to think that this is true of the kind of failure of anticipation that occurs in depression; one which is not currently forthcoming.

There is perhaps one way a defender of a desire-based theory of Agential Pathology could argue that they in fact *are* able to satisfy the impossibility constraint. This would be to suggest, as Smith does, that desires are nothing over and above "dispositions to act in certain ways under certain conditions"[26], where such conditions can only obtain if the subject has suitable beliefs (1987: 52-53). For a desire to be absent on such a view is simply for a bare disposition to act to be absent. A

[26] Smith can comfortably accommodate into his main theory the refinement that hopes and wishes *also* satisfy such a description by saying that the structure of motivating reasons is that of a belief-proattitude complex (rather than strictly belief-*desire* complexes).

defender of such a theory can then suggest that if an agent is aware of having a pervasive lack of action-dispositions, as is plausibly characteristic of cases of agential pathology, then the actions themselves will feel impossible to perform. In order to feel able to do something, I must, the defender of this position will claim, feel disposed to do it. If I do not feel so disposed, the action will feel impossible, or at least very difficult to perform.

I do not think this suggestion is very plausible. It seems reasonable to grant that in order for it to be possible for me to *actually* perform an action in the moment, I must *at some point* be disposed to perform it. It is hard to imagine a case of an agent acting without first being disposed to do so. A bowl, after all, cannot break without being first disposed to breaking. But it seems implausible to suppose that my *feeling* of what I could possibly do is determined by what I currently feel disposed to do, or even by what I am currently disposed to do (regardless of how I feel). I am currently aware that I *could* steal the laptop left unattended next to me in the library. But I do not feel the slightest disposition to do so. Nor does it really make sense to claim that I am so disposed without knowing it. If I were, I would have grabbed it and be half way out of the door by now.

One could, of course, claim that my dispositional properties come very cheaply. Perhaps *I am* disposed to stealing the laptop under certain extreme conditions, which are not at all close to currently being satisfied, and this explains my feeling (under current conditions) that I *could* do so. But this move points to a broader and perhaps even more serious problem for the theory as a whole. If we allow that I count as being disposed to steal the laptop simply because there is some set of extreme, yet imaginable, circumstances in which I would do so, then the advocate of this theory must accept that I also count as currently having a *desire* to do so, since desires (according to them) *just are dispositions*. Such a claim is clearly untenable on any day-to-day understanding of 'desire', since I would count as desiring at any moment almost anything that I could in principle do.

Perhaps a defender of this position could retreat to the claim that desires necessarily involve dispositions, but that such dispositions are insufficient. This would allow them to maintain that current dispositions (or current felt dispositions) determine the

feeling of what actions are possible, and that losing desires involves losing such dispositions, without concluding that any old disposition counts as a desire. But this does not resolve the main issue at hand. If the necessary dispositions still come cheaply (that is, any set of possibly enabling circumstances will suffice for having a dispositional property of the kind necessary for action), then any case of agential pathology will nevertheless count as one in which the person is disposed to act. This is because they presumably *would act* if the circumstances included their not being depressed. On this view, the lack of a disposition to act *cannot explain* why a person feels certain actions are impossible, simply because *they are, in fact, so disposed*. If, on the other hand, the necessary dispositions are not that easy to have, then we are back to square one. It is highly intuitive that I can have and/or feel no disposition to do something that I nevertheless feel it would be possible for me to do. So, the loss of a disposition-involving desire cannot (at least by itself) explain my sense that a particular action is impossible for me to perform.

The lesson here is that although there is a clear sense in which my doing anything depends on me being at some point disposed to do so, my sense of what is possible and impossible for me to do at any given moment is not plausibly determined by my dispositions at that present moment.

Finally, and briefly, no desire-based theory looks able to satisfy the lethargy constraint. For one, a lack of a desire to do something, even many things, is grossly insufficient to explain why I feel lethargic. I lack the desire to do lots of things, but don't consequently feel tired or weighed down. For another, an inability to satisfy my desires, even a persistent one, is similarly unrelated to lethargy. Such a situation might be expected to prompt frustration in me, but it has no apparent systematic connection to lethargy. So, either the desire-theorist must explain these two apparently disparate states, or a satisfying theory of agential pathology will have to be sought elsewhere. For reasons of space, and because these theories fail on other grounds regardless, I will not attempt to do the work of the desire-theorist for them here.

4.2.2. Belief Theories

Iain Law further points out that depression/agential pathology does not typically leave a person's evaluative outlook untouched, contrary to Smith's claim (2009). He points out that a widely accepted view of depression, Aaron Beck's cognitive approach, suggest that depression is rooted in the 'cognitive triad'; negative beliefs about oneself, the surrounding world, and the future, which produce "the other signs and symptoms of the depressive syndrome" (Beck et al 1979: 11). That is, depression is the sort of thing that pretty much *necessarily* entails a radical change in a person's evaluative outlook. So the characterisation of depression offered by Smith is faulty in any case; predictable patterns of belief change are not just typical, but *characteristic* (perhaps even *constitutive*) of depression, on Beck's view.

This point, by itself, is insufficient for our purposes (though of course it somewhat disarms Smith and Stocker's intended argument). Even if changes in belief are characteristic of depression, they may not explain agential pathology specifically. Law, however, has a persuasive argument to the effect that they can and, at least sometimes, do.

Take some negative belief typical of depression on Beck's understanding, for instance, 'I am a boring person, and people are uninterested in spending time with me'[27]. Now imagine a person, *A*, who holds such a belief about themselves, is at a party and notices that a person across the room from them, *B*, looks lonely, shy, and anxious. It is perfectly conceivable that *A* has the desire to help *B*, and yet on the basis of their negative self-belief, fails to do so. After all, *A* is likely to believe that no matter how shy and lonely *B* is, *B* will nevertheless not want to talk to them. In this case, instead of *A* lacking some motivating desire, one of their beliefs acts as a *defeater* to whatever other motivations they may have (Law 2009). For Law's purposes, it is crucial that this case be describable without recourse to desire-talk at all (the goal, after all, being to account for lack of motivation in depression without conceding a necessary distinction between belief and pro-attitude). For our purposes, however, that is unnecessary. What is crucial here is simply that an interfering belief, rather

[27] Many different beliefs will do the work for the purposes of this example, but I will stick with this one throughout for clarity.

than an absent or weakened desire, may also result in a depressed person failing to act.

It is important to note that, unlike the theories of Smith, Stocker, or Roberts, this sort of suggestion seems to be able to neatly satisfy the impossibility constraint. It is highly plausible to think that negative beliefs about oneself or the world characteristic of depression could include content like 'I am unable to do *x*', where *x* is some activity that the depressed person previously undertook with ease. In the case above, for instance, the belief that one is boring could lead to the conclusion that one is unable to help the shy person – only a much more interesting person could be of any use in this situation. It is also sensible to think that these negative beliefs could be sufficiently varied or broad in scope as to interfere with action across a very wide range of circumstances. If one, for instance, sincerely judges oneself to be useless at anything one tries to do, one is likely to fail to act in a wide range of circumstances, because one believes that a very wide range of activities are simply impossible for you to perform. The idea here is that a persistent judgment that one is unable to do certain things, may be expected to quickly lead to a more all-encompassing *feeling* that those things are unusually difficult or even impossible to do.

On the other constraints, however, this sort of account comes up short. Much as with desires, there seems to be nothing of the required systematic connection between a surplus of negative, action-inhibiting beliefs and persistent heaviness or tiredness that would be required to satisfy the lethargy constraint. As before, perhaps such a story can be told convincingly. But there is no obvious connection in our everyday experience with beliefs, or belief-change.

In some respects, the prospects of the belief theory improve when it comes to the practical significance constraint. One might simply think that amongst the action-inhibiting beliefs held by the depressed person are included beliefs with content like '*x* is unavailable for my use', '*x* is irrelevant to me'. This might, indeed, sometimes be the case, but two considerations tell against it being the whole story.

Firstly, because we have previously noted that diminution of practical significance in agential pathology persists even in the face of conflicting knowledge, it seems like a

belief theorist will be forced to suggest that many of these beliefs will be held while a contradictory belief is also held; agents in the grip of agential pathology will often (indeed, typically) be hypothesised to both believe and disbelieve that some object is available to them for use. This already seems like an extravagance; while depressed people might be thought to entertain contradictory thoughts on occasion, this hardly seems to be the rule. Negative thinking does not imply, or even suggest, contradictory thinking, and the latter is not generally thought to be an especially common feature of depressive illness[28].

Secondly, this picture entirely fails to explain why the diminution of practical significance in depression seems to be unmediated; the experience we are looking to explain is not characterised by deep and abiding cognitive conflict, or deliberative indecision over the state of the world. It is simply one in which how the world is presented to people in a direct, experiential sense sometimes conflicts with what they know to be the case. And this is not how having conflicting beliefs generally feels.

4.2.3. Degenerated Intention Theory

A final kind of mental state theory is the *degenerated intention* view. This position derives from a hypothesis put forward by Julius Kuhl (1984). Let an intention[29] be *degenerated* if it fails to represent a) when, b) where, or c) by way of what possible concrete actions the relevant task will be carried out (Kuhl 1984). Intentions of types a and b are likely, if not relinquished, to claim 'space' in working memory at inappropriate times or locations, and thus interfere with the moment-to-moment fulfilment of other, better specified intentions. Intentions of type c can, naturally, not result in action and may also, if not relinquished, interfere with other active intentions (as well as the development of new ones) simply by increasing the load on

[28] On the basis of my own experience, I will admit that this does happens sometimes, especially (in my case) when I am on my way out of a depressive episode and am explicitly challenging negative beliefs when they arise in consciousness. But diminutions in practical significance do not track these situations. Objects can appear unavailable to me in the relevant sense both when I am utterly convinced that they are, and when I begin to explicitly doubt my own capacities to a degree that makes it seem plausible that they in fact are not.

[29] We should think of intentions here as the sorts of thing that are able to be the immediate mental antecedents (Nanay 2013) of concrete bodily actions (like intending to pick up a glass of water), rather than abstract motivations to achieve certain goals in the future (like intending to bring about the end of Patriarchy).

working memory, and causing one to be moved under certain circumstances to satisfy the degenerated intention in question here and now, possibly at the expense of other, satisfiable, intentions.

The view as it relates to agential pathology, then, is that people with depression are especially prone to 1) taking on degenerated intentions and 2) failing or struggling to relinquish them. There is some empirical evidence that this is indeed the case. Kuhl & Helle 1986 demonstrated that service users admitted to in-patient psychiatric services were significantly more prone than controls to taking on degenerated intentions suggested by investigators, as well as failing to relinquish them when the impossibility of satisfying them became apparent. Since this is likely to interfere with the formation of new intentions and the execution of already existing ones, due to the additional load on working memory and persistent interference of degenerated intentions, people with depression will struggle to act (assuming that intentions in the sense used here are necessary, or at least typically necessary pre-conditions for action).

It seems that this account satisfies the impossibility constraint. The presence of a degenerated intention plausibly leads to a feeling that what one intends to do cannot be fulfilled, or acted upon. This may be for a variety of reasons. Perhaps while one recognises that one intends to act so as to satisfy some goal, one also experiences constant interruptions in their attempts to satisfy said goal. Alternatively, one may struggle to translate an intention to satisfy some goal into a concrete action plan, and so fail to act upon the intention. Either of these situations may contribute to an experience of unusual difficulty, or even impossibility in satisfying one's intentions, in the former case because of interruptions and in the latter case because of the lack of any actionable content in the intention at all. It is an interesting question as to whether the inability to form novel intentions may translate into a feeling of impossibility – certainly it might if one is in a situation where one would normally expect to form such an intention, but instead feel overwhelmed by other things 'to do'. Yet it is clear that the mere absence of an intention to act in a particular manner does not translate into a feeling that such an action is impossible. I do not currently intend to go to the pub, but I have a strong feeling that it is possible.

It is also plausible that this account satisfies at least some of the requirements of the practical significance constraint. Contemporary accounts of motivation often highlight the role that intentions to act may play in directing attention towards objects and features in our external environment relevant to their satisfaction (see e.g. Wu 2011, or the previous chapter of this thesis), as well as in the formulation of concrete action plans and the initiation of actual motor activity. That is, many accounts suggest that an intention to drink from the mug in front of you not only functions so as to formulate and execute a sequence of motor commands that will cause you to grasp the mug and lift it to your lips, but also directs perceptual attention to the mug and the features (such as its size and distance from the agent) relevant to grasping it (Wu, 2011; Nanay, 2013). If this is right, it is plausible that such perceptual selection could be experienced as enhanced practical significance of the object and/or features in question. This is because, ordinarily, if an intention perceptually selects a certain object, it raises your awareness of it at the same moment that you have a concrete action-plan (represented by the same intention) that involves interacting with it in some way. This perceptual selection is also likely to occur in tandem with the initiation and online control of that action, also accomplished by way of the same intention (or at least some process it sets in motion). Thus the relevant instances of perceptual salience are closely tied to a particular, concrete action that the agent is *psychologically ready* to undertake. If this is right, the degenerated state of such an intention could explain diminished experiences of practical significance, either by precluding perceptual selection entirely, or by failing to associate it sufficiently closely with action-readiness.

Note that this proposal does *not* explicitly involve the suggestion that degenerated intentions change the contents of an individual's perception (this will be the sort of suggestion made later on in section 4.4.). Rather, it is a claim about the close *association* of attention (guided by an intention) and the action-readiness such intentions bring. On this view, practical significance is simply the result of this kind of associative relation, and I need not make any claims about what kinds of properties are, generally speaking, perceptually represented.

It is doubtful, however, that accounts that focus on degenerated intentions can satisfy the lethargy constraint. Even if the strain on working memory such an account implies might suggest a kind of mental exhaustion associated with agential pathology, it does not obviously explain the feeling of physical, highly embodied exhaustion and heaviness that characterises it. Consider, degenerated intentions are those that lack particular contents such that they are hard, if not impossible to satisfy. This experience might naturally be understood as frustrating. But I can be (and fairly often am) frustrated in the satisfaction of my intentions without experiencing an embodied feeling of deep, almost painful lethargy. If I intend to pick up my glass but then observe that somebody has just knocked it over as they walk past my desk, this might plausibly involve a failure to satisfy an intention that nevertheless persists for a time as I take in the fact that the planned action is no longer possible. But it doesn't make me feel any more tired than I did before. Perhaps there is some kind of unusually close association between either degenerated intentions or their frustrating consequences, that does explain such a connection. But I propose that it is the job of those who would defend such a theory to propose and explain this connection, since it is not immediately obvious how it might be accomplished.

This is a problem that all the mental state theories have faced. The general issue, it seems, is that no mental state is sufficiently closely tied to embodiment to bridge the gap in any obvious way. To this extent, the way in which agential pathology characteristically involves lethargy will require an explanation that goes beyond the attribution to agents of certain mental states or their absences. It is to an account that focuses on the body itself that I shall now turn.

4.3. Somatic Theories

Some have proposed that agential pathology in depression is better accounted for by theories that place the body, and its role in shaping our sense of what is possible and important for us in our surrounding environment, at the centre of explanation.

One such view is due to Thomas Fuchs (2005). He distinguishes between two ways in which the body might be involved in experience. Primarily one's body is the medium through which one has experiences of the wider world. Here, the body is the thing that enables you to experience other things, rather than being itself an

object of experience. In this situation, it is not correct to say that the body is *unexperienced* or has receded from experience altogether. Rather, it is typically something of which we have background awareness; while we are primarily aware of the things that our bodies are interacting with (touching, seeing, hearing, etc), we are aware of our body *as the thing enabling the interaction*. This particular kind of feeling of one's body, or its constituent parts, as the enabler of experience of the external world is termed a *noetic* feeling (Colombetti & Ratcliffe 2012).

The body, or its activities and parts may, however, also be experienced as an *object* of awareness. Here, the body is primarily felt to be the thing being experienced, rather than the thing enabling experiences of external features of the environment. Here, the body is at the foreground of experience, and is typically experienced much as any other physical object can be. Call these kinds of bodily feelings *noematic* (Colombetti & Ratcliffe 2012).

The contrast between noetic and noematic feelings can best be characterised by example. Suppose you are running your hand over a wooden table. Your experience includes the feel of the wood under your hand, but also a background awareness of your hand itself, especially where it is in contact with the wood. Though your hand has not vanished from experience (you are not unaware of it), you are primarily aware of it simply as the thing enabling you to feel the wood grain. Or consider a case where you are waiting for a person to arrive for an important appointment with you, and they are running late. The objects of your awareness are unlikely to include your body. You will instead be focused on the world; the door through which they may walk at any moment, the rapidly cooling cup of coffee you bought for them, your watch, and so forth. But all of these aspects of the experience are shot through with a sense of tense anticipation and/or anxious energy, which is in part the product of the tightness and restlessness of your body entering into the experience in the background. The tense anticipation of the whole situations is made manifest *through* a tense bodily feeling. These are both noetic bodily feelings.

Contrast these with noematic feelings; you remove your hand from the wood and focus your attention on it while you do nothing else with it. Here, your hand is the object of your awareness, it is in the foreground of your experience, and it is not felt

to be the *mediator* of your experience of something else. Or consider the experience of feeling your stomach rumble, or feeling a sharp pain in your head that lacks any apparent external cause. In both of these cases the body is experienced as an object of experience, not as something that enables you to experience external objects. Or consider a situation where you are performing a 'full-body scan', perhaps as a form of mindfulness therapy or meditation. As you shift your attention from your feet, moving it up your legs, to your stomach, chest, and so on, you experience each of these body parts noematically; that is, as distinct objects of awareness rather than as mediums of interaction with the wider environment.

There is a key point to introduce here. It is generally thought that both the feeling of the body as a mediator of experience and the ability of an agent to use it effectively as such (i.e. to effortlessly engage in habitual, day-to-day activity) are diminished by its being a primary object of awareness. That is, the more the body is felt to be at the foreground of experience, the less it is experienced as a medium of engagement with the external world, and the more effort is required to use it for that purpose. Consider, for instance, a case where a typist reflects for too long on the precise movements and sensations of their fingers. The more the fingers (rather than, say, the keyboard or computer screen) become the objects of awareness, the less they are experienced as a tool for typing, and the less effective they become as tools to actually type with.

According to Fuchs, depression, (or 'melancholia' as he sometimes calls it) is characterised by a diminution of noetic bodily feelings (which he refers to as feelings in which the body is 'transparent'), and a concurrent enhancement of certain kinds of noematic bodily feeling[30]. In particular, the body ceases to be experienced as a medium of contact with the world, and is increasingly involuntarily experienced as a *heavy and unwieldy* object of foreground awareness; one that impedes the agent's 'access' to the wider world. He labels this process *corporealization* (in the non-standard sense of increasingly resembling a corpse). Since the body is no longer

[30] It is worth noting that noematic bodily feelings are not necessarily disabling in this way, or to any significant degree. Rather, the *involuntary* and *pervasive* way in which the body becomes an object of experience is the key to understanding the role this change plays in bringing about agential pathology.

experienced as the kind of thing that enables action, but rather as an impediment to it, the agent struggles to act with the fluidity that they once did. Fuchs writes,

> To act, patients have to overcome their psychomotor inhibition and push themselves to even minor tasks, compensating by an act of conscious effort for what the body no longer does relatively effortlessly. (2005: 99)

Here we see how this theory places the body at the centre of the explanation of an agent's struggle to act. The difficulty emerges precisely because 'acts of will' (paradigmatically mental efforts) must compensate for the fact that the body no longer functions as an effective interface between the agent and the world, and must instead be forced into action despite being increasingly experienced as a barrier between the agent's intentions and the wider environment, rather than an enabler of them.

Fuchs leaves unclear how we should understand the *mechanics* of this retreat of noetic feeling and concurrent encroachment of noematic bodily awareness. He does, however, tell us that the normal neurobiological underpinnings of these sorts of bodily feelings is the integration of various kinds of proprioceptive and otherwise kinaesthetic information (Fuchs 2005: 96). Thus we may sensibly infer that he thinks disturbances to these feelings involve disturbances to the same systems. The details, however, are not important for our purposes.

Fuchs' core proposal is that the process of corporealization inhibits an agent's ability to *initiate* and *sustain* action. Smooth everyday action is typically (in part) enabled by a feeling of the body as an active mediator of world-directed action. In cases of agential pathology, the body is increasingly experienced instead as a passive object of foreground awareness; something which must be effortfully forced to move, rather than something that enables engagement with the outside world.

This suggestion neatly satisfies both the impossibility and lethargy constraints. Action feels impossible or difficult because under normal circumstances, noetic experience of the body functions to enable relatively effortless action; as the body is experienced more and more as an object that stands 'in the way' between an agent and their environment, this enabling function is not fulfilled, and action requires

increased effort. Further, this difficulty is marked by lethargy, because the character of the encroaching noematic feeling is of the body as leaden and unresponsive.

Yet this account is still insufficient to properly characterise the experience of agential pathology. As I pointed out above when I insisted on the practical significance constraint, it is not simply the ability to *initiate* or sustain action which is inhibited in cases of agential pathology, but the practical significance of objects and features of the external environment. Critically, Fuchs' core proposal cannot yet explain why agential pathology takes on this distinctive character, simply because it concerns itself only with explaining the characteristic failure to initiate and sustain action, rather than wider features of experience that relate to and support actions that one wants to, intends to, or merely could perform.

A variant of Fuchs' core proposal, however, developed in detail by Matthew Ratcliffe (2015) and Benedict Smith (2013), suggests that the kinds of changes to bodily feelings invoked by Fuchs interfere with more than just an agent's ability to initiate action. They also, so Ratcliffe and Smith claim, interfere with an agent's sense of what they could *possibly* do, and wider kinds of practical significance of environmental features and objects.

Their thought proceeds as follows. In addition to enabling the body to be a medium through which the world can be interacted with, Smith argues that particular kinds of bodily feelings also render the body "that through which [worldly] things are experienced as meaningful in various ways" (2013: 626). The body is not just a medium of interaction between an agent and the world around them, but "is directly involved in shaping the meaningful contours of the world in ways that are usually inextricable from experience" (2013: 627). In particular, bodily feelings are the primary medium through which we experience the opportunities for action that the world presents us with. Crucially, the claim is not simply that *what we can do* is in part determined by our bodily make-up (which, I take it, is uncontroversially true), but rather that our implicit *sense* of what we can do, and what we are attracted to do, in a given situation are partly determined by the quality of our experience of our own bodies.

This claim seems at least *prima facie* plausible. It seems at least possible that ordinarily, if I experience the keyboard in front of me as something that I can type on, that sense is at least in part determined by a background awareness of my hands as things that can be used to type. That is, my experience of this action-possibility is partly determined by experiencing my body, or some relevant part of it, as the sort of thing that is able to perform that sort of action (typing). This is another kind of noetic bodily feeling; an awareness of the body as that through which *specific* things can be done. Cases of awareness of possibilities for action where the body does *not* play this subtle role might also be thought to be unusual in the following sense; they are cases where I actively consider the various possibilities offered to me by the environment, and make an effort to 'think outside the box'.

For instance, it is plausible that I have a noetic bodily feeling that partly determines my sense that the keyboard is something that I can type with, but less plausible that I have such a feeling partly determining my sense that I could pick up the keyboard and use it to batter the person sitting next to me around the head. This is at least in part because the first kind of experienced possibility is something that I experience as part of the flow of ordinary day-to-day action, and the second is an abstract judgment of something that I could do. That is, the first kind of experienced possibility is of practical significance to me, and the second is not. It is the practically significant action-possibilities in my environment that are partly revealed to me through distinctively bodily (rather than psychological) experiences, according to Smith (2013) and Ratcliffe (2015). So, if depressive experience involves a process of corporealization (specifically, a retreat of noetic bodily feelings), then it stands to reason (or so Smith and Ratcliffe posit) that the practical significance of various environmental action-possibilities will retreat along with it. In such a case we would expect increased psychological effort to be required to act (which Fuchs observes), and for depressed persons' sense of action-possibilities to take on a more explicit and

deliberative character (that is, more like the 'battering-my-neighbour' possibility than the 'typing' possibility)[31].

The central claim of these accounts is that corporealization not merely creates difficulties initiating actual actions, but also disturbs our sense of what we can do, and what is otherwise significant to us in our environments. That is because, so Ratcliffe and Smith claim, bodily feelings are the primary medium through which we experience environmental significance of various kinds.

This variant of the somatic account appears, *prima facie*, able to satisfy the practical significance constraint in principle. The hypothesis that bodily feeling determines the practical significance of objects in and features of the wider environment is exactly what this account adds to Fuchs' basic somatic story. A wide range of ways in which features of the environment can appear significant depend critically, according to this view, on the presence of the right kind of bodily feelings and orientation. Feelings that relate the agent not to the world through their body (noetic), but to the body itself (noematic) are also likely to be of the sort that will impede experiences of practical significance. If the objects and features of the world are experienced by us as practically significant (and, in particular, actionable) in virtue of the background experience of a transparent, mediating body, then the involuntary encroachment of opaque, objectified bodily experiences will interfere with those same experiences.

Two questions remain completely unanswered, however, which together raise serious questions about the somatic theory's explanatory capacity. These are as follows. Firstly, why should someone believe that the character of our background bodily feelings does, in fact, determine all (or even a majority) of our experiences of practical significance (i.e. do Smith or Ratcliffe have an argument available for this particular claim beyond repeating the introspective claims of historical phenomenologists)? Secondly, how are Ratcliffe and Smith to *explain* this determination, even if it is assumed to hold?

---

[31] I cannot comment on how likely this is to be true in general, though it would certainly seem to be empirically verifiable/falsifiable. From personal experience, I can go so far as to say that it certainly rings true for *some* cases where I struggled to perform simple actions when depressed.

On the first point, I can appreciate the *prima facie* attraction of thinking that bodily feelings shape my sense of objects' practical significance in some cases. For instance, when I am holding a pen in my hand, whether my focus is on my fingers or the pen (as-gripped-by-my-fingers) clearly affects how prepared I feel to write with it. That much is introspectively reasonable (at least to me). But it is not remotely obvious that this generalises, especially to objects that I am not currently in bodily contact with. The sense in which my shampoo seems significant to me when I am showering, and insignificant to me when I am shaving in the same room, is not obviously connected to any change in bodily feeling. Changes in feeling somewhat like this are a huge proportion of those we are seeking to account for with a theory of Agential Pathology; we want to explain what has changed when I become depressed, am showering, and now experience my shampoo bottle as practically insignificant and 'inert'.

Perhaps unlike advocates of mental state theories, Ratcliffe and Smith will be hard-pressed to suggest that the change is unconscious. While I may easily conceive of myself to be in mental states of which I am not aware, it seems almost immediately contradictory to suggest that I am having a bodily feeling of which I am unaware. Of course, bodily feelings can be more-or-less foregrounded in experience, but I am inclined to think that if I don't *feel* anything, then I cannot count as having a *bodily feeling*. Naturally, those more sympathetic to Ratcliffe or Smith than me need not try to make such a claim. But in that case, I think the burden is on them to precisely identify the kind of bodily feeling that could account for such a case. As it stands, characterising these changes of practical significance as all or even mostly determined by changes in bodily feeling strikes me as simply phenomenologically inaccurate. Whatever the difference is, it does not feel to me to be somatic.

Secondly, even if I entertain, for the sake of argument, the hypothesis that many or most of these kinds of practical significance are determined by bodily feelings, it is not obvious how I, or anyone else, are to go about explaining that determination. It is tempting to say that these various kinds of practical significance (say, the *availability* or *utility* of objects for instance) are *represented* in some aspect of my experience. But while it is not wholly implausible to think of some kinds of bodily feelings as

representing things about the world (perhaps pains or itches do this (see Klein 2007)), it is highly unnatural to think of a wider class of bodily sensations as representing *anything*, let alone properties of objects in, and features of, the wider environment. For one thing, most bodily feelings seem to lack the requisite structure to represent complex environmental properties. This is, broadly speaking, the slight advantage that mental state theories have in dealing with the practical significance constraint – they are naturally thought of as structured, representational entities. Absent this option, it is hard to see how this determination relation could be unpacked in a psychologically plausible way.

Note that my claim here is not that somatic accounts are fundamentally unable to satisfy the practical significance constraint. My claim is more modest; no explanation is *currently* forthcoming as to how any *existing* somatic account could satisfy it in its full complexity.

So what could account for the kinds of practical significance this account fails to account for? One answer is a mental state theory of some kind. But I think there is another option. One might say that agential pathology is the result of pathological *perceptual* states. These have the advantage of being naturally thought of as representational states, without possessing some of the disadvantages that mental state theories did on the topic of practical significance. Namely, they would also account for the *unmediated* character of the disturbances to practical significance in depression. But are they plausible candidates for the central components of a theory of agential pathology in general? It is to this question that I now turn.

4.4. Perceptual Theories

Perceptual changes characteristic of depression are not often the focus of research. More often, psychologists take negative beliefs, low mood, anhedonia, and (occasionally) somatic symptoms to be their primary explanatory target (for a few notable exceptions, see Bubl et al 2010; 2009). This lack of focus should be somewhat surprising after one has read a few first-personal accounts of depression. Many people with experience of severe depression speak of abnormalities in the way they perceive the world when depressed. For instance,

> The world now looks remote, strange… uncanny. Its color is gone, its breath is cold, there is no speculation in the eyes it glares with. (James 1902: 151)
>
> It feels as if I am a ghost – I cannot touch or see the world clearly and it all becomes grey and transparent. (Anon, cited in Ratcliffe 2015: 33)
>
> …wherever I sat—on the deck of a ship or at a street café in Paris or Bangkok—*I would be sitting under the same glass bell jar*, stewing in my own sour air. (Plath 1963: 178, *emphasis mine*)

These reports give *prima facie* evidence that severe depression is associated with certain perceptual abnormalities. If we take these quotes at face value, we are inclined to think that to the depressed person the world *looks* different; somehow distant, cut off, or even not really there. I think the reason why these phenomena do not receive more attention is that they are deemed a comparatively uninteresting consequence of other symptoms; they are thought to be the direct consequence of, say, a low mood, and it is assumed that they will dissipate when that underlying symptom is relieved (for a recent exception see Golomb et al 2009). More importantly, I think they are generally assumed to hold no independent explanatory power of their own. That is, they are not thought to figure in an explanation of any other feature of depression[32]. As far as depressive symptomatology goes, perceptual abnormalities are deemed to be *epiphenomena*.

In this section, I will argue that such a view is mistaken. The perceptual abnormalities in depression can, when suitably unpacked, afford us an explanation of (some features of) Agential Pathology. In particular (though not exclusively), they afford us an explanation of the characteristic diminishment of various kinds of practical significance, which the other two classes of theory have failed to deliver.

Moving beyond the suggestive character of the quotes above, I propose to investigate one central kind of perceptual abnormality characteristic of depression, which I term *active distance*. I argue that it is likely critical to understanding the character of agential pathology.

---

[32] See Fitzgerald (2013) for a recent exception. In that paper, it is argued that perceptual abnormalities might mediate anhedonia in Major Depressive Disorder.

4.4.1 Active Distance & Ecological Percepts

I begin by unpacking the notion of active distance. This refers to the experience reported by people with depression whereby the world, its features, and occupants look to be somehow distant, detached, and separated from the observer (Benson et al 2013; Kendler 2016). As Benson and colleagues note, depressed people often report an overwhelming sense "of physical distance having been introduced between the self and the world" (2013: 65). Since we typically judge distance perceptually, this is naturally thought of as being an experience that is, minimally, partly perceptual in character.

A quick sidenote. One might wonder how this experience is to be distinguished from derealisation, a diagnostic feature of many *other* psychiatric conditions, but not Major Depressive Disorder (MDD). My personal opinion is actually that there is *no* clear difference between this common depressive experience and what is termed 'derealisation' elsewhere in psychiatry. But if this is right, then why is derealisation not included in the diagnostic criteria for MDD? The state I am describing is, after all, a common experience of people with depression. The answer I think is predominantly sociological; Kenneth Kendler accurately identifies a trend in DSM criteria since DSM-III to diagnostically exclude many features that were previously widely agreed to be central to depression (e.g. Muncie 1939). The reasons for this are many and varied, but the central upshot is clear; the DSM-V may well not accurately capture many central features of depression, including the derealisation described above.

This may be true, but it is, I admit, a philosophically unsatisfying answer. Here is a reply that captures a more interesting dimension of the experience; though it *is* a form of derealisation, it is one that specifically pertains to action, in the sense that the world appears *unreal* specifically with respect to the agent's ability to act on it. This is not true of derealisation more generally, which is typically more all-encompassing. So the form of derealisation associated with agential pathology is of a specifically action-oriented kind. This distinguishes it from derealisation as a broader psychiatric phenomenon.

Returning to the main point of this section, there are at least two ways in which one might initially interpret this kind of experience of worldly 'distance'. Firstly, one might think that it means the same thing as when I note that a skyscraper two miles away looks to be 'distant'. That is, it looks to be behind closer objects; if I fixate on a different object in the middle-distance and move my head in a particular direction, motion parallax ensures that it seems to move in the same direction; and it looks smaller than I know it to actually be. Call this *literal distance*.

This, I think, is an obviously uncharitable way to interpret the term 'distant' in the above quotes. Having the world appear to be 'distant' in this sense would be a very distinctive kind of experience, which would presumably lead to numerous other phenomenological features of such reports (e.g. abnormalities of motion parallax, occultation, texture gradient, and so forth), which we do not observe.

One might be tempted to think that any other way in which we might interpret the claim that the world appears to be 'distant' is hopelessly metaphorical. It might mean that the person doesn't feel socially connected to their peers, or that they struggle to emotionally connect with features of their environment in the way that they once did. I don't deny that this is likely to be the case for some instances wherein a depressed person reports that some feature of the world seems 'distant' to them, but I nevertheless think there is a more robust sense of distance available to us here.

This other sense of distance I term *active distance*. The depressed person, according to this interpretation, is reporting a kind of experience whereby the world, its features, and occupants seem to be inaccessible from a "practical, active point of view" (Slaby, Paskaleva & Stephan 2013: 44). For instance, a bottle appears as actively distant if it looks to me to be out of reach. Alternatively, my shower appears to be actively distant if it feels impossible for me to make use of. Since literal proximity is generally a key feature of whether or not I can physically interact with an object, a case where I have an experience of an object or feature being unengageable where proximity *isn't* the issue may nevertheless get reported as a kind of distance, especially if the experience is otherwise alien. Active distance, understood in this way is not *just* ego-centric, in a way literal distance also is. It is, specifically, tied to an agent's awareness

of their own *capacities to interact with the world*. It is, in other words, not merely ego-centric, but action-centric.

One reason to interpret the experience of 'distance' reported in Agential Pathology in this way, is that depressed people's descriptions of feeling distant from the world are characteristically expressed in ways that suggest an *active separation from the world*. Consider Sylvia Plath's memorable image of being surrounded by the glass walls of a Bell Jar; this description is not meant simply to indicate separation from the world, but a kind of separation where one feels unable to step into the world, or act in it.

Alternatively, consider the following report of a suicidally depressed person taken from a study conducted by Outi Benson and colleagues,

> I want to reach out to the world, but it isn't there to reach out to… (2013: 65)

Here the focus is on the world as being unavailable, specifically as somewhere to *reach out to*. The world's separation from the subject is described in a specifically *active* way. Sometimes, indeed, this is not described specifically in terms of distance (though this is a common way of expressing the point), but rather in terms of objects being sapped of some essence, in a way that makes them appear impossible, difficult, or uninviting to act upon. For instance,

> It was as if the whatness of each thing … the essence of each thing in the sense of the tableness of the table or the chairness of the chair … was gone. There was a mute and indifferent object in that place. *Its availability to human living … in the world was drained out of it.* Its identity as a familiar object that we live with each day was gone … *the world had lost its welcoming quality*. (Hornstein 2009: 212-13)

Here, the person describing their experience are clear to clarify that what seems to be lost from the objects is a quality of *availability* or *welcoming-ness*. These sorts of qualities connect most naturally to what we feel able or invited to *do* with such objects.

This is all well and good. But how is one to make sense of the claim that something like active distance, as described here, is represented in depressed people's *perceptual* experience (rather than simply their cognition more widely)? My answer to this will

be that there are numerous kinds of perceptual representation, arguably present in everyday perception, the *absence of which* has the potential to explain why the world seems actively distant to a depressed person[33]. I will refer generically to all of these kinds of perceptual representations as *ecological perceptual representations*[34].

Firstly, and most simply, one might think that ordinary perception involves the representation of action-properties. That is, one might think that ordinary perception (at least sometimes) involves representing apples as *edible*, ladders as *climbable*, mugs as *graspable*, and so forth. Bence Nanay has made two different arguments for this claim (2011; 2012). Specifically, he argues that a typical agent *a*, for many different kinds of action *Q*, regularly perceptually represents objects as *Q-able* for *a* (henceforth I abbreviate this to simply *Q-able*). On this story, we have one plausible way of cashing out the claim that the world appears *actively distant* to a depressed person; for a wide range of objects and values of *Q*, a depressed person fails to perceptually represent those objects as *Q-able*. Thus the world *appears* to the agent to be inaccessible to action.

Moreover, Nanay (2012) argues that to perform an action, *Q*, with an object, *x*, an agent, *a*, must represent *x* as *Q-able*, in the sense of representing that it is not impossible for *a* to *Q* with *x* (perceptually or otherwise). That is, I regularly perceptually represent apples as edible, and in order to actually eat the apple I *must* represent it (perceptually or otherwise) as edible for me, in the sense of representing it as not impossible for me to eat it.

If correct, this does not imply that it is *necessary* to perceptually represent an object as *Q-able* in order to *Q* with it, but it is does suggest that perceptual representation of *Q-ability* is a common way in which one necessary condition on *Q-ing* is satisfied in

---

[33] This contrasts with an alternative possible position that I will not consider where one thinks some property, or property-complex, can be positively represented in perception such that objects and features of the world are perceived to be actively distant. I do not consider this proposal, simply because I do not know how to begin to cash it out, though I do not rule out that there might be a way to do so.

[34] I name them in this way as a nod to the tradition of Ecological Psychology (Gibson 1979), which has historically been the area of Psychology most inclined to closely tie perception and action together in the way these accounts do. I do not intend to suggest that any of these accounts share everything, or indeed much, else in common with contemporary Ecological Psychology.

practice. If so, then it stands to reason that persistent failure to perceptually represent objects as *Q-able* could be expected to lead to a persistent failure to act. Thus this account of active distance is, in the basic sense, a viable explanation of Agential Pathology; it is able to explain why agents suffering this specific impairment would fail to act.

Secondly, one might think that ordinary perception involves not just an *informing*, but also a *guiding* form of intentionality. It is often taken for granted that perceptual representations are solely belief-like in the sense that they only purport to tell us what features the world has and what objects it contains (they exhibit only *informing* intentionality). But one might think that perceptual representations can sometimes function more like desires or intentions by 'telling' us, directly, what to do with those features and objects (they exhibit *guiding* intentionality as well). This is a claim defended by both Sebastian Watzl (2014) and Susanna Siegel (2014).

There are many different kinds of contents one might attribute to perceptual representations in order to make sense of this. One might, for example, suggest that the contents of certain perceptual representations are imperatival; of the form [Do Q!] (a suggestion that mirrors certain recent suggestions in the Philosophy of Pain/Itches, see Watzl 2014: 418, Hall 2008, and Klein 2007). Alternatively, one might follow Susanna Siegel (2014) and say that the contents of certain perceptual representations are *soliciting*. That is, they represent some object *x* as [to-be-Q'd]. For instance, a slice of chocolate cake may be perceptually represented as [to-be-eaten].

Finally, one might go even further than this, as Siegel also does. Consider an experience where you are in the supermarket and walk past a colourful display advertising a new brand of coffee. It is plausible that (perceptually or otherwise) the coffee on display is represented to you as [to-be-bought]. But this is compatible with your feeling no urge to actually buy any of it. Indeed, your awareness that the display has been consciously designed precisely so that you will represent the coffee as [to-be-bought] may put you off actually purchasing any. Alternatively, you might experience an urge to actually buy the coffee, *on top of* your experience that it is to-be-bought. Siegel argues that this urge can be explained by contents that are also represented perceptually, 'answerability contents', of the form [It is answered that: *x*

is to-be-Q'd] (Siegel 2014). This is meant to propositionally encode the idea that not only is some object represented as soliciting a particular action, but that one's perceptual experience represents that one feels moved to actually perform that action. The notion of 'answerability' is meant to resemble the feeling one gets when one hears one's name shouted across a crowded room at a party. As one hears it, one cannot help but feel as if one is in some sense 'responding to' or 'answering' the call, even if that is by way of purposefully *ignoring* the person who shouted. The idea is that, in some cases, perceptual experiences represent this kind of answerability to objects' solicitations. This final perceptual suggestion is, however, particularly controversial and not widely discussed. Since this suggests that our understanding of the potential for the perceptual representation of answerability is relatively incomplete, I shall not discuss it further in this paper. That said, if we can clarify this idea further, it is certainly a promising source of explanations of agential pathology.

If any of these views are onto something, then we have another way of cashing out the notion of active distance (which is, of course, not incompatible with the first). On this view, active distance would be a product of the diminution or absence of these guiding perceptual representations. The world would look distant and inert to the depressed person because their visual experience is lacking its typical *guiding* form of intentionality. Consequently, their perceptual experience does not inform them of what they can do/solicit them to act/create an urge to act in the way it normally would.

Further, on the assumption that some version of this phenomenon of action-guiding perceptual representations is relatively commonplace, this view can again satisfy the basic requirement of a theory of agential pathology. It can explain why persons so afflicted fail to act, or act less than others. If most of us in our day-to-day lives regularly act with the support of action-guiding perceptual representations, their absence would be expected to hinder us in our efforts.

4.4.2. What can Perceptual Theories Explain?

The above discussion identified two broad classes of perceptual theory – those concerned with the perceptual representation of action-properties (informing intentionality), and those concerned with the perceptual representation of properties

that directly guide action (guiding intentionality). What perceptual theories are able to explain depend somewhat on which of these versions one considers; that is, exactly what kind of ecological perceptual representations are hypothesised to be disturbed in agential pathology. I shall highlight and evaluate the significance of these differences as I go along.

Since the biggest motivation for introducing perceptual theories in the first instance was to satisfy the practical significance constraint, it should come as little surprise, I hope, that they do well on this score. Not only can perceptual theories account for why feelings of practical significance in general are presented to agents as unmediated (perception in general is experienced as unmediated in the relevant sense), they also have significant resources for dealing with the full range of experiences of practical significance (and their absence). Even versions of the perceptual theory that merely posit the absence of action-properties from perception are able to explain why objects strike a person suffering from agential pathology as unavailable, or unusable; the normal representations of their Q-ability are absent from the contents of that person's perception. If we also permit that perception may exhibit guiding forms of intentionality, then experiences as of objects being uninviting or inert with respect to one's goals also make sense; it is plausible that what is missing in such cases is a representation of these objects as soliciting, inviting, or demanding certain kinds of action.

Whether versions of the perceptual theory that posit guiding forms of intentionality are strictly necessary to satisfy the practical significance constraint will depend on how we interpret claims that the world seems uninviting or inert. If people experiencing agential pathology are, in these cases, giving somewhat metaphorical descriptions that intend only to latch onto a sense that the world looks as if it is unavailable for action, then guiding forms of intentionality are not obviously explanatorily necessary. If, however, we think that such descriptions are quite literal; that they highlight an experienced absence of 'invitingness', then no perceptual

theory will be able to fully satisfy the practical significance constraint unless it posits the disappearance of specifically guiding forms of intentionality[35].

Regarding the impossibility constraint, the disappearance of action-properties from perception is the most relevant explanatory possibility. If, typically, we perceptually represent objects as Q-able, then the absence of such representations will plausibly correspond to a sense that such actions just can't be performed, or at least are unusually difficult. That the possibility of acting on objects is present certainly won't be as immediately obvious or transparent as it once was.

Given, however, that an absence of a positive representation to the effect that *x* is possible does not immediately suggest an experience as of *x* being impossible, one might wonder how much of the feeling of impossibility is explicable purely in terms of the absence of action-properties. I think the answer to this depends on exactly *how* widespread such representations are, and to what degree we typically represent action-properties non-perceptually.

If, on the one hand, we generally represent *everything* (or at least all or most objects) we perceive as having not just colour and shape but also action-properties, then their absence would be highly affecting, and it seems plausible that we would not be able to consciously represent all such properties as, say, beliefs about those objects instead. In such a case, the retreat of these properties from the contents of perception would likely lead to a profound sense of action quite generally being impossible or unusually difficult, because such actions' possibility would be absent in a way that is both widespread and difficult to compensate for. If, rather, we only *selectively* perceptually represent certain objects as having action-properties (as Nanay 2012 suggests) then their absence would be both less stark, and more plausibly compensated for by conscious, belief-like representations of those action-properties instead. Therefore, whether the perceptual absence of action-properties is able to satisfy the impossibility constraint by itself depends significantly on how

[35] That is, of course, unless one thinks that perceptual experiences of invitingness can be reduced to the perceptual representation of action-properties. I do not know whether such a view is plausible, and will not pursue it here for reasons of space.

widespread one thinks such representations generally are. If one thinks they are widespread, then such a posit *can* do that work. If not, then one might wish to supplement this posit with one regarding interfering beliefs to fully explain the feeling of impossibility.

Much like mental state theories however, perception does not seem sufficiently embodied for changes to perceptual contents to be able to satisfy the lethargy constraint[36]. Thus, it seems that perceptual theories are no more able to stand alone in explaining agential pathology than the two other theory-types above. Moreover, it seems as if somatic theories are importantly indispensable in giving plausible explanations of agential pathology, since neither mental state nor perceptual theories fare well with respect to the lethargy constraint. I now turn to a more general discussion of the lessons we can learn about explaining agential pathology from my analysis of these three theory-types, and suggest directions for future empirical and theoretical research.

4.5. The case for a hybrid theory of agential pathology

What lessons should we learn from the evaluation of the preceding three types of theory? The central take away is that none of them are well suited individually to explaining the full character of agential pathology. Mental State theories in general tend to struggle with the lethargy and practical significance constraints, and only certain variations are able to satisfy the impossibility constraint. Somatic theories, on the other hand, have particular trouble with the practical significance constraint. Finally, Perceptual theories have no obvious route to satisfying the lethargy constraint, and only certain variations show any promise with respect to the impossibility constraint.

One option here would be to conclude that we need to seek out a novel kind of single-factor theory. It is, naturally, unclear what such a theory would amount to.

[36] The possible exception to this is a perceptual theory which posits the absence of otherwise commonly represented answerability contents (in Siegel's sense), since that notion connects perceptual contents with urges (which are paradigmatically embodied feelings). Since the notion of 'answerability contents' is not yet sufficiently well-developed, however, it is difficult to see how such a theory might go, never mind evaluate its explanatory capacity. Consequently, I leave this possibility aside here.

But, perhaps more importantly, this move is suspiciously unmotivated, because taken together the theories above have *all* the explanatory resources necessary to satisfy all three constraints. The obvious move, then, is to propose some variety of hybrid theory, incorporating some variant of several of the above accounts.

The constraints that we have in place currently underdetermine a specific hybrid theory that performs best. Here are a few of the most promising options that one could propose, along with the specific explanatory justifications for their inclusion, and some of their strengths and weaknesses:

> **A.** Agential pathology is explained by a combination of *interfering beliefs concerning the agent's lack of ability* (satisfying the impossibility constraint), *a sudden and involuntary retreat of noetic bodily feelings in favour of noematic ones* (satisfying the lethargy and impossibility constraint), and *a disappearance of action-properties from the contents of perception* (satisfying the practical significance constraint). This explanatory strategy has the advantage of not over-burdening individual factors with the job of satisfying more than one explanatory constraint by themselves. It also involves only the least controversial (though certainly not uncontroversial) kind of ecological perceptual representations. Nevertheless, it does require that we endorse a significant degree of theoretical complexity; at least three distinct factors are involved in agential pathology, on this story.
>
> **B.** Agential pathology is explained by a combination of *the disappearance of both action-properties and guiding intentionality from the contents of perception* (satisfying the impossibility and practical significance constraints) *and a sudden and involuntary retreat of noetic bodily feelings in favour of noematic ones* (satisfying the lethargy and impossibility constraints). This strategy has the benefit of relative simplicity, since it seems to involve only two sorts of factor. Nevertheless, it requires that we posit the existence of a highly controversial kind of ecological perceptual representation; the sort that is capable of directly guiding or even initiating action. This is a theoretical cost that many will not wish to pay. In order for these ecological perceptual representations to satisfy the impossibility constraint we will likely to have to put them at the

centre of our explanations of everyday action, meaning that their role will become even more controversial.

C. Agential pathology is explained by a combination of *interfering beliefs concerning the agent's lack of ability and degenerated intentions* (satisfying the impossibility and practical significance constraints), and *a sudden and involuntary retreat of noetic bodily feelings in favour of noematic ones* (satisfying the lethargy constraint). This strategy also has the benefit of relative simplicity, and moreover involves positing no controversial ecological perceptual content at all. However, it is likely to struggle with the practical significance constraint; it is unclear whether changes to the content of intentions or to bodily feelings will be capable of explaining all the different kinds of changes to practical significance reported in agential pathology. Determining this will require significant further work in analysing the precise nature of disturbances to the experience of practical significance in agential pathology.

There may well be other explanatory constraints (phenomenological or otherwise) that would help decide between these theories. One proposal, for instance, would be to choose the least qualitatively or quantitatively complex theory. These theories would also, presumably, be open to empirical confirmation or falsification. Moreover, it seems entirely possible that different cases of agential pathology are to be explained by appeal to a somewhat different combinations of factors (i.e. agential pathology is not merely, for all we know, multiply psychologically realisable, but *in fact* multiply psychologically realised). Perhaps this idea goes some way to explaining the different degrees to which people emphasise different phenomenological components of the condition when reporting on it.

It is worth noting that somatic theories exhibit a certain indispensability to viable explanations of agential pathology (though they are insufficient). This is closely tied to the lethargy constraint; no other hypothesised factors involved in agential pathology are clearly able to satisfy it. Nevertheless, it is equally important to not attempt to stretch somatic theories beyond their explanatory limits, as I think Smith

(2013) does. As we saw above, it is at the very least unclear how they account for the wide range of changes identified by the practical significance constraint.

These options, then, are not the final word on how best to explain agential pathology. Rather, I hope that they can act as guides to future research in the area. Regardless of what other conclusions we may draw, the central point here is that the prospects for single-factor explanations are dim, vindicating the suspicions of Smith (2013). Agential pathology is a deeply disparate, and most likely multiply psychologically realised phenomenon that cannot be fully explained by appeal to any single type of psychological process or state.

There is however, a further important lesson to draw from the above. Recall that Oyebode (2015) characterised agential pathology in depression as involving disturbances to motivation and will. An important lesson of the above is that these constructs do not fully encompass the factors involved in agential pathology.

Let's take motivation first. As understood in the previous chapter and elsewhere in the literature (Pacherie 2012), the functional profile of motivation is concerned with the initiation, concrete planning, and online control of action[37]. The functional profile of many of the processes disturbed in agential pathology, however, is concerned with awareness of the coarse-grained action-possibilities in one's immediate environment. For instance, on Smith's account (2013), agential pathology is partly explained by disturbances to bodily feelings that normally determine one's sense of what one can do. And on the simpler versions of the perceptual theory, agential pathology is partly explained by the absence of perceptual contents that represent objects' action-properties. Neither of these processes are properly thought to directly involve action-initiation, guidance, or control. Rather, they are best understood as pre-conditions of the execution of these functions; you cannot initiate, concretely plan, or control actions if you are not appropriately aware of the basic actions you could perform. In general, a similar point holds assuming a pre-theoretic

[37] Oyebode has a simpler understanding that reduces motivation's functional profile to simply action-initiation (2015). I expand this understanding partly to bring it more in line with existing literature, but also so as to ensure that I am not treating motivation unduly narrowly when positing that many of the processes involved in agential pathology are not motivational in character.

understanding of motivation; if I am not aware that doing something is possible, while I might, trivially, not be motivated to do it, that lack of motivation is not considered to be the most salient feature of the situation. People will inform me of the possibility before, for instance, trying in earnest to encourage me to pursue it. That is, they will make me properly aware of the action-possibility before trying to 'motivate' me.

Even fewer of the processes thought to underpin agential pathology are plausibly associated with the will, which in Oyebode's intended sense refers to the ability to make choices and express preferences. Naturally understood, bodily feelings are not directly tied to this sort of capacity (though they might, again, be preconditions of exercising it). Nor are the sorts of perceptual contents that perceptual theories posit are lacking in agential pathology. Perhaps preference expression involves the ability to form appropriate desires, and perhaps making choices (at least in the practical sense of decision-making) should be thought of as necessarily involving the formation of non-degenerated intentions. So some mental state theories might be thought of as suggesting that agential pathology is underpinned by disturbances to the will. But few dimensions of the other theories are easily interpreted in this way. All this together suggests that agential pathology is not fully captured by reference to the constructs of motivation or will.

This is not merely of theoretical interest. When we consider how psychiatric researchers might develop or evaluate interventions to relieve agential pathology, or even how they might communicate more effectively with, and better understand the experiences of, service users, it is critical that they have the right kinds of psychological constructs in mind. If the suggestions made in this paper are correct, it will not be sufficient to investigate how to intervene on the psychological mechanisms underpinning motivation or will, in the psychologically standard sense of those terms. Nor will it be appropriate to communicate to service users that the difficulties they are experiencing are disorders of motivation or will, because it will

not be true in either the technical nor everyday understanding of those terms[38]. Both practical research and everyday service user interaction need to re-orient towards the possible underlying causes of agential pathology that are psychologically prior to any of these constructs.

[38] This is perhaps part of the reason why many depressed people object to being described as lacking in motivation. This term, in both its technical and everyday senses, rarely captures the actual experience of agential pathology.

4.6. References

Beck, A. T. Rush, A. J. Shaw, B. E. & Emery, G. (1979). Cognitive therapy of depression. New York: Guilford Press.

Benson, O. Gibson, S. & Brand, S.L. (2013). "The experience of agency in the feeling of being suicidal". *Journal of Consciousness Studies 20*(7-8): 56-79

Bubl, E. Kern, E. Ebert, D. Bach, M. Tebartz van Elst, L. (2010). "Seeing gray when feeling blue? Depression can be measured in the eye of the diseased". *Biological Psychiatry 68*(2): 205-208

Bubl, E. Tebartz van Elst, L. Gondan, M. Ebert, D. & Greenlee, M.W. (2009). "Vision in depressive disorder". *The World Journal of Biological Psychiatry 10*(4): 377-384

Colombetti, G. & Ratcliffe, M. (2012). "Bodily feeling in depersonalization: A phenomenological account". *Emotion Review 4*(2): 145-150

Fitzgerald, P.J. (2013). "Gray coloured glasses: Is major depression partially a sensory perceptual disorder?". *Journal of Affective Disorders 151*(2): 418-422

Fuchs, T. (2005). "Corporealized and disembodied minds: A phenomenological view of the body in Melancholia and Schizophrenia". *Philosophy, Psychiatry, & Psychology 12*(2): 95-107

Gibson, J.J. (1979). "The Ecological approach to visual perception". Boston: Houghton Mifflin

Golomb, J.D. McDavitt, J.R.B. Ruf, B.M. Chen, J.I. Saricicek, A. Maloney, K.H. Hu, J. Chun, M.M. & Bhagwagar, Z. "Enhanced visual motion perception in Major Depressive Disorder". *The Journal of Neuroscience 29*(28): 9072-9077

Hall, R.J. (2008). "If it itches, scratch!". *Australasian Journal of Philosophy 86*(4): 525-535

Hornstein, G. (2009), "Agnes's Jacket: A psychologist's search for the meanings of madness". New York: Rodale Books

James, W. (1902). "The varieties of religious experience: A study in human nature". New York: Longmans, Green & Co.

Kendler, K.S. (2016). "The phenomenology of Major Depression and the representativeness and nature of DSM criteria". *The American Journal of Psychiatry 173*(8): 771-780

Klein, C. (2007). "An imperative theory of pain". *The Journal of Philosophy 104*(10): 517-532

Kuhl, J. (1984). "Volitional aspects of achievement motivation and learned helplessness: Toward a comprehensive theory of action control". *Progress in Experimental Personality Research 13*: 99-171

Kuhl, J. & Helle, P. (1986). "Motivational and volitional determinants of depression: The Degenerated-Intention hypothesis". *Journal of Abnormal Psychology 95*(3): 247-251

Law, I. (2009). "Motivation, depression and character" in Broome, M. & Bortolotti, L. (eds.). "Psychiatry as Cognitive Neuroscience: Philosophical perspectives". Oxford: OUP. 351-364

Lewis, G. (2006), "Sunbathing in the rain: A cheerful book about depression". London: Jessica Kingsley

Muncie, W. (1939). "Psychobiology and Psychiatry: A textbook of normal and abnormal human behaviour". St Louis: Mosby

Nanay, B. (2011). "Do we see apples as edible?". *Pacific Philosophical Quarterly 92*(3): 305-322

Nanay, B. (2012). "Action-oriented perception". *European Journal of Philosophy 20*(3): 430-446

Nanay, B. (2013). "Between perception and action". New York: OUP

Oyebode, F. (2015). "Sims' symptoms in the mind: Textbook of Descriptive Psychopathology". London: Elsevier

Plath, S. (1963). "The Bell Jar". New Hampshire: Heinemann

Ratcliffe, M. (2015). "Experiences of depression: A study in phenomenology". Oxford: OUP

Roberts, J.R. (2001). "Mental illness, motivation and moral commitment". *Philosophical Quarterly 51*: 41-59

Siegel, S. (2014). "Affordances and the contents of perception". in Brogaard, B. "Does perception have content?". Oxford: OUP. 51-75

Slaby, J., Paskaleva & Stephan, A. (2013). "Enactive emotion and impaired agency in depression". *Journal of Consciousness Studies 20*(7-8): 33-55

Smith, B. (2013). "Depression and motiovation". *Phenomenology and the Cognitive Sciences 12*(4): 615-635

Smith, M. (1987). "Reason and desire", *Proceedings of the Aristotelian Society 88*: 243-258

Smith, M. (1994). "The moral problem". Oxford: Blackwell

Starkstein, S.E. Petracca, G. Tesón, A. Chemerinski, E. Merello, M. Migliorelli, R. & Leiguarda, R. (1996). "Catatonia in depression: prevalence, clinical correlates, and validation of a scale". *Journal of Neurology, Neurosurgery & Psychiatry 60*: 326-332

Stocker, M. (1979). "Desiring the bad: An essay in moral psychology". *Journal of Philosophy 76*: 738-753

Watzl, S. (2014). "Perceptual Guidance". *Ratio 27*(4): 414-438

Wu, W. (2011). "Confronting many-many problems: Attention and agentive control". *Nous 45*(1): 50-76

Chapter 5: Depression, empathy, and experiential difference

5.0. Abstract

Matthew Ratcliffe has argued that experiencing depression will tend to hinder more than it helps a person's ability to empathise with *other* people's experiences of depression (2015: 247-8). Call this the *Bad Similarity Claim* (henceforth, BSC). In this paper, I evaluate three arguments for BSC. The first will be Ratcliffe's own, based on his somewhat idiosyncratic *Difference View* of empathy. I conclude that, even granting Ratcliffe his view and supporting claims, the argument fails to secure BSC, properly understood. The second and third are my own suggestions, based on the more standard but also more restrictive *Imagination View* of empathy (Coplan 2011). I argue that though one of them also ultimately fails, the second offers some evidence that BSC could be true. I also argue that the *way* in which BSC might turn out to be true (i.e. what *kind* of empathic failure is most likely to be involved) gives us some interesting insight into the nature of interpersonal disability in depression more generally; failures of empathy between two depressed people, if they are indeed any more likely than would be expected pre-theoretically, will probably involve the empathiser getting *too close* (emotionally speaking) to empathise properly, rather than failing to bridge the gap between themselves and the other.

5.1. Introduction

In his book-length phenomenological study of depression, Matthew Ratcliffe has argued that experiencing depression does not improve a person's ability to empathise with other people's experiences of depression (2015: 247-8). On the surface, this claim presents depression as being rather unlike most other sorts of psychological similarity, which we have good evidence to think generally aid empathising with relevant aspects of a person's experiences (Hoffman 2000; Eisenberg 2000; Chouliaraki 2006; Gutsell & Inzlich 2010). Call this the *Bad Similarity Claim* (BSC). BSC also has significant and surprising consequences for how effective we should expect group therapy and peer support to be for Depression, what our view should be on the probable utility of collective advocacy amongst people with Depression, amongst other important issues.

This paper aims to investigate BSC more thoroughly, both within and outside of the context in which Ratcliffe proposes it. In section 5.2. I evaluate three arguments for it. Firstly, I present and examine the argument that Ratcliffe himself makes for BSC, based on his own *Difference View* of empathy. Even granting Ratcliffe the Difference View, I conclude that this argument fails, because Ratcliffe does not pay sufficient attention to potentially interesting features of the case where one depressed person tries to empathise with another. Instead, he overgeneralises from an argument he makes for a non-specific empathic deficit in depression. Secondly, I propose two novel arguments based on what I term the *Imagination View* of empathy, due to Amy Coplan (2011). I conclude that though one of these arguments also fails to provide good evidence for BSC, the other shows some promise. Thus I conclude that if one is tempted by the Imagination View, one should also be amenable to the possibility of BSC.

In section 5.3. I also argue that the *way* in which BSC might turn out to be true (i.e. what *kind* of empathic failure is most likely to be involved) gives us some interesting insight into the nature of interpersonal disability in depression more generally; failures of empathy between two depressed people, if they are indeed any more likely than would be expected pre-theoretically, will probably involve the

empathiser getting *too close* (emotionally speaking) to empathise properly, rather than failing to bridge the gap between themselves and the other.

This section provides the background necessary for fully understanding the arguments for BSC in section 5.2. There are two main parts. In the first sub-section I render BSC in a sufficiently precise way that it can be a clear target of evaluation and argument, as well as making some remarks about why such a claim is of special interest. In the second, I introduce the two theories of empathy that BSC will be evaluated in light of.

5.1.1. The Bad Similarity Claim

In this section I will present and explain a precise version of BSC. This claim, which Ratcliffe posits as a consequence of his views on empathy and empathic difficulties in Depression (2015: 247), will be the target of discussion in the rest of the paper.

5.1.1.1. What is BSC?

Ratcliffe (2015: 247-48) mentions BSC somewhat in passing, saying initially that "there are reasons to doubt…that depressed people are better able to understand the experiences of other depressed people" (247). But that is not to say that he asserts it carelessly, or does not attempt to make an argument for it. He has clearly given his statement some thought, and believes that the preceding two chapters have provided an argument for it, or at least that it is a corollary of conclusions drawn in those chapters. He writes,

> Depression does not enhance the ability to engage with *someone else's* depression, to recognize the particularity of her experience. *As I have argued, the potential for this kind of interaction is absent from the world of many depressed people.* (Ratcliffe 2015: 247; *emphasis mine*)

To be clear, Ratcliffe does not mean to deny that shared depression may enhance a generic kind of understanding, or engagement with experience, such as might coincidentally emerge from typifying (this person falls under the type 'depressed', as do I) or projecting (this person is depressed and consequently feels the same way I do). Instead, the kind of understanding he is talking about is best thought of (as he does) as a distinctly *empathetic* kind of understanding, whereby someone is able to

"recognize the particularity of [the other's] experience" (247). Nor does he intend to deny that *past* experience of depression might aid empathy with somebody who is currently depressed (2015: 247). His claim is intentionally restricted to individuals who are *both* depressed at the time one tries to empathise with the other.

It is crucial to get very clear on exactly what claim Ratcliffe *does* intend to make here, since he sometimes appears to vacillate. On the one hand, what he initially asserts is a simple negated proposition; that depression *does not enhance* the ability to empathise with somebody else's experiences of depression. But the support he invokes for this claim in the following sentence seems to suggest something stronger; that depression might rob people of *the mere potential* to empathise. On this second reading, depression *actively impedes* the ability to empathise both in general, and (presumably) with somebody else's experience of depression in particular.

Generally speaking, as noted above, we expect most sorts of psychological similarity to aid, though certainly not guarantee, empathy (Hoffman 2000; Eisenberg 2000; Chouliaraki 2006; Gutsell & Inzlich 2010). Intuitively, this is because the more psychologically similar we are to a given individual, the easier it will be to put ourselves 'in their shoes', as it were. That is, we think of most psychological similarities as *good* similarities, empathetically speaking. At the very least, we expect empathy to be easier in cases where a person is trying to make sense of an experience that has a direct connection to some psychological feature both agents share. Ratcliffe wants to suggest to us that, in some sense, depression is unlike other psychological similarities in this respect. What exactly does this amount to?

Firstly, I do not think that the first reading of Ratcliffe suggested above turns out to be a very charitable one, simply because Ratcliffe endorses, with good evidence (see e.g. Brampton 2008: 176), that people with depression generally experience active difficulties empathising with others (Ratcliffe 2015: 237). Thus there is little reason to believe that he thinks that one's own experience of depression *merely fails to help* to empathise with another's experience of depression. Rather, it seems that he thinks that this *more general deficit* is not (significantly) improved in the case that both individuals in question are depressed.

Secondly, I should note that there is a way of reading Ratcliffe's claim where it *does not even* rule out that depression is a good similarity. Assume that we are operating on the basis of the second reading mentioned above, and that we interpret the suggested impediment to empathising with *someone else's depression* as merely an instantiation of the *more general* impediment to empathy that occurs in depression. That is, the impediment to empathy that applies generally (and only this impediment) also applies specifically to the case where the other person is also experiencing depression. Such a reading permits, logically speaking, that depression, *as well as* hindering empathy in general, nevertheless plays a strong supportive function in empathy in the specific case where the person they are empathising with is also depressed. On this picture, while a depressed person's general empathetic capacity is hindered when they attempt to engage empathetically with *anyone* (including another depressed person), it might still be true that their shared experience provides an overwhelming aid to empathy in the specific case where the target of their empathy is also depressed.

While this second reading strictly permits of such a situation, it is, I think, clearly against the spirit of Ratcliffe's claim. There would be no point in speaking negatively of the specific situation of *shared* depression if one merely wished to make a point that was covered by noting that depressed people experience a generic empathetic difficulty. Nevertheless, I don't think Ratcliffe necessarily intended to rule out that depressed people might experience *some benefit* when attempting to empathise with another depressed person. But he certainly intended to suggest that he had provided good reason to think that whatever benefits might accrue were fewer and less potent than the relevant disadvantages. In my view then, Ratcliffe's claim is ambiguous between a stronger and weaker version of the claim that depression is a bad similarity, empathetically speaking.

> [**BSC-Weak**] The fact that two people, x and y, are both depressed at time t impedes x's ability to empathise with y at t to a greater degree than it enhances that ability.

[**BSC-Strong**] The fact that two people, *x* and *y*, are both depressed at time *t* impedes *x's* ability to empathise with *y* at *t*, and does not enhance that ability to any degree.

The weak version of this claim says that the empathetic deficit common to depression *outweighs* the advantages usually generated by the fact that two people share that psychological feature. The strong version says that it *eliminates* those advantages. When I use 'BSC' in what follows, I am referring to the inclusive disjunction of BSC-Weak and BSC-Strong.

When evaluating the purported arguments for BSC in what follows, I shall consider them in light of both disjuncts. For now, however, I wish to briefly turn my attention to why the status of BSC in general is of significant interest and concern.

5.1.1.2. Why is BSC important?

First of all, whether BSC is true or not will have importance for how we understand and evaluate group therapy and peer support. It would be natural to be suspicious of these interventions for the treatment of Depression if we had good reason to believe that people with Depression typically struggle to empathise with each other's experiences, or even if they were simply no better at it than anybody else. Naturally this suspicion would need to be tempered with an alternative explanation for why such practices seem to be effective, as we saw in the previous sub-section. But minimally, it might lead us to think that the mechanism of interpersonal interventions' efficacy has little or nothing to do with being better understood by one's peers. This would push research into these interventions in a notably different direction than if we had good reason to think that understanding and empathy were at their heart.

Secondly, BSC's truth seems to bear on how we should understand the role of 'experts by experience' (i.e. current or ex-service users) in representing depressed people's interests in formal psychiatric contexts. Experts by experience are typically thought to be able to represent the interests of other people with relevantly similar experiences of mental illness. Presumably, this is not to be justified by the thought that all experiences of particular mental illness-types (such as Depression) are very

similar; it has been noted that Depression is remarkably heterogeneous in its expression (Goldberg 2011). Rather, to my mind, the best epistemic (as opposed to moral or political) justification for the systematic inclusion of experts by experience on psychiatric decision-making bodies is that they are uniquely placed to understand and properly interpret the experiences of people *relevantly like themselves*. That is, their presence is thought to be partly justified by the fact that they are better able to empathise with others in somewhat similar positions to them, not because they know everything that there is to know about a particular condition simply in virtue of having been diagnosed with and directly experienced a single token instance.

Naturally, this observation does not entail that the truth of BSC would warrant the exclusion of depressed experts by experience from psychiatric decision-making bodies. For one thing, there may be good moral and political reasons for doing so that do not depend on their having any kind of especially privileged epistemic position, or at least not one based on improved *empathy* with other members of their community (see Crichton, Carel & Kidd 2017; Miller Tate 2018). Nevertheless, we should be cautious to get our evaluation of BSC right, since it clearly bears on an important question of service user self-advocacy in Psychiatry.

Finally, BSC seems liable to produce a stigmatising effect. The reason for this is relatively simple; if BSC were true, depressed people would look to be somewhat uniquely, and negatively, different from those who are not depressed, with respect to a capacity that is highly socially valued. The ability to empathise, or a lack of it, is clearly a psychological capacity that society (to a significant degree, understandably) values very highly. Moreover, to the degree that we are typically willing to forgive failures of empathy, it is in cases where the person you are trying to empathise with is very unlike yourself. To claim, then, that depressed people are either unable, or less able than we would expect, to empathise with other depressed people, is somewhat dangerous; it is to claim that depressed people are different from the rest of us in both a fundamental and morally charged manner. This may not be sufficient justification for not claiming it, and to assert BSC is certainly not to assert, or even imply, that any negative perception of depressed people that results is justifiable (minimally because one probably ought not to think that depressed people are

responsible for the deficit in question). But it does put a certain kind of moral burden on our evaluations of this claim. If we are to go around asserting that depressed people are unlike people who are not depressed in this kind of highly charged way, then we had better be sure that we are *correct* about it.

All things being equal, then, we should find BSC independently surprising, and think that getting our evaluative judgments about it correct is a matter of more than just purely academic significance.

Moving on, in the final section of this introduction, I will introduce the two views of empathy that I will be looking at in the remainder of the paper. The differences between them will be highlighted, so as to make it clearer in the final analysis why these views might reasonably be thought to offer competing perspectives on the plausibility of BSC.

5.1.2. Two Views on Empathy

I do not claim that the following two views somehow represent an exhaustive summary of positions one might take on empathy, and thus do not claim that what I have to say about them in relation to BSC is the final word on the matter (or even the final word absent further developments in the empathy literature). Nevertheless, they represent an important subsection of the empathy literature, and investigating them sheds light on exactly how one might go about finding evidence supporting or falsifying evidence for BSC.

5.1.2.1. The Difference View

Ratcliffe bases his rather idiosyncratic view of empathy on an analysis of interpersonal deficits in depression, and a rejection of so-called *simulationist* views of empathy; those that require an empathiser to somehow simulate the mental state of the target of their empathy. Here, I call this the Difference View of empathy.

Ratcliffe (2015: 238-240) suggests that a minimal kind of empathy requires satisfying only the following single condition,

[DV] *A* empathises with *B* iff:

(i) *A* is open to profound experiential difference between *A* and *B*.

What this condition amounts to, and why Ratcliffe thinks satisfying it amounts to at least a (if not *the central*) kind of empathy requires a little unpacking. He proceeds by first considering what it is to *feel empathised with* (rather than to empathise). He points out that, especially in certain clinical contexts, feeling empathised with does not involve anything like thinking that another person has first-personally replicated your experience. Rather, it can involve believing that the other person accepts that the two of you experience things in profoundly different ways (2015: 238-239). As Ratcliffe says, quoting Halpern (2001), from the empathisers viewpoint,

> …it is not so much a matter of undergoing a similar experience as "acknowledging that you *don't* fully understand how the patient feels and are curious to learn more". (Ratcliffe 2015: 239)

Moreover, he notes that the sense of *connection* with another human being that such a recognition of difference can bring about in patients is what many people with depression emphasise as missing in their interpersonal interactions. He notes that for these people,

> Someone else is recognized as empathetic when she manages to foster at least some sense of interpersonal connection, the possibility of which might previously have been experienced as *absent from the world*. (Ratcliffe 2015: 240)

These considerations lead Ratcliffe to the conclusion that opening oneself to the possibility that another person experiences the world profoundly differently to yourself counts as a form of empathy[39]. Hence the Difference View gives an expansive account of empathy, encompassing any state or process involving the satisfaction of the above condition.

---

[39] As a sidenote, I struggle with this line of reasoning. Though I can see the *prima facie* attraction of the intuition that reflecting on what it is like to be empathised with might tell us something about what it is to empathise with someone else, I worry that this methodology easily steers us wrong when applied elsewhere. For instance, the feeling of one's peers being critical of a decision you have made *can* involve (or even just *be*) embarrassment, but it does not follow from this that embarrassing a peer suffices for being critical of a decision they have made. I don't consider this concern here, partly so as to avoid going too far off-track, and partly because I suspect that there is something off about drawing this kind of analogy. But I do think that Ratcliffe and his sympathisers need to do more to justify this method of developing an understanding of empathy.

Ratcliffe is careful to head off concerns that satisfying this condition is so simple as to be *trivial*. He writes,

> In our everyday encounters with others, we of course appreciate that our own experiences differ from those of others in all manner of ways. Even so, we continue to take much for granted as shared…it is *us* who inhabit a realm of interconnected artefact functions, norms, and social roles…However, not all attempts to empathise can presuppose so much…when empathizing with a young child, someone from a different culture, or someone with a very different set of interests and values, less can be assumed…Empathizing with others thus involves suspending, to varying degrees, a background of norms, roles, artefact functions, self-interpretations, projects, values, and various other experiential contents. (Ratcliffe 2015: 240-241)

What is crucial here is that suspending many kinds of presumption about what others' experiences share with ours is not a trivial task; it can easily involve suspending that which seems obvious to us and which we would rarely think another might not realise; for instance that the machine at the entrance to the train station is for buying tickets.

This should suffice as an overview of Ratcliffe's Difference View of empathy. I now turn to an alternative, significantly more restrictive view. This should serve as an interesting foil to Ratcliffe's view, thus ultimately giving us a clearer sense of how BSC stacks up in light of two competing views of empathy.

5.1.2.2. The Imagination View

What I call the Imagination View of empathy is due to Amy Coplan (2011). She presents this view as a self-consciously *narrow* understanding of empathy, which seeks to very sharply distinguish the notion from other, related phenomena (2011: 5). Indeed, Coplan's search for specificity leads her to declare that,

> … it is less important that we call this process empathy than that we stop conflating it with several related processes…" (2011: 5)

Coplan's aim is certainly to identify a social cognitive phenomenon that is empathetic in a broad sense, but more importantly to identify a highly precise phenomenon of this broad kind. Her goal contrasts sharply with Ratcliffe's approach; he identifies what he thinks is a very minimal and general notion of empathy (the recognition of some experiential difference) that presumably can then be built upon to produce a number of more sophisticated and specific empathic outcomes.

Coplan's proposed necessary and sufficient conditions for this kind of empathy are as follows,

[IV] *A* empathises with *B* iff:

(i) *A* is in an affective state *S*

(ii) *S* is type-identical to some other state *S'* that *B* is in

(iii) *A* is in *S* by imagining *B's* situation, experiences, and characteristics as if *A* were *B*

(iv) *A* maintains strict differentiation between themselves and *B*

The first two conditions specify what we might call *affective matching*; a person empathising with another thereby enters into an affective state that is of the same kind as theirs. This is a significantly more traditional understanding of empathy than Ratcliffe's, and explicitly rejects his emphasis on acknowledgment of interpersonal difference. Contrary to Ratcliffe, a minimal requirement of empathy, according to Coplan, is the actual *bridging* of the psychological gap between one person and another (2011: 6-7) not the mere acknowledgment of it.

Moreover, by giving the third condition, Coplan insists that this affective matching is achieved in a very particular way. Not only must a person imagine[40] what another is

[40] Naturally, Coplan has a specific notion of imagination in mind. It suffices here to say that when I speak of imagination in relation to IV in the rest of this paper, I am speaking of *recreative imagination*, not merely *suppositional imagination* (Goldman 2006; Coplan 2011). That is, Coplan requires that an empathiser recreates an affective experience of another, rather than just supposing or entertaining the idea of entering into that affective experience.

experiencing, but they must imagine it *as if* they were that other person. To introduce some terminology, this act of imagination must be *other-oriented* rather than *self-oriented*. In Coplan's words,

> In other-oriented perspective-taking, when I successfully adopt the target's perspective [or, rather, imagine B's situation, experiences, and characteristics], I imagine *being the target undergoing the target's experiences* rather than imagining being myself undergoing the target's experiences. (2011: 13)

This differentiates the Imagination View of empathy from one on which empathy can involve imagining what *you* would feel if *you* were "in another's shoes" or "in their position". This kind of imagination would be self-oriented.

Finally, Coplan requires, by way of her fourth condition, that the empathiser preserves "a separate sense of self." (2011: 15). They must not end up "experiencing the other's perspective as [their] own". To do so would be, to borrow a term from Coplan, one of *enmeshing*, rather than empathy; a temporary disturbance to one's sense of self and grip on one's own experiences brought about via the act of other-oriented imagination. Although they *imagine* being the other, an empathiser must,

> …keep separate [their] awareness of [themselves] and [their] own experiences from [their] representations of the other and the other's experiences… (Coplan 2011: 16)

This means, drawing from the work of Martin Hoffman that the empathiser is able to (amongst other things) "distinguish what happens to others from what happens to themselves" (2000: 63). Amongst other things, this suggests that empathisers must not misattribute an affective state of *their own* brought on by imagining the situation of another as being *of the other* (that is, a state of their own that is. merely *precipitated* by their act of imagination, as actually *belonging* to the subject of their imaginative act).

This condition is, I shall argue, very important when evaluating what the Imagination View might tell us about BSC. I suggest that this kind of misattribution is particularly likely when one depressed person tries to empathise with another.

Given these two views, two questions then present themselves for consideration in the remainder of this paper:

> a. Given the *Difference* view of empathy, how much credence ought we to have in BSC?
>
> b. Given the *Imagination* view of empathy, how much credence ought we to have in BSC?

In section 5.2, I shall evaluate the status of BSC in light of both of these views of empathy. I shall conclude that Ratcliffe's offers little or no evidence in its support, and thus that those attracted by it should, in fact, be suspicious of BSC. The Imagination View, however, offers at least some evidence in BSC's favour, though this evidence is not decisive. Finally, in 5.3, I reflect on what this state of affairs teaches us about the nature of empathic and other interpersonal difficulties in depression, and how we should move forward from here in understanding them.

5.2. Evaluating the Bad Similarity Claim

5.2.1. On the Difference View

Ratcliffe states that "there are reasons to doubt that" people with depression "are better able to understand the experiences of other depressed people." (Ratcliffe 2015: 247). The reasons he goes onto provide are varied, but many of them do not clearly amount to arguments for BSC. For instance, he notes that;

> [r]ecognizing that another person is 'depressed like me' *could* involve typifying rather than empathizing, where one infers that both parties fall under the type 'depressed' and therefore have the same kinds of experience. There is also a risk of imposing one's own experience of depression on another person.
>
> (Ratcliffe 2015: 247; *emphasis mine*)

I do not find myself in strict disagreement with Ratcliffe here – mutual recognition amongst depressed people *could possibly* (generally speaking) amount only to typifying or projection rather than empathising. But not only is this not anything close to an argument for BSC (merely hypothesising about what other kind an interpersonal relation *might* fall under does not support the claim that it is not empathy), I think we have good reason to doubt that the relevant kinds of interpersonal interactions between depressed people *are*, generally speaking, either of those two things.

Naturally I allow that typifying or projection will *sometimes* occur between depressed people. But, for instance, in cases of peer support, a hypothesis of typifying or projection does not obviously explain the degree of *being understood* and *connected* that people with depression report in such settings (see Behler et al 2017). This experience of participants in group therapy sharply contrasts with what Ratcliffe notes, correctly, is the typical experience of depressed people when interacting with *non-depressed* family members, friends, or strangers; they feel deeply disconnected and/or misunderstood. For example, respondents to a survey asking about people's interpersonal experiences when depressed wrote,

> #15. …my friends are supportive but struggle to know what to say
>
> #34. I find other people irritating when depressed, *especially those that have never suffered with depression*, and find the 'advice' often given by these is unempathetic and ridiculous
>
> #153. Nobody understands or loves me
>
> (Ratcliffe 2015: 202)

Now, I admit that empathy is not the only route to enhanced understanding of, or connection with, another person. So perhaps depressed people's experience of understanding and connection in peer support is not an outcome of empathy, but instead some other process or state. But whatever this other thing is, we should not think that it is typifying or projection. Experiences of depression are notoriously heterogeneous, and typifying is essentially a process of inductive generalisation; if depressed people were typifying in circumstances of peer support, we would not

expect the results to enhance a feeling in many, if any, others that they are well understood. Instead, we would expect very generic statements that don't really 'speak to' or connect with people as individuals, or (to borrow a phrase from Ratcliffe) the "particularity of [their] experience" (2015: 247). Similar remarks apply to projection. If members of a heterogeneous group project their experiences onto other members of the group, we would not expect those actions to produce in others a feeling of being understood, but rather feelings of confusion or distance. Naturally, in either of these cases, people might sometimes come to feel understood as a matter of dumb luck, but nothing about either typifying or projecting explains these systematic positive experiences of peer support.

After offering these alternative suggestions for understanding interpersonal relationships between depressed people however, Ratcliffe gives a clue as to what his central reason for believing BSC is[41]. He writes,

> Depression does not enhance the ability to engage with *someone else's* depression, to recognize the particularity of her experience. As I have argued, the potential for this kind of interaction is absent from the world of many depressed people.
>
> (Ratcliffe 2015: 247; *emphasis in original*)

When Ratcliffe speaks of recognizing the particularity of someone's experience he is referring to empathy. One feature that Ratcliffe thinks empathy requires is an ability "to relate to others in a distinctively personal way" (2015: 206); that is as proper subjects or persons rather than as novel kinds of objects. Ratcliffe posits that this ability depends critically on the further ability to experience people as loci of particular kinds of (distinctively interpersonal) *possibility* (209-212). These possibilities are those of particular kinds of interaction which are necessarily interpersonal in character (e.g. conversation). So, he thinks that an ability to relate to others in a distinctively personal way depends on the ability to experience others as

[41] In what follows, I am not reporting Ratcliffe's argument exactly as he presents it, for the simple reason that he does not ever present the argument formally, or even in a single location. What I am attempting here is a charitable reconstruction of his reasoning across two chapters of his 2015.

the loci of interpersonal possibilities, such as conversation. He further argues that this latter ability is typically diminished or lost in cases of Depression, either in the sense that one loses the ability to experience people in this way, but retains a sense of what exercising that ability would be like, or more completely, in the sense that one loses all sense of what it is to experience people in such a way.

For our purposes, this suggests that Ratcliffe endorses the following claim.

> [DEPRESSED ENGAGEMENT] For subjects *x* and y, if *x* is depressed, then *x* will typically struggle to experience others as loci of interpersonal possibility, and thus to engage with *y* in a distinctively personal manner.

I shall return to this claim and discuss its precise features below in order to support what I think is a central premise in Ratcliffe's argument for BSC.

Another claim that is critical for Ratcliffe's argument is one that emerges directly from the Difference view of empathy that he endorses, which I discussed in section 5.1.2.1. I will not repeat the justification or discussion of it here. I will, instead, simply go ahead and label it as Premise 1 of Ratcliffe's argument.

> [P1] *x* empathises with *y* only if *x* is open to the possibility of profound experiential difference between *x* and *y*.

Essentially, what Ratcliffe needs is a principle that connects the notions of distinctively personal engagement and openness to the possibility of experiential difference; some kind of linking principle. And while he certainly *seems* to have one in mind, the exact content of this principle is rather unclear. He writes,

> Openness to phenomenological difference, *of a kind that is inseparable from a distinctive kind of second-person attitude*, is necessary for empathy and sufficient for some empathetic achievements. (Ratcliffe 2015: 230)

Let us assume that the distinctive kind of second-person attitude Ratcliffe mentions here is the same attitude that is involved in engaging with somebody in 'a

distinctively personal manner'[42]. Then it seems like he endorses something like the following sketch of a linking principle.

> [LINK] x being open to the possibility of experiential difference between x and y is inseparable from x engaging with y in a distinctively personal manner.

What might Ratcliffe mean here by claiming that these two things are inseparable? One could read him as making an identity claim; that being open to the possibility of experiential difference between you and somebody else *just is* to engage with them in a distinctively personal manner. But this is implausible, and would be an uncharitable reading. There are clearly cases where I, for instance, experience a person as a locus of interpersonal possibility – and thus engage with them in a distinctively personal manner – but am in no way open to the possibility of significant experiential difference between the two of us. As discussed in section 1.2.2, we ordinarily assume that others occupy a shared world with us in the sense that they attribute similar kinds of significance to surrounding artefacts, take themselves to be bound by similar norms, and occupy similar kinds of social role as we do (Ratcliffe 2015: 240). To be open to the possibility of experiential difference between me and somebody else is, minimally, to be open to suspending these presumptions in their case. But my distinctively personal engagement with somebody need not entail that kind of openness. Indeed, my experience of somebody as being a person with whom I could have an enjoyable conversation may *depend* on just these kinds of presumptions. So the relationship here cannot be one of identity, nor can it be the case that distinctively personal engagement generally depends on openness to experiential difference.

My reading is that Ratcliffe intends the dependency relation to run in the opposite direction; openness to experiential difference depends on distinctively personal

[42] This is plausible for the following reason; neither first-personal nor third-personal attitudes require the recognition of another subject. A first-personal attitude recognises only myself, whereas third-personal attitudes require only *something* else (an 'it') separate from myself. A second-personal attitude, however, requires a 'you' to which it is presumed to be directed. So, plausibly, to engage with something (or, rather, somebody) second-personally, as a 'you', is to engage with them in a distinctively *personal* manner.

engagement, in virtue of the former being a sub-type of the latter. Openness to experiential difference is a determinate, distinctively personal engagement a determinable. That is, openness to experiential difference is to distinctively personal engagement as scarlet is to red.

We should think of Ratcliffe as positing that openness to experiential difference is a sort of sensitivity to a particular kind of interpersonal possibility that *in turn* produces a kind of distinctively personal engagement. When one is open to the possibility that a person *does not* share with you some aspect of the background world that you otherwise generally assume *is* shared between you and others, you are open to them as a source of personal epistemic transformation, whereby you will come to understand the contingency of the background world you presume to share with others. Artefacts to which you ascribe particular meanings and kinds of significance may be viewed in other ways; norms to which you take people to be beholden may not regulate others' behaviour; social roles you inhabit may not be complemented by those inhabited by this other person. This kind of possible transformation is still uniquely interpersonal, because it is only by engaging with other *subjects* that you can come to truly understand a difference in a world that is presumed to be *shared*. That is, you are not merely coming to understand a difference between your own understanding of the world and that of another, but that between the 'shared' world you typically presume to inhabit with them and the 'shared' world they presume to inhabit with you. None of this is to say that you are open to the possibility that you are *wrong* in some way; you may be open to this kind of possible difference between you and others while nevertheless thinking that you have somehow got the world *right* and that they have it *wrong*. Nonetheless, openness to the possibility of alternative ways of understanding background reality *is* a recognition of the possibility that the world you presume others share with you is *not* shared by all. And recognition of somebody as the locus of this kind of possibility is to engage with them in a distinctively personal manner, because only another *subject* could be the source of this kind of possibility. A mere object could not suffice.

Call this kind of interpersonal possibility an *epistemically self-transformative possibility*. The claim is that being open to experiential difference *just is* the recognition of epistemically self-transformative possibilities, which in turn necessitates distinctively personal engagement with others.

If all this is correct, then we can sensibly interpret Ratcliffe as endorsing the following two claims.

> [IDENTITY] *x* being open to the possibility of experiential difference between *x* and *y* is identical to x recognising y as the locus of epistemically self-transformative possibilities.
>
> [NECESSITATION] *x* recognising *y* as the locus of epistemically self-transformative possibilities necessitates *x* engaging with *y* in a distinctively personal manner.

It is worth pointing out here that NECESSITATION will not be directly important for Ratcliffe's argument. It is worth noting here, however, for the purposes of ensuring that the link between epistemically self-transformative possibilities and Ratcliffe's notion of distinctively personal engagement is as clear as possible.

Now, when we originally noted that Ratcliffe endorsed DEPRESSED ENGAGEMENT, we were making a very broad and non-specific claim. Recall what it states;

> [DEPRESSED ENGAGEMENT] If *x* is depressed, then *x* will typically struggle to experience others as loci of interpersonal possibility, and thus to engage with *y* in a distinctively personal manner.

What we have since discovered is crucial to understanding the significance of DEPRESSED ENGAGEMENT. Experiencing others as loci of epistemically self-transformative possibilities is a *particularly important* way of producing distinctively personal engagement, because it specifically makes *empathy* – a particular breed of distinctively personal engagement – possible on the Difference View. What is important for Ratcliffe's argument is the following specific principle, which falls out of a proper understanding of DEPRESSED ENGAGEMENT.

[P2] If *x* is depressed, then *x* will typically struggle to recognise *y* as a locus of epistemically self-transformative possibility.

This emerges from DEPRESSED ENGAGEMENT simply because recognising somebody as a locus of epistemically self-transformative possibility is a particular way of recognising them as a locus of interpersonal possibility *simpliciter*.

Here then, I think, are the crucial elements of Ratcliffe's intended argument for BSC. For persons *x* and *y*,

[P1] *x* empathises with *y* only if *x* is open to the possibility of experiential difference between *x* and *y*.

[P2] If *x* is depressed, then *x* will thereby struggle to recognise *y* as a locus of epistemically self-transformative possibility.

[IDENTITY] *x* being open to the possibility of experiential difference between *x* and *y* is identical to x recognising y as the locus of epistemically self-transformative possibilities.

Ergo,

[C1] If *x* is depressed, then *x* will thereby struggle to be open to the possibility of experiential difference between *x* and *y* [P2, IDENTITY, by substitution of identicals].

Assume *x* is depressed. Then,

[C2] *x* will thereby struggle to be open to the possibility of experiential difference between *x* and *y*. [C1, by MODUS PONENS]

[C3] *x* will thereby struggle to empathise with *y* [P1, C2, by MODUS TOLLENS[43]]

[43] Naturally, this is not strictly an application of *modus tollens*, since that deals with strict negations of consequents, not 'struggles' with them. Nevertheless, it is of that same basic form, in the sense that it moves from undercutting a conditional's consequent to undercutting that conditional's antecedent.

[C4] If *x* is depressed, then *x* will thereby struggle to empathise with *y* [C3, by CONDITIONAL PROOF]

Note that this argument does not rely on *y* being depressed. Ratcliffe's argument supports a quite general empathic deficit in Depression. This is a result that he seems to welcome. In support of it, he points out that lots of people who have recovered from Depression admit that they struggled to empathise with other people while depressed (Ratcliffe 2015: 237).

One might naturally ask why I am more interested in BSC than C4. C4 is, after all, the more general claim, and thus would, *ceteris paribus*, normally be considered the more philosophically interesting target. My reasons are twofold; firstly, there is independent evidence that C4 is true, if one restricts the domain of *y* to the neurotypical population (see, e.g. Ratcliffe 2015: 237; Wolkenstein et al 2011; Thoma et al 2011), and in any case I believe that matter to be one that can only be settled by more empirical research. Secondly, for the reasons I gave in sections 5.1.1.2., BSC is a particularly interesting and potentially troubling claim. It marks Depression out, not simply as a condition that *generally* makes empathy harder to achieve, but as being a condition that makes empathy harder to achieve, *even with respect to the experiences of others that are characteristic of the condition itself*.

I think that we can charitably attribute to Ratcliffe something like the following view; that BSC follows as a corollary of C4, or at least that BSC is strongly evidenced by C4. My position is that both of these suggestions are false. The thought that BSC logically follows from C4 is based on a misreading of the latter, and the evidence it provides is weak at best, given other considerations. Thus, the Ratcliffean argument should lend no significant credibility to BSC, even if we accept the premises.

5.2.2.1. Rejecting the Ratcliffean Argument

Recall how I stated the strong and weak versions of BSC;

> [BSC-Weak] The fact that two people, *x* and *y*, are both depressed at time *t* impedes *x's* ability to empathise with *y* to a greater degree than it enhances that ability.

> [BSC-Strong] The fact that two people, *x* and *y*, are both depressed at time *t* impedes *x's* ability to empathise with *y*, and does not enhance that ability to any degree.

One crucial thing to note is that BSC only follows as a logical consequence of C4 if we read the latter far too loosely; that is, as if it states that depression impedes empathy in all cases *and additionally* that this impediment either outweighs or eliminates any advantage that depression might provide when empathising with another depressed person. If *x* impedes *y*, after all, that does not rule out the possibility that *x* also enables or aids *y* in some other way, perhaps to some significant degree. This is not a reading of C4 that the argument licenses; we are only able to obtain the conclusion that the empathiser's depression somewhat impedes the process of empathy. That it might, in some circumstances, make some positive contribution, is a possibility we can't rule out.

So, BSC is not a corollary of C4. But this is perhaps an unreasonably strict criterion to expect; instead, the defender of the Ratcliffean argument should suggest that C4 provides strong evidence for BSC. This, I argue, is also not the case.

To see why, consider what it would take for C4 to count as evidence for BSC. *x* is a depressed person trying to empathise with *y*, who is also depressed. According to C4, *x* will be impeded in their efforts by the fact that they are depressed. So far so good. But for C4 to thereby count as evidence for BSC, we must also think one of two additional things: Either 1) that in this particular situation (where *y* is also depressed) the similarity between *x* and *y* will not provide any benefit (for BSC-Strong to come out true) or 2) That whatever benefit *is* conferred is unable to outweigh the impediment (for BSC-Weak to come out true). To the contrary, I suggest that we have good reasons to doubt both of these claims, even if we operate strictly on the assumption that the Difference View is accurate.

The first reason is that we have empirical evidence from studying of peer support groups for those with depressive illness that interactions between depressed people exhibit significantly more successful empathy and understanding than those between depressed and non-depressed people (Behler et al 2017). Depressed people

often report a feeling of being deeply, and uniquely, understood when discussing experiences of depression with other depressed people, and often attribute this understanding to their shared experiences. For instance, as one 41-year-old woman wrote of her experience of peer support for people suffering from depressive disorders;

> I was able to come here and explain what was going on and I felt like the people are really sincere here and *truly do know what is going on*.
>
> (Behler et al 2017; *emphasis mine*)

And as Behler and colleagues summarise;

> Participants repeatedly stated that Peer Support provided *an experience of mutual understanding from people who shared similar experiences* without the perceived constraints…existing in most professional relationships.
>
> (Behler et al 2017: 222; *emphasis mine*)

Behler and colleagues explicitly attribute this mutual understanding to a high degree of empathy experienced between group members (2017: 221). Naturally, it is conceivable that the high degree of understanding reported by these service users is *not* the result of empathy. But the suggestion that enhanced empathy between people who share current experiences of depression is what accounts for this feeling of mutual understanding is, I think, a natural one. Yet if this thought is correct, then we have good reason to suspect that not only is there a empathetic benefit to be gained by shared experience of depression, but that it is a large one; large enough to be a transformative experience for many who attend group therapy. Ratcliffe's reasoning gives us no reason to doubt that this is the case.

Moreover, there is evidence that support from peers who currently share experiences of Depression is a highly effective intervention that significantly aids recovery from such conditions. A systematic meta-analysis (Pfeiffer et al 2011) found strong evidence that peer support interventions (i.e. interventions where no specific group-therapeutic approach was utilised, and where support groups functioned just as forums for open discussion of experiences and difficulties) were more effective in

promoting recovery from Depression than usual care (i.e. one-to-one therapy, or psychopharmacological interventions). Such interventions were also *just as* effective on average as group Cognitive Behavioural Therapy (Pfeiffer et al 2011), which is itself recognised as an effective intervention for reducing depression and related symptoms of low self-compassion and worthlessness in a variety of circumstances (Duarte et al 2009; Gilbert & Procter 2006). Given the increased efficacy of these interventions when compared with individualised interventions, it stands to reason that the mechanism of improvement has a distinctly interpersonal dimension.

Once again, while it is possible that the mechanism that produces this effect is not dependent on successful empathy *per se*, that is a natural hypothesis. Indeed many already-hypothesised mechanisms of improvement in peer support, including decreased feelings of isolation and sharing of effective health and self-management information (Dennis 2003) plausibly depend on successful empathy between peers. This is because depressed people often attribute their feelings of isolation to a *lack* of empathy from others (Ratcliffe 2015: 224-225), and it is hard to envisage why sharing strategies of self-management would be helpful unless this sharing were based on a relatively sophisticated understanding of others' needs. So, it is not obvious that either of these mechanisms could be responsible for improvement in peer support contexts unless depressed people were empathising with each other *more often or to a greater degree than* people without depression are able to.

That said, I doubt that such evidence is good enough for our purposes on its own. Partly, this is because it is not clear that such studies make use of a concept of empathy consistent with the Difference View (to the extent that they make use of a rigorously articulated concept of empathy *at all*). Further, it would be illegitimate to ignore the fact that there seem to be viable interpretations of the data that don't make use of the concept of empathy at all, but rather any other route to interpersonal understanding.

My second reason to think that Ratcliffe should believe that the situation where both parties are depressed is empathetically special, in a positive sense, is not only consistent with, but actually hinted at by the details of the Difference View. It is true

that what is critical to empathy on this view is openness to experiential difference, which we saw was to be identified as sensitivity to the possibility of epistemic self-transformation. Similarity between people is not, consequently, directly an aid to empathy, because empathy is not primarily about bridging gaps between people, but openly acknowledging them. Nevertheless, I think there is a sense in which we can expect psychological similarities to aid empathy; by making it *easier* for individuals to be open to what experiential differences there are between them and the person in front of them. That is, I suggest that psychological similarities such as Depression dispose people to better recognise another's status as a locus of epistemically self-transformative possibilities.

I propose that there are at least two ways in which psychological similarity might be expected to enhance people's capacity to be open to experiential difference. The first way is that such similarities make the acknowledgment of difference *affectively* easier. When we have reason to believe that people are like us, it intuitively makes any potential differences between us feel less extreme – that is, if I believe that you and I share a psychological feature (a collection of intentions, desires, beliefs, emotions, etc), I am less likely to feel as if the differences between us are insurmountable. Plausibly in such circumstances, I will be better disposed towards putting in the effort to suspend assumptions I may have, because I will not worry that it will reveal sensitive or emotionally-charged differences between us, or that my efforts at even beginning to understand you are hopeless.

The second way is that such similarities make the acknowledgement *cognitively* easier. If you and I share a range of psychological features, there will generally be fewer assumptions that I must suspend in order to be open to the remaining differences. Consequently, the difficulty of the act of opening myself up to profound experiential difference is reduced. The more psychologically similar I am to you, the fewer background assumptions I will need to suspend in order to be genuinely open to the way in which your world works differently to mine (all else being equal). This is because the more psychologically similar we are, the fewer are the ways in which our experiences of the world *could* radically diverge, assuming (as I think we should) that it is various kinds of psychological differences (beliefs, desires, intentions,

perceptions, and so forth) that underpin profound experiential difference. And, presumably, having to suspend fewer ingrained assumptions makes the cognitive process a simpler one than it might otherwise have been.

Thus, I contend that we have good reason to suspect that the case of one depressed person trying to empathise with another depressed person is special, empathetically speaking. Although people with depression struggle with acknowledging experiential difference *in general*, they are likely to struggle much less when confronted with people who share (otherwise rare) psychological features common to cases of Depression. Hence we have no reason to believe that C4 provides good evidence for BSC, even if the argument provided for the former is sound.

5.2.3. On the Imagination View

So, Ratcliffe's own view of empathy provides little support for BSC. Given that his view is rather idiosyncratic, any argument for BSC on the basis of a different view is likely to look rather different. But that does not mean it is not worth attempting. BSC is interesting independent of the Difference View of empathy. As it turns out, I shall argue that, somewhat ironically, Ratcliffe's endorsement of BSC would be rather more defensible were it to be supported by a view of empathy rather far removed from his own.

All that lies ahead of us. Let us begin by recalling the Imagination View of empathy discussed in section 5.1.2.1.

> [IV] *A* empathises with *B* iff:
>
> (i) *A* is in an affective state *S*
>
> (ii) *S* is type-identical to some other state *S'* that *B* is in
>
> (iii) *A* is in *S* by imagining *B's* situation, experiences, and characteristics as if *A* were *B*
>
> (iv) *A* maintains strict differentiation between themselves and *B*

The conjunction of conditions (i) and (ii) can be called the *Affective Matching* (AM) condition. (iii) I shall call the *Other-Oriented Imagination* (OOI) condition. (iv) is the *Self-Other Differentiation* (SOD) condition. I shall take this view of empathy to support

BSC if I can find good evidence for both of the following two claims; that (a) one or more of these conditions are significantly harder to satisfy when depressed, and (b) this difficulty is not at all reduced (for BSC-Strong to be true) or likely to be overcome (for BSC-Weak to be true) when the other person is also depressed.

I shall suggest two routes one might go in presenting such evidence. Firstly, one might draw on the specific form taken by broader social cognitive deficits in depression to undermine OOI. I call this the Argument from Mental State Reasoning. Secondly, one might draw on the phenomenon of depressive emotions to undermine SOD. I call this the Argument from Depressive Emotions). I shall conclude that the first route is not very promising, but that the second provides some evidence for BSC. Though neither option *secures* BSC, the second supports the idea that if one endorses OOI and SOD as necessary conditions on empathy, then one should also take BSC seriously. This is because it highlights a potential defeater to these conditions that applies specifically to cases where one depressed person tries to empathise with another depressed person.

5.2.3.1. Decoding, Reasoning, and Undermining OOI

A common distinction in the literature on social cognition, and especially Theory of Mind (Frith & Frith 2006) is that between mental state decoding, and mental state reasoning. Both are processes that (when they work properly) aid attribution and understanding of others' mental states, and they normally operate in tandem to create a relatively detailed understanding of another's mind (Tager-Flusberg & Sullivan 2000).

Mental state decoding is usually defined as the process of attributing mental states to others on the basis of directly observable social information, such as posture or facial expression. This is typically thought to require relatively basic mental processes. Mental state reasoning, on the other hand, involves attributing mental states on the basis of further contextual information about the whole person and their situation, which is typically thought to require higher-order, relatively complex, psychological processes (Wolkenstein et al 2011: 104).

Larissa Wolkenstein and colleagues (2011) found that depressed people's social cognitive abilities were significantly impaired (compared with those of healthy controls) with respect to mental state reasoning, but not mental state decoding (2011: 109). That is, they struggled to integrate contextual information into the process of mental state attribution, but did not show any significant deficit in their ability to attribute basic mental states on the basis of directly observable social information (with the possible exception of those in the study who were most *severely* depressed).

What might this knowledge of the specific nature of social cognitive deficits tell us about how depression might impede empathy, on the Imagination View? My suggestion is that we might interpret the mental state reasoning deficit as impeding OOI, by thinking of it as a failure of imagination.

It seems plausible that the metal state reasoning deficit might reflect a difficulty accurately or precisely *imagining* another's situation, broader mental state profile, personality traits, and so forth. This would explain why such information was not suitably integrated into the process of mental state attribution; it was simply not available for the depressed person to make use of. Instead, they were forced to make use of only the rather limited information provided by processes of mental state decoding. On such an interpretation, the mental state reasoning deficit noted by Wolkenstein and colleagues would make it more difficult for a person to affectively match with another *in virtue of imagining* their situation, experiences, and characteristics, simply because these acts of imagination are harder for such a person to successfully perform. OOI would thus be undermined.

I shall consider this difficulty in relation to the special case where both parties are depressed in section 5.2.3.3. For now, I shall consider an alternative way in which empathy may, on the Imagination View, be disturbed in depression, this time by examining a route to undermining SOD.

5.2.3.2. Depressive Emotions and Undermining SOD

One of the most common features of depression is regular, intense, negative emotional experiences. These most commonly resemble sadness (Oyebode 2015), guilt (Ratcliffe 2010), or anger (Abi-Habib & Luyten 2013), and are not generally

clearly caused by the kinds of events in the immediate environment that would be expected to trigger ordinary instances of such emotions. Call tokens of these emotions 'depressive emotions'.

What might this feature of depression tell us about how depression might impede empathy, on the Imagination View? My suggestion is that a strong disposition to experience intense negative emotions in the absence of intense external triggers, might, in combination with a successful act of other-oriented imagination, produce a significant risk of failing to satisfy SOD.

This would work in the following way. Imagine that a depressed person, *x* has successfully engaged in an act of OOI with respect to a potential target of empathy, *y*. Grant, as is likely, that *x* regularly experiences depressive emotions. We have no reason to suspect that a successful act of OOI should prevent the onset or lessen the severity of depressive emotions (otherwise, a depressed person could simply prevent or overcome such an episode by imagining themselves to be somebody else). Hence, *x* may feel intense negative emotion (as a result of their depression) while imagining themselves to be *y* (with all of *y*'s relevant characteristics, contextual features, and so on, in place).

I propose that in such a situation there is a significant risk of enmeshing (as defined in section 5.1.1.2), in the following sense; *x* is liable to misattribute their own depressive emotion to *y*, since it is an experience that intrudes without obvious cause during an act of OOI. One reasonable explanation of that occurrence is that it is simply part of y's experience of depression (that x presumes themselves to be effectively imagining). Hence this situation involves a significant risk of a failure to satisfy SOD, and thus a significant risk of empathetic failure (insofar as one endorses SOD). In such a situation, *x*'s misattribution of their own experience to *y* blurs their subjective differentiation of themselves and *y*, even if they retain a sense that they are not *identical* to *y*. While they presumably (normally) retain enough of a sense of self to not mistake themselves for *y*, this is not the same as retaining a *strict* sense of difference. This is violated as soon as they fail to recognise to whom a particular experience properly belongs.

Most depressed people experience depressive emotions. Indeed, one might even deem such experiences necessary if one is to count as 'depressed'. But even if one thinks that some depressed people might not experience depressive emotions (or not very often), this is not a strong objection to the argument. This is because the same could be said of almost any feature of depression one might appeal to in this debate. It will suffice for our purposes, I think, that depression very commonly involves such experiences, and hence that any resulting empathetic difficulties will thus be correspondingly common.

One might also object that even those who *do* experience depressive emotions do not typically experience them constantly, and hence that they can, at most, only get in the way of empathy when they are actually ongoing. In a sense, this is true. Nevertheless, depressive emotions are a common enough feature of depression that one can sensibly think that most people with depression are at least *significantly disposed* to such states of mind. Assuming the truth of the SOD and OOI conditions, this ongoing risk of depressive emotional experiences is an extra risk of empathetic failure that a depressed person must deal with, which their non-depressed peers need not generally consider. Thus depression, quite generally, makes empathy more difficult to achieve, at least on average.

I have so far presented two arguments for ways in which empathy (understood in a manner consistent with the Imagination theory) might be disturbed in depression. What remains to be shown is whether either of these arguments point to difficulties that cannot be reduced, or even that might be enhanced if the target of that empathy is also depressed. I shall argue that the first argument provides no such evidence, but that the second does. This suggests that the Argument from Depressive Emotions gives us reason to take BSC seriously, but nevertheless fails to definitively *secure* its truth. It also suggests directions for future research that may be able to verify or falsify BSC, at least on the basis of the Imagination View.

5.2.3.3. Rejecting the argument from mental state reasoning

Difficulties integrating contextual information into reasoning about others' mental states can be readily interpreted as a difficulty successfully satisfying the OOI

condition. This certainly suggests that people with depression will have a generic difficulty with empathy. But we already know that is insufficient to get us anywhere near to BSC. Do we have any reason to think that this difficulty cannot be offset in the case where the target of the empathy is also depressed?

My answer is no. In fact, I think that quite the opposite is likely; people with depression are *less likely* to struggle with other-oriented imagination when the other person is also depressed. The reason for this suggestion is simple; the less radical the difference between somebody's situation, experiences and characteristics and my own, the easier it will be for me to imagine these features of their life. And one depressed person's situation, experiences and characteristics are likely to be more similar to those of another depressed person than somebody who is not depressed (at least on average). That is to say, even if a depressed person generally finds other-oriented imagination harder than most, this burden is nevertheless likely to be significantly lessened when they attempt to imagine the experiences of another depressed person. This certainly undercuts BSC-Strong, and probably even gives us strong reason to doubt BSC-Weak. This is because the experiences of one depressed person are likely to *significantly* more closely resemble those of another than those of a person is not depressed. We should believe this simply because of the *enormity* of the *gap* between the degree of understanding and sense of similarity depressed people report when interacting with those who are not depressed, as opposed to those who are (Behler et al 2017).

Consequently, imagining another's situation will be significantly easier for a depressed person in the case that the target of their empathy is also depressed.

Thus this argument provides little or no support for BSC; it falls into basically the same trap as Ratcliffe's attempt does, by identifying a kind of empathetic difficulty plausibly associated with depression that is significantly *less likely* to obtain when the target of that empathy is also depressed.

5.2.3.4. Supporting the argument from depressive emotions

So much for the argument from mental state reasoning. What about our final option? The Argument from Depressive Emotions, as presented above, suggests that people

with depression are at a heightened risk of enmeshing (a failure of SOD) when attempting to empathise. In order to provide some evidence for BSC, however, we at least need reason to believe that this risk is not mitigated by the target of the empathy also being depressed. In this final substantive section, I shall explain why I think we have a reason to believe that in such a situation the risk is actually *worsened*.

Consider the following toy case describing a failure of one depressed person to empathise, in the sense of the Imagination View, with another depressed person:

> **Imagine Jo is depressed, and is attempting to empathise with Kit, who is also depressed. Jo successfully imagines Kit's experiences and characteristics *as if they were Kit*, including those characteristic of Kit's depression. This act of imagination, in turn, might reasonably be expected to trigger a depressive emotion in Jo. Given that Jo is presently engaged in a sophisticated act of imagining being Kit, they may misattribute this depressive emotion as belonging to Kit (when it actually belongs to Jo). Such a case is one of enmeshing. Ergo, at that point Jo fails to empathise with Kit.**

Notice that at least two key moves in this case look significantly more plausible because *both Jo and Kit* are depressed.

Firstly, it is significantly more plausible that Jo's act of imagination should trigger a depressive emotion when Kit is also depressed, due to the fact that the kinds of experiences being imagined are presumably those particularly likely to trigger such emotions. On the plausible assumption that depressive emotions have *some* cause, and that they are not the ordinary external events that trigger non-depressive instances of such emotions, the actual cause presumably typically has something to do with the experiences and psychological characteristics of the depressed person themselves, which are exactly what are being imagined.

Secondly, it is also significantly harder for Jo to reliably determine to whom the depressive emotional experience properly belongs in such a case. If Kit were not depressed, then Jo would be less inclined to attribute a depressive emotion to them

(because they are characteristic of depression, and rare in the absence of this or similar conditions). But in this case, either of them could plausibly be the 'owner' of the depressive emotion in question, significantly increasing the risk of misattribution, and hence enmeshing.

So, this case gives us reason to think that experiences typical of depression can cause not only a particular kind of empathetic failure, but moreover, that this sort of failure is *more, not less* likely when the target of that empathy is also depressed. Thus, it gives us some reason to accept BSC-Weak. Though I have not done enough to guarantee it, if one accepts my premises (some of which are, I grant, controversial), we should take seriously the possibility that depressed people have a significant *disadvantage* that comes specifically with trying to empathise with other depressed people; a heightened risk of enmeshing.

Given further study, this suggestion could even undercut our earlier presumption that existing successes in group therapy, peer support, etc, for treating depression are to be explained by a specifically *empathetic* mechanism between service users. Perhaps these cases are ones where a lot of enmeshing is going on instead. It's not at all clear that this would be a *bad thing*, of course. In cases where the depressive emotion that is misattributed is nevertheless type-identical to one that the target of empathy is actually experiencing, the empathiser might (in a sense, by luck) still achieve insight and understanding that could be therapeutically useful. One might even think that such cases would be relatively common, on the assumption that there is a relatively systematic connection in depression between a person's situation, experiences and characteristics, and the specific depressive emotions they experience. A depressed empathiser who enmeshes in the way described above might do so precisely because they successfully imagined the experiences that gave rise to the depressive emotion in question in the target of their empathy.

If enmeshing of this sort was found to be common in group therapeutic and peer support contexts, then this would provide a principled reason to doubt a major source of evidence for depressed people having *any specifically empathetic* advantage with respect to other's depression, while nevertheless explaining why the situation appears to involve empathic success. And that would be good evidence for BSC-Strong.

I will conclude this section by noting and defusing a potential worry with this line of argument. Suppose that when Jo experiences a depressive emotion, *e*, in response to imagining Kit's situation as if they were Kit, it just so happens that Kit is in an emotional state *e1*, which is type-identical to *e*. Such a case is fairly unlikely, but it is possible and, indeed, more likely when each party is prone to similar kinds of depressive emotional episodes. Here, I wish to claim that AM is satisfied, but SOD is not. Indeed, the likelihood of such scenarios is why I am hesitant about claiming that this argument supports BSC by way of undercutting AM.

One may ask, however, how this situation differs from any other case of empathy, according to IV. It does not seem as if it does. Jo imagines Kit's situation as if they were Kit, and as a direct response to this act of imagination they experience an emotion which is type-identical to Kit's. Surely this counts as empathy, if anything does, according to IV? But then that must mean that I have overstretched the notion of enmeshing. What I have identified going on between Kit and Jo must not have been a real case of enmeshing at all, since it is indistinguishable from a paradigmatically acceptable case of empathy, according to IV. Another way of putting it is that there is no substance to my distinction, in this case, between the relevant emotion being properly attributable to Jo and it being properly attributable to Kit.

The difference, I suggest, becomes clearer when we reflect on *how* the act of imagination brings about the affective matching. In a successful case of empathy, according to IV, your act of imagination works as a kind of open channel between you and the person you empathise with. You are in an affective state *because they are*, by way of an act of imagination. The emotion you are experiencing is, in a sense, attributable to them. In the case of enmeshing you are *not* in that state because the target of your empathy is; your act of imagination set off something in you that *just so happened* to bring you into a matching state. The emotion you are experiencing is not at all attributable to them, only you. These differences are important to recognise, according to IV, *precisely because* you must distinguish them in order to reliably avoid violating SOD.

5.3. Concluding Remarks

This paper has come to two central conclusions. The first is wholly negative; Ratcliffe (2015) does not provide us with a convincing argument in favour of BSC. The second is more positive; the Imagination View of empathy, due to Coplan (2011), offers us a route to a more convincing argument in favour of BSC, though it is far from decisive. Moreover, it has become clear that the status of SOD has important consequences for how we think about empathetic pathology in general. If enmeshing *really does* constitute a failure of empathy, then our views on how empathy may be impaired need to be revised. In a phrase, empathetic failures might be thought to emerge from either a failure to get *close enough* to one's target, or from a failure to retain *sufficient distance*. Empathy, on this sort of view[44], requires the achievement of a relatively narrow degree of personal engagement.

One might think that the kind of evidence that the Imagination View provides for BSC is of a sort that undermines many conclusions one might have thought could be drawn from BSC. In particular, one might think that it doesn't expressly undermine the idea that depressed people have an advantage when it comes to gaining deep, personal understanding of another's depression. This is because it seems like enmeshing can lead to this outcome as well, and is particularly likely to do so where both parties are depressed (and consequently are likely to share dispositions to specific depressive emotions where one party is imagining what it is like to be the other). This thought is, I think, correct up to a point. Pervasive enmeshing is not, for instance, obviously a bad thing in group therapeutic and peer support contexts (though it is not obviously a wholly *good* thing either). To the extent that an advocate of BSC wants it to count as a reason against such practices then, this will feel like a pyrrhic victory.

---

[44] Coplan is certainly not the only theorist of empathy to propose such a condition, though not all clearly do so intentionally. For instance, Vignemont & Singer (2006) propose a necessary condition on empathy such that the empathiser *knows* that the person they are empathising with are the source of the mental state that they are matching. This is intended to distinguish empathy from emotional contagion, but on a factive reading of *knows* also precludes cases of enmeshing as described above. This is because in such a case, the empathiser cannot *know* that the other is the source of their matched state since it is not *true*.

But this thought must also be carefully reigned in; heightened risk of enmeshing is not without potential for negative consequences, some of which stand independently of SOD. For instance, successful affective matching is a considerably dicier affair under conditions where the risk of enmeshing is high, for the following reason; any matching that occurs is no longer plausibly driven by a mechanism that tracks the mental states of the other person. That is, even if successful affective matching sometimes occurs in cases of enmeshing, and even if we think this is more likely where both parties are depressed, such cases are precarious. They occur for reasons that are not connected to the proper functioning of the ability to imagine, observe, or reason about another's mind[45]. This relative precarity of a depressed individual's ability to affectively match with others is a significant disadvantage when it comes to developing understanding others' experiences, independently of whether we think enmeshing by itself necessarily counts as a failure of empathy.

What broader conclusions ought we to draw from all of this, beyond our way of thinking about BSC?

One thing that I wish to emphasise is that the results above suggest that the full nature of empathetic and/or other kinds of interpersonal difficulty in depression is currently poorly understood. Hence, existing empirical research on the phenomenon might not be properly examining all its dimensions. If we assume that there *is* some such deficit that presents relatively consistently across instances of depression (as I think we should), then there still remains an open question regarding what exact form/s that might take.

Through focusing my investigation around BSC, I have uncovered at least one interesting possibility; that an important component of depression's interpersonal pathology might not be a matter of failing or being unable to experience the perspective of another, but rather of being unable to properly disentangle these

[45] Indeed, one might think that this is a reason to accept SOD in the first place. One might think that a state that aims at providing understanding of the kind that empathy does can no more be subject to this kind of luck in meeting its other success conditions than can knowledge *generally* be the result of lucky processes. That is, for a state of any kind to count as empathy, it must be reached exclusively by sensitivity to other's mental states, not by lucky misattribution of one's own.

experiences from one's own. This is due in significant part to the involuntary and intense character of many characteristically depressive emotional experiences. Empirical research into empathic difficulties in depression (and more generally) need to acknowledge that a failure of empathy need not merely be a matter of failing to experience something of what it is like to be somebody else. One may rather fail to distinguish one's own experiences from those of the other, and thus be unable to acquire the kind of vivid understanding of their experience that is typically empathy's goal. And even if such failures are not failures of empathy in the strict sense, they nevertheless result in more precarious routes to achieving significant interpersonal understanding. This is critical for understanding what happens to people's interpersonal capacities when they become depressed.

5.4. References

Abi-Habib, R. & Luyten, P. (2013). "The role of dependency and self-criticism in the relationship between anger and depression". *Personality and Individual Differences 55*(8): 921-925

Behler, J. Daniels, A. Scott, J. Mehl-Mardona, L. (2017). "Depression/Bipolar peer support groups: Perceptions of group members about effectiveness and differences from other mental health services". *The Qualitative Report 22*(1): 213-236

Brampton, S. (2008). "Shoot the damn dog: A memoir of depression". London: Bloomsbury

Chouliaraki, L. (2006). "Spectatorship of suffering". London: SAGE

Coplan, A. (2011). "Understanding empathy: Its features and effects". in Coplan, A. & Goldie, P. (eds.) "Empathy: Philosophical and psychological perspectives". Oxford: OUP

Crichton P, Carel H, Kidd I.J. (2017), "Epistemic injustice in psychiatry", *Psychiatric Bulletin*. 41(2): 65-70.

Dennis. C-L. (2003). "Peer support within a health care context: A concept analysis". *International Journal of Nursing Studies 40*(3): 321-332

Duarte, P.S. Miyazaki, M.C. Blay, S.L. & Sesso, R. (2009). "Cognitive-behavioral group therapy is an effective treatment for major depression in hemodialysis patients". *Kidney International 76*(4): 414-421

Eisenberg, N. (2000). "Emotion, regulation, and moral development". *Annual Review of Psychology 51*(1): 665-697

Frith, C.D. & Frith, U. (2006). "How we predict what other people are going to do". *Brain Research 1079*(1): 36-46

Gilbert, P. & Procter, S. (2006). "Compassionate mind training for people with high shame and self-criticism: Overview and pilot study of a group therapy approach". *Clinical Psychology & Psychotherapy 13*(6): 353-379

Goldman, A. (2006). "Simulating minds: The philosophy, psychology, and neuroscience of mind-reading". Oxford: OUP

Gutsell, J.N. & Inzlicht, M. (2010). "Empathy constrained: Prejudice predicts reduced mental simulation of actions during observation of outgroups". *Journal of Experimental Social Psychology 46*(5): 841-845

Halpern, J. (2001). "From detached concern to empathy". Oxford: OUP

Hoffman, M.L. (2000). "Empathy and moral development: Implications for caring and justice". Cambridge: CUP

Miller Tate, A. (2018), "Contributory injustice in psychiatry", *Journal of Medical Ethics* (online first), doi: 10.1136/medethics-2018-104761

Pfeiffer, P.N. Heisler, M. Piette, J.D. Rogers, M.A.M. & Valenstein, M. (2011). "Efficacy of peer support interventions for depression: A meta-analysis". *General Hospital Psychiatry 33*(1): 29-36

Ratcliffe, M. (2010). "Depression, guilt and emotional depth". *Inquiry 53*(6): 602-626

Ratcliffe, M. (2015). "Experiences of depression: A study in phenomenology". Oxford: OUP

Tager-Flusberg, H. & Sullivan, K. (2000). "A componential view of theory of mind: Evidence from Williams Syndrome". *Cognition 76*: 59-89

Thoma, P. Zalewski, I. von Reventlow, H.G. Norra, C. Juckel, G. & Daum, I. (2011). "Cognitive and affective empathy in depression linked to executive control". *Psychiatry Research 189*: 373-378

de Vignemont, F. & Singer, T. (2006). "The empathic brain: How, when and why?". *Trends in Cognitive Sciences 10*(10): 435-441

Wolkenstein, L. Schönenberg, M. Schirm, E. & Hautzinger, M. (2011). "I can see what you feel, but I can't deal with it: Impaired theory of mind in depression". *Journal of Affective Disorders 132*(1-2): 104-111

Chapter 6: Epistemic Oppression in Psychiatry

6.0 Abstract

In this paper, I outline three kinds of epistemic oppression that have previously been identified in the theoretical literature (hermeneutical injustice, contributory injustice[46], and testimonial smothering). I suggest that they are frequently experienced by service users in formal psychiatric contexts, drawing on the testimony of service users with a range of diagnoses and experiences of mental distress to support this contention. I further identify strategies to remove these kinds of injustices from psychiatric services and defend my proposals against a variety of objections. The general lesson of this paper is that in interactions with service users, it is of paramount importance that practitioners make a greater effort than is currently typical to adopt and present a picture of mental distress that reflects understanding of the underlying complexities, goes beyond a neurological or otherwise over-individualised style of explanation, and centres service user perspectives on the relevant conditions.

6.1. Introduction

Service user involvement in mental health service provision is increasingly viewed as best practice, at least formally, for healthcare service providers across the UK and mainland Europe (Newman et al 2015; Tait & Lester 2005; Rutter et al 2004). In principle, this reflects an acknowledgment that service users not only deserve (as a matter of justice) a say in what sort of care they receive, but that the kind of knowledge they can contribute to such discussions is unique and/or valuable. That said, few beyond service user or survivor activist groups (see, e.g. Mental Health Resistance Network, Recovery in the Bin, Hearing Voices Network) are calling for a 'level playing field' between the opinions of clinicians and those of service users. As Peter Campbell points out regarding the UK context (in a 2001 prediction that has

[46] The arguments I present in connection with contributory injustice in this paper have been previously published (with myself as the sole author) in the Journal of Medical Ethics (doi: 10.1136/medethics-2018-104761). What appears on this topic in this chapter is an expanded version of that paper.

been largely borne out), although some of the interests of service users are now better reflected in the priorities of mental health service provision,

> [a]ction by service users has not touched the clinical authority of mental health workers – an authority, the Green Paper (Secretary of State for Health, 1999) suggests, that will be reinforced in a new Mental Health Act." (2001: 88)

In practice, service users still face many barriers to actually getting their voices heard, whether in primary care contexts when articulating their own immediate experiences and needs, or through programs supposedly designed with the specific intent of enabling their participation (Tait & Lester 2005). Several key negative upshots of this exclusion of service users from deliberation over and delivery of care are well documented, including impoverished states of clinical knowledge regarding the requirements and state of patient wellbeing (Thornicroft & Tansella 2005; Simpson & House 2002) and worse health outcomes for the service users themselves (Simpson & House 2002). Less acknowledged and analysed is the structure and intrinsic harm of these practices of exclusion themselves. My aim in this paper is to take a first step towards plugging that gap.

I generically refer to ways in which service users are ignored, invalidated, and silenced in clinical psychiatric contexts as *epistemic oppression*. Recent work on epistemic oppression in mental health services has focused on Miranda Fricker's (2007) notion of *testimonial injustice* (Crichton et al 2016). On Fricker's construal, testimonial injustice occurs iff a hearer assigns insufficient credibility to a specific instance of a speaker's testimony, due to a prejudicial assessment of some aspect of the speaker's identity. For instance, negative stereotypes regarding women or people of colour may deflate their credibility on particular topics (or quite generally) in the eyes of those who hold them.

Paul Crichton, Havi Carel, and Ian Kidd argue that mental health service users are particularly vulnerable to testimonial injustice – specifically at the hands of the medics, psychologists, and nurses tasked with their care at all points of their interaction with the healthcare system (2016:1). Reasons for this are, according to them, to do with both the general stereotyping of those with any mental illness, as

well as the stigma attached to *particular* psychiatric diagnoses (termed *global* and *specific* factors respectively) (2016: 3-5). This leaves them even more vulnerable to testimonial injustice than people with somatic illnesses, who are themselves more vulnerable than the general population (Blease et al 2017; Crichton et al 2016: 1; Kidd & Carel 2016; Carel & Kidd 2014).

In their short paper, Crichton and colleagues have undoubtedly identified an issue of great importance to both psychiatrists and theorists of social justice, not to mention *anybody* concerned with the just treatment of those suffering from chronic and severe mental distress. To be clear, I do not take exception with anything that their paper says. Rather, I think that there is significantly more to say, and significantly greater consequences of identifying the full range of Epistemic Oppression, than Carel, Crichton, and Kidd are able to make clear. In particular, reform of Psychiatry to mitigate Epistemic Injustice will need to go beyond regular user-led meetings between Doctors and service users and the inculcation in clinicians of individually virtuous epistemic practices (though this will undoubtedly be important).

This paper aims to do two things. Firstly, I will trace and analyse three distinct kinds of systemic epistemic oppression as they occur in contemporary psychiatric contexts, through service user testimony and other sources of evidence. This will require explaining and deploying three extant categories of Epistemic Oppression explored in the philosophical literature; *Hermeneutical Injustice* (Fricker 2007), *Contributory Injustice* (Dotson 2012), and *Testimonial Smothering* (Dotson 2011). As it will turn out, in order to make proper sense of one of the phenomena I am interested in, it will be necessary to adapt the first of these three concepts in a small but interesting way.

Secondly, I will argue that the existence of these three forms of oppression, and the presumed desirability of their elimination, provides strong moral reason to significantly reform psychiatric practices and institutions. In particular, eliminating or diminishing the forms of Epistemic Oppression I identify will require 1) all available information and conceptual tools relevant to the needs of patients being made part of the same shared pool, 2) service users' conceptual tools in particular, developed through experience, being incorporated into the pool of shared epistemic resources of service users and clinicians, and 3) a fundamental recalibration of the

power relationship between patients and clinicians from a hierarchical one towards a horizontal one. That is, not only must current best practice actually be implemented, but more radical adjustments to the material and interpretive environment must occur.

To be absolutely clear, my aim is not to argue unequivocally for the reform of present psychiatric institutions merely to avoid the injustices identified here. Perhaps there are strong conflicting obligations (though I seriously doubt that the degree to which these injustices are occurring is justifiable). My goal here is more modest; I wish to raise the profile of three moral (rather than pragmatic) reasons for the reform of psychiatry, and ensure that they are not overlooked in the ongoing debate.

6.2. Three kinds of Epistemic Oppression

In this section, I shall outline the particular notions of Epistemic Oppression that I wish to highlight in the practice and institution of contemporary Psychiatry. In doing so, I construct a loose theoretical framework through which to interpret and analyse the experiences of psychiatric service users. Proceeding in order, I shall explain the notions of Hermeneutical Injustice (Fricker, 2007), Contributory Injustice (Dotson, 2012), and Testimonial Smothering (Dotson, 2011).

6.2.1. Hermeneutical Injustice

Hermeneutical injustice, for Fricker (2007), always occurs against a certain background of collective hermeneutical resources. Gaile Pohlhaus characterises these in the following way,

> Knowing requires resources of the mind, such as language to formulate propositions, concepts to make sense of experience, procedures to approach the world, and standards to judge particular accounts of experience. (Pohlhaus, 2012: 718)

According to Fricker, hermeneutical injustice is

> the injustice of having some significant area of one's social experience obscured from collective understanding owing to a structural identity prejudice in the collective hermeneutical resource. (Fricker, 2007: 155)

A structural identity prejudice should be thought of as a gap or failure in the collective hermeneutical resources that renders a particular identity group's social experience difficult or impossible to understand, or even articulate, using those resources. Unlike the negative prejudicial stereotypes that are at play in Fricker's account of testimonial injustice, structural identity prejudices are "not primarily located within the cognitive landscape of individual perceivers…" (Dotson, 2012: 29). Rather, they are consequences of an insufficiency in a collection of resources that are collectively developed and used. Moreover, structural identity prejudices typically stem, Fricker argues, from the *hermeneutical marginalization* of the relevant identity group (2007). For a group to be hermeneutically marginalized is for them to be disadvantaged in their ability to contribute to the generation and alteration of the collective hermeneutical resources. They are, simply put, less able to contribute to the development and refinement of those linguistic and conceptual resources required for knowledge. Hence, concepts and language that speak to unique experiences of theirs will be less likely to become common parlance.

Hermeneutical injustice, then, occurs to a person x when there exists a gap in the collective hermeneutical resources that prevent some aspect of x's experience from being understood. Fricker also requires that x is significantly disadvantaged by this. If only these conditions are met, then we call the hermeneutical injustice *incidental*. If however, in addition, x is persistently hermeneutically marginalized, so that they are unable to make a contribution to the collective hermeneutical resource that might alleviate their disadvantage, we say that the hermeneutical injustice is *systematic*.

For instance, in Ian McEwan's novel Enduring Love a man named Jed is stalking Joe, the main character. Jed is a religious fanatic who holds the delusional belief that Joe is in love with him. Joe's attempts to report his constant proselytising and declarations of affection are met with either a lack of interest (from the police, who refuse to act on the grounds that such actions are not criminal) or amusement (from Joe's partner Clarissa). Joe is seen either as complaining about nothing, or failing to 'see the joke'. But the experience significantly distresses Joe, and so counts as a disadvantage for him. Here, clearly, the social background understanding of harassment is failing to provide Joe with the interpretive resources necessary to

properly make sense of his experience, or explain it to others (because of others' dismissal, Joe begins to question his state of mind). But this is not due to any persistent hermeneutical marginalisation. He is a white, middle class, cis man who would, in any other context, have as much ability as anybody to participate in the creation of resources used to make sense of the world. Though severe, the injustice faced by Joe "is localized and one-off" (Fricker, 2007: 158). The hermeneutical injustice described here is incidental.

The same cannot be said for many women in the past who struggled to interpret and communicate their experiences of sexual harassment in the workplace against a background social understanding that "sexual propositions in the workplace are never anything more than a form of 'flirting', and their uneasy rejection by the recipient only ever a matter of her 'lacking a sense of humour'." (Fricker, 2007: 153). Material power (by way of typically being in low-paid, under-valued, and exhausting jobs) and identity prejudice both contributed to systematically exclude women from participating in the processes of meaning creation that would have revealed the wrongness of such acts sooner. The hermeneutical injustice described here is systematic.

One crucial takeaway here is the structural character of hermeneutical injustice, whether incidental or systematic. The structural conditions – the interpretive resources collectively available in a given context – ground individual manifest cases of hermeneutical injustice.

As Anderson notes, Fricker proposes a corrective against instances of hermeneutical injustice that is nevertheless rooted in individual practice. The virtue of hermeneutical justice,

> consists in a disposition to attribute the inarticulate struggles of speakers to make sense of their experiences to hermeneutical injustice rather than to innate epistemic deficiencies… (Anderson, 2012: 167).

This, according to Fricker, will result in the listener being significantly less likely to be affected by identity prejudice in such a way that causes sufferers of hermeneutical injustice to be mistreated when trying to articulate marginalised experiences.

The latter point may well be true, but it does not resolve the issue that has been identified. Paradigmatic cases of hermeneutical injustice occur when a particular society's 'interpretive resources' are in some way deficient, or unjustly distributed, and thus when the experiences of some group of people are incapable of being articulated and/or understood. A disposition to attribute such a lack of understanding to hermeneutical injustice is an improvement over thinking such speakers deficient, hysterical, or deranged, but it is hardly a solution to the core problem; the listeners' failure to understand the reporting of certain marginalised experiences, and the speakers' inability to articulate them clearly. Put simply, knowing that hermeneutical injustice is occurring is insufficient for anyone to *understand* the experience that needs interpreting. That a hermeneutically marginalised speaker is treated as (and perhaps believes themselves to be) inarticulate is merely derivative of the main injustice; that they *cannot understand clearly, nor be understood*.

Knowing that women were attempting to articulate *some* poorly understood dimension of experience when reporting sexual assault many years ago does not by itself reveal to us what it is about the experience that is significant, wrong, or anything else. While such an acknowledgment *may* provoke further investigation that results in a fundamental shift in the interpretive landscape, it will be this shift that resolves or lessens the hermeneutical injustice, not the initial virtuous disposition. Moreover, there is no guarantee that such events will occur. At best then, Fricker's solution is highly partial. It does not sufficiently acknowledge that the structural character of hermeneutical injustice will eventually require a structural resolution; a change in the distribution or quality of a society's interpretive resources. Individual virtues, while surely helpful in prompting efforts to affect such changes, will not do the work by themselves.

I mentioned above that this concept will require a little adaptation for the purposes of my argument. This alteration needs to be made at the point where the distinction is drawn between incidental and systematic hermeneutical injustice. The details of precisely why this amendment needs to be made will become clear as the argument unfolds. For now, however, I can detail what the change will be.

Fricker seems to take any case where hermeneutical injustice occurs independently of the persistent hermeneutical marginalization *of the people who experience the injustice* to be incidental. Yet it seems conceivable that the persistent hermeneutical marginalization of a group who do not themselves suffer from the relevant hermeneutical injustice may nevertheless render the injustice systematic. These are cases where 'Group A' is producing, or simply possess, interpretive resources relevant to the understanding of the experience of 'Group B', but Group A are prevented from contributing these resources to the pool available to persons in Group B (and others). Accommodating these cases is a relatively simple matter of removing the requirement that the hermeneutical marginalization needed for a case of hermeneutical injustice to count as systematic need apply to the same group to whom the injustice is being done.

I should also make it clear that I believe that Fricker's notion of a 'collective' hermeneutical resource is insufficiently expansive to accommodate all cases of hermeneutical injustice. A group need not be marginalized with respect to the collective hermeneutical resources of an *entire community* in order for systematic hermeneutical injustice to occur. It will suffice that they are marginalized with respect to *some* hermeneutical resources that are shared between themselves and a dominant group in a particular context. In such cases, we say that a group is hermeneutically marginalized in a particular context *c*. The general cases that Fricker considers are simply those where the relevant resource is widely shared; where an identity group is hermeneutically marginalized in the same way in all or most social contexts.

Thus, for the purposes of this paper, I shall say that an identity group *g* suffers *systematic hermeneutical injustice* in a context *c* iff,

> a) There is a significant gap in the hermeneutical resources shared between *g* and some other identity group *h* with whom *g* regularly interacts in *c*,
>
> b) This gap prevents members of both *g* and *h* from understanding some aspect of *g*'s members' experience,
>
> c) This lack of understanding significantly disadvantages members of *g*, and

d) This gap occurs and persists due to the hermeneutical marginalization in *c* of some group *i*, whether or not *i* = *g*.

6.2.2. Contributory Injustice

Kristie Dotson (2012) highlights a further limitation in the extant concept of hermeneutical injustice. She argues that it presupposes that the hermeneutical resources at issue are collective or shared (as even my revised notion does). Consequently, theorising around hermeneutical injustice presupposes that gaps in interpretive resources will be shared by the epistemically privileged and the epistemically marginalised alike. That is, the notion of hermeneutical injustice does not clearly allow cases where the marginalised and dominant groups make use of resources that vary in their capacity to make sense of marginalised experiences.[47]

Dotson notes that, in particular, it is often unrealistic to assume that marginalised groups will remain just as unable to interpret their experiences as the dominant group are (Dotson, 2012: 32). More precisely, there will, she posits, be large numbers of cases where the marginalised group have access to interpretive resources able to make sense of their experiences, but the dominant group will not share in them. Thus, the marginalised may quite effortlessly understand and describe some experience of theirs, but be unable to communicate this understanding to the dominant group. This may result in any number of harms. With Dotson, I shall call the injustice in this case *contributory injustice*.

To illustrate, imagine a case where a rape survivor has gone to speak with the police about her assault (the example is due to Ishani Maitra (2010)). She is physically and mentally agitated, finds it difficult to form coherent sentences, and struggles to maintain eye contact with the person she is speaking to. She perfectly well understands this behaviour; it is a result of her trauma. But to the police officers, the behaviour is odd and somewhat inexplicable. They may eventually conclude that the behaviour indicates that she is lying, or at least deeply confused and mistaken, though this is not strictly necessary for the contributory injustice to have occurred. In

[47] That's not to say that Fricker doesn't describe cases where the hermeneutical gap is not a shared one, but simply that the way she characterises hermeneutical injustice obscures this dimension of the relevant cases.

this case, the rape survivor experiences contributory injustice because her experiences and behaviours are presently unintelligible to her interlocutors, but not to herself or others in the relevant community. Her experiences, while perfectly within the reach of her interpretive resources, are beyond those of the police officers.

Notably, in the case of contributory injustice, it seems *indispensable* to understanding the phenomenon that the group suffering it be hermeneutically marginalized. If they were not, their ability to contribute to the shared hermeneutical resource would presumably cause the relevant interpretive resources to become part of a shared pool. So cases of persistent contributory injustice will only be those where the unjustly treated group are unable to bring the benefits and tools of their understanding to the attention of the epistemically privileged.

Elizabeth Anderson's discussion of structural epistemic injustices is importantly distinct from Dotson's point here, although it seems to overlap in certain interesting ways (2012). Firstly, Anderson discusses differential access to socially acknowledged markers of credibility. These markers may, in themselves, be reasonable heuristics for evaluating a speaker's credibility (at least in certain contexts). For example, certain markers of having received a good education may, Anderson thinks, be reasonably used to evaluate a speaker's credibility on matters requiring such an education. These include such qualities as using standardised grammar[48] (2012: 169). Hence, if used appropriately, deploying these markers to evaluate a speaker's credibility is not in itself unjust. It is a reasonable response to the hard task of making reliable credibility judgments in a fast-moving, complex environment. Since, however, certain disadvantaged groups in our society are denied equal access to a quality education, they are also denied access to certain credibility markers. So, while no agent is individually responsible or expressing prejudice when they use these markers to make credibility judgments, the judgments themselves will systematically epistemically marginalise certain people. In this case, differential access to markers of

[48] For what it's worth, I'm not convinced this example is a good one. Standardised grammar is a pretty poor indicator of having received a good education that would seem to lead to injustice more often than not. But my disagreement with a specific case should not be allowed to obscure the central point Anderson wants to make.

epistemic credibility transfers injustice from one structural context (educational disenfranchisement) to another.

This is plausibly why the police officers in the above example may easily be imagined to conclude that the rape survivor is lying. Her behaviour lacks certain markers associated with credible testimony, for instance maintenance of eye contact, clear and articulate sentences, a general sense of calm and so on, that are in themselves (perhaps) typically reliable and justifiable indicators of credibility. Those giving insincere, false testimony (i.e liars) perhaps *really do* have a disproportionate tendency to appear agitated and confused in ways similar to those exhibited by the rape survivor. Thus, while we may entirely fairly lay blame at the feet of the police officers in this case (since we would expect them to be suitably informed about the likely behaviour of rape survivors), the problem goes deeper than the actions of the individual officers. The rape survivor, as someone who has suffered severe trauma, lacks access to certain markers of credibility, and her impoverished access to these credibility markers transfers the injustice of the actions, events and social structures responsible for her trauma into the epistemic domain. This transmitted epistemic injustice, of course, feeds back into the original domain of rape culture, as it makes it significantly harder (indeed, all too often impossible) for the rape survivor to obtain any kind of recompense for her attacker's actions.

Dotson sensibly cautions against treating this structural dimension of contributory injustice as providing much absolution for the police officers in this situation though (2012: 39). For one thing, their epistemic situation is not one in which they had no way of obtaining the necessary interpretive resources. Given that rape survivors share information regarding the effects of their traumas in arenas that are not typically hard to seek out, a lack of crossover in the relevant interpretive resources between survivor and police is plausibly a result of culpable negligence, or even active practices of ignorance (c.f. Mills 1997; Medina 2012).

Moreover, the general reliability of a credibility marker is not obviously enough to ensure that its particular *use* will be just. Whether it is reasonable to judge a person's credibility on the basis of some particular feature also depends on the context being *apt*. In the case at hand, it is not at all obvious that the context of a rape report is one

in which otherwise operative standards of acceptable credibility markers are apt. They are certainly likely to systematically mislead.

The lesson here is that contributory injustice may exist at a point between more agential forms of epistemic injustice, such as testimonial injustice, and more structural kinds. Even when operating in a difficult epistemic context, dominant groups may be sufficiently agentially involved in certain instances of contributory injustice for those instances to be jointly grounded in deficient interpretive resources (structural) and agents' actions and omissions (individual).

In summary, for the purposes of this paper, I shall say that a group *g* suffers *contributory injustice* in a context *c* iff,

> a) There is a significant gap in the hermeneutical resources possessed by a dominant group *h* with whom *g* regularly interacts in *c*,
>
> b) This gap prevents members of *h* (but not *g*) from understanding some aspect of *g*'s members' experience,
>
> c) This lack of understanding significantly disadvantages members of *g*, and
>
> d) This gap occurs and persists due to the hermeneutical marginalization of *g* in *c*.

6.2.3. Testimonial Smothering

So far, we have focused on cases where the person to whom the epistemic injustice has been done typically does, or can, at least make an *effort* to be understood. It is important, however, to consider cases where a person is somehow inhibited or discouraged from providing their testimony in the first place.

Dotson (2011) identifies a pair of phenomena that she collectively refers to as 'epistemic violence'. The most important contribution for our purposes is her notion of testimonial smothering. Testimonial smothering, for Dotson, occurs when, for some potential piece of testimony, *x*, and a speaker, *s*, a) *x* is risky for *s*, b) the potential speaker's audience demonstrates testimonial incompetence with respect to *x*, c) this testimonial incompetence emerges from pernicious ignorance on the audience's part, and d) the potential speaker truncates their testimony to exclude *x* as a result of a, b, and c.

Let's unpack these elements. A piece of testimony is risky, for Dotson, just in case there is a high chance that the testimony will be less than fully intelligible to a speaker's audience (it is *unsafe*), and in virtue of this carries a significant chance of causing harm should it be uttered. For example, since reports of domestic violence perpetrated by black men carries a significant chance of perpetuating a stereotype of black men being particularly violent and abusive, and is likely to be understood in this light by certain (particularly white) audiences, it counts as risky testimony (particularly when delivered to a predominantly or entirely white audience).

An audience demonstrates testimonial incompetence just in case they fail to demonstrate that 1) they find some piece of testimony fully comprehensible, or alternatively that 2) they would be able to detect, though not necessarily correct, inaccuracies in their own comprehension should the need arise. This second element is crucial, as it makes clear that testimonial competence does not require that one finds a piece of testimony fully intelligible, as long as one is potentially aware of one's own ignorance. For instance, even though I do not find advanced talk about mathematical logic wholly intelligible, I do find it comprehensible, and am sensitive to indicators that I have misunderstood certain claims about it. As a result, I am able (though may fail in some cases) to demonstrate testimonial competence with respect to the topic of mathematical logic.

Imagine, though, that I had never studied any formal logic at all, but somehow developed tremendous overconfidence in my ability to understand advanced mathematical logic. I simply believe that whatever I think about any given element of the discipline is correct, and am insensitive to indicators that I am wrong. In this case, I would likely display testimonial incompetence with respect to the topic of mathematical logic. Unless, however, my testimonial incompetence seemed to emerge from an ignorance that was otherwise harmful in that context, condition c would not be satisfied (nor, in all likelihood, would testimony directed at me on that topic be risky, though it would be *unsafe*). Pernicious ignorance is a reliable ignorance about a topic that is likely to cause harm in some context. Since my imagined total ignorance about mathematical logic would be likely to be more

annoying than harmful, this kind of testimonial incompetence would be very unlikely to lead to testimonial smothering.

Given that conditions a, b, and c are met, and that the speaker truncates their testimony in response to them, then Dotson labels what has occurred testimonial smothering. Given the sorts of conditions at play, it is clear that most, if not all, actual cases of testimonial smothering will derive from various structural inequalities, as well as systematic prejudices and biases. Pernicious ignorance and testimonial risk are most commonly outcomes of exactly these societal level variables. Moreover, they are often outcomes of gaps in interpretive resources characteristic of contributory or hermeneutical injustice.

To summarise again, for the purposes of this paper I shall say that a speaker, *s*, experiences *testimonial smothering* with respect to some piece of testimony, *x* iff,

a) *x* is risky for *s.*

b) the potential speaker's audience demonstrates testimonial incompetence with respect to *x.*

c) this testimonial incompetence emerges from pernicious ignorance on the audience's part, and

d) the potential speaker truncates their testimony to exclude *x* as a result of features a, b, and c obtaining.

With the framework of this section in place, I shall now move on to tracing these three kinds of Epistemic Oppression in contemporary psychiatric contexts. To do this, I shall use the framework to interpret and analyse the testimony of psychiatric service users regarding their experiences of treatment in psychiatric institutions. This will involve identifying and unpacking examples of these three kinds of Epistemic Oppression as they occur in contemporary psychiatric contexts.

6.3. Tracing Epistemic Oppression in Psychiatry

Over the last 15 years or so, there has been a significant amount of research done regarding psychiatric service user satisfaction and experiences (Tait & Lester, 2005). By focusing on service users' reports of their experience offered in this body of literature, I will aim in this section to identify plausible cases of the three forms of

epistemic oppression I have identified above. I will also supplement this information with that drawn from the concerns and platforms of service user advocacy organisations.

6.3.1. Hermeneutical Injustice in service user experience

Crichton, Carel and Kidd note that clinicians often feel pressure from their colleagues in other parts of the medical profession (and indeed within Psychiatry) to adopt a biologically reductionist perspective on mental illness (2016: 3-4). This plausibly has a number of potentially deleterious effects on the treatment of patients, both personal and medical (Crichton et al 2016; Bracken et al 2012), but amongst them are insufficient clinical respect for the social determinants and context of mental illness (Kirmayer & Crafa, 2014; Priebe et al, 2013), and the treatment of patients as objects of clinical assessments rather than participants in the process of diagnosis and treatment (Crichton et al 2016: 3)

There is a wealth of evidence that many different kinds of social inequality, division, and isolation contribute to the production, maintenance, and deepening of psychological distress characteristic of mental health conditions such as depression, anxiety, schizophrenia, and others. These effects track inequality and distress across the fault lines of income (Drentea & Reynolds 2012; Dooley et al 2000), race (Tolmac & Hodes 2004; Biafora 1995), gender (Rogers & Pilgrim 2014), sexuality (Dorais 2004), and dysfunctional social/familial relationships more generally, amongst others.

Despite this, many clinical and academic mental health professionals continue, with some notable exceptions[49], to advocate for both a continuation and expansion of a technical perspective on mental illnesses, which decontextualizes mental distress and locates their causes (including faulty mechanisms and processes) within the individual, usually at the level of their neurology, though occasionally abstracting to their individual cognitions and feelings (Bracken et al 2012).

For the majority of people seen by psychiatric services in the UK, the primary sources of information about how they should think about and interpret their

[49] See professional groups such as Psychologists Against Austerity and the Midlands Psychology Group.

experience of mental distress are the medical professionals whom they come into contact with. And as we have seen, these individuals have a strong tendency to downplay, or flatly ignore, the social determinants of mental illness in the explanations they give.

This should be expected to trickle down to many (though not all) service users. Some service users may welcome such explanations (Crichton et al 2016: 4). In particular, those whose problems really *do* have little to do with their environment and social relationships may well find such explanations satisfactory, and may be equipped with sufficient interpretive resources to make sense of their experience. But those for whom the surrounding environment *really is* a major determinant and sustaining factor in their distress may find such explanations unsatisfactory, as they will render certain dimensions (the environmental ones) of their experience unintelligible, or at least less intelligible. They may have a harder time making sense of how their condition came about, why it is chronic rather than merely acute, why pharmacological interventions are providing limited (or no) relief from symptoms, why the cognitive re-evaluation exercises recommended to them by their therapist are proving so difficult to rehearse, and so forth. All of these limits in understanding may be shared by the medical professionals responsible for their care, and will typically prove disadvantageous to the service user.

For instance, especially when combined with the negative dispositions towards oneself characteristic of conditions like depression, anxiety, and others, a lack of proper explanation for various features of one's pathology may lead an individual to believe themselves intractably broken, faulty, without any chance of recovery, or even responsible for their own condition. This in turn may easily lead to the worsening of distress. Alternatively, though relatedly, the inability to appreciate the environmental dimensions of their condition may impede service users from taking certain actions to improve their situation, specifically those which involve taking

steps to alter one's social position[50]. This means that a lack of the correct interpretive resources may directly impede some people's recovery.

Despite a close resemblance, this does not obviously count as a standard instance of systematic hermeneutical injustice, as Fricker (2007) defines it. The reason for this is that it is not clear whether allowing service users an equal role in meaning-making around their own experiences (i.e. eliminating their hermeneutical marginalization) would resolve the issue. The social causation and maintenance of distress are not necessarily apparent to service users even under conditions of epistemic democracy, though undoubtedly more equitable epistemic relationships between service users and their clinicians would go *some way* towards resolving the problem. That is, it seems *prima facie* unlikely that the hermeneutical marginalization of service users is the sole or even primary way in which the interpretive gap that we have identified is sustained. Nevertheless, it seems in all other respects to be a *paradigm* case; both a disadvantaged and dominant group are deprived of particular interpretative resources that are required to make full sense of the disadvantaged groups experiences, to the serious detriment of that group.

This is the heart of the motivation for the alteration to Fricker's model of hermeneutical injustice that I suggested at the end of section 2.1. I suggest here that the hermeneutical marginalization of *some* group is largely responsible for the interpretive gap in Psychiatry that we have identified, but that this is not strictly (or merely) the service users themselves. Rather, it is the hermeneutical marginalization of those researchers, clinicians, social workers, and others who are sceptical of a narrow, biologically reductionist model of psychiatric illness that is largely responsible for the interpretive and explanatory gap faced by service users and clinicians alike.

Two illustrations of this will suffice. The first I have already mentioned; pressure to integrate with the rest of medicine, as well as a wider commitment to biological

[50] This is not to say that everyone who finds themselves in social circumstances deleterious to their mental health would have the ability to lessen their distress simply in virtue of being aware of this fact, though some might.

reductionism in the contemporary psychiatric landscape, will tend to propagate a narrow biomedical model of mental distress amongst clinicians (as pointed out by Crichton et al 2016: 4). Secondly, though relatedly, the reaction to the contemporary concern about diagnostic validity (and other crises) in Psychiatry has generally been to double-down on the conception of the discipline as a form of applied cognitive neuroscience, rather than to broaden its investigative remit beyond individual brains (e.g. Bullmore, Fletcher & Jones 2009; Insel & Quiron 2005, see Bracken et al 2012 for a critique of this position). This doubling down is reflected even in otherwise revisionary research programmes. The Research Domain Criteria framework (RDoC), though critical of contemporary Psychiatry in various respects, suggests that the solution to the identified problems is to focus on cross-diagnostic symptomatic constructs, rather than pre-existing diagnostic categories (e.g anhedonia, rather than Major Depressive Disorder) at a number of analytical levels, from behaviour and self-report, to specific neural circuits, all the way down to molecular and genetic bases of these same circuits (Kirmayer & Crafa 2014; Cuthbert & Insel 2013). Any consideration of social context however, is strikingly absent, except very narrowly with respect to the neural bases of social cognitive abilities.

These two considerations illustrate that socially sensitive perspectives of both practicing clinicians and social workers, as well as research psychiatrists and epidemiologists, are systematically (though perhaps not entirely) excluded from mainstream and revisionary approaches to psychiatric research and practice. This should strike us as a case of hermeneutical marginalization, though not one that impedes the understanding of these practitioners' and researchers' experiences. Rather, it is the experiences of *service users* that are rendered unintelligible.

Since the case that we have identified is naturally thought about as a case of systematic hermeneutical injustice done to service users, but is not (wholly) based on the marginalization of that same group, we have reason, I think, to drop Fricker's requirement on hermeneutical injustice. In any case, the situation is one of pervasive, and harmful, marginalization and injustice of a hermeneutical stripe.

6.3.2. Contributory Injustice in service user experience

Next, I will consider a plausible case of contributory injustice in mental healthcare. This is to be found in the reports of service users treated for auditory hallucinations – specifically those characterised by hearing voices. I will focus on this case here, though much of what I point out is reflected in other service user experiences as well, including those of individuals diagnosed with Borderline Personality Disorder (Rogers & Dunne 2011; Horn, Johnstone & Brooke 2007)

It is a common experience of psychiatric service users, particularly those who spend time voluntarily or involuntarily as an in-patient on a residential ward, to feel that they are understood by staff solely through the lens of their medical diagnosis. In general, moreover, users are critical of the tendency of staff to behave towards them, in all respects, in a manner dictated by a medical understanding of mental distress. This framework employed by staff is often felt to be rigid, uncompromising, and unable to accommodate or 'speak to' users' experiences. This is particularly true of service users diagnosed with psychotic disorders. For instance, as two service users report,

> When I talk to my psychiatrist it's just purely a medical model … but there is so much more to psychosis than just chemicals. (Oakland & Berry, 2015: 125)
>
> I talked to the doctors, and they say things like 'Do you hear voices' and 'Do you think people can take thoughts in and out of your mind' and the sort of standard questions – and they are very rigid about it and they try to fit you in the framework of questions, the standard structure and they won't listen to anything outside of that – they are more interested in trying to diagnose you… (Lovell, 1995: 147)

Participants in the above studies were frustrated at a perceived rigidity in their clinicians' perspectives on voice-hearing, which emphasised neurological explanations and interventions, as well as diagnostic categorisation. This focus was judged to be inflexible and unsuited to properly understanding service users' experiences of voice-hearing.

This is not due to a lack of alternative ways of conceptualising psychosis developed by those who experience it. The activity of the Hearing Voices Network (HVN) is a

case in point. HVN is a loosely connected collection of support groups that aim to provide environments that are,

> …accepting [of] all explanations of voice-hearing… [which] encourage people to share coping strategies in order to increase a sense of control over the experience of hearing voices. (Oakland & Berry, 2015: 119)

HVN advocates an explanatory pluralism regarding voice-hearing. They note that a very wide range of explanations may be offered for such phenomena, encourage dialogue and discussion between service users (and sometimes service providers) who advocate different positions, and are accepting of *any* such explanation, provided its advocate remains respectful in any disagreement. Explanations offered in HVN groups may be spiritual, biochemical, paranormal, or cognitive/affective in nature, and may also make reference to notions of dissociation, trauma, or physical health problems (HVN 2018). The overarching goal is to enhance the wellbeing of those who live with voices and visions (Oakland & Berry 2015).

Service users who participate in the activities of HVN report an acceptance of in-depth, and non-medicalised explorations of hearing voices (2015: 124). Moreover, they report "[i]ncreased empowerment, activity and control" as a consequence of, in part, "[t]he ability to share personal theories of voice-hearing..." (2015: 127), as well as the peer group being "a safe environment where…their expertise in managing their experiences is acknowledged" (2015: 126). For many service users, participating in the group represented

> …the first time their experiences had been believed after significant periods of isolation in their experiences, *including dismissal of experiences by health care workers*. (2015: 126)

The novelty of the experiences within the support group, as well as the significant improvements in participating individuals' states of mind, suggests that the openness and acceptance of service user perspectives, as well as the concomitant benefits, are not reliably present within formal psychiatric services. This is to service users' *significant disadvantage*. Perhaps more importantly, the difference here does not seem solely attributable to testimonial injustice (or a lack of it). While service users'

at HVN did report a lack of 'belief' in their experiences within formal psychiatric services, it is not sensible to assume that other participants in HVN groups necessarily believe all the claims that others make about voice-hearing. After all, at least *some* of these claims may be entirely incompatible with those that others' make. Rather, the benefits HVN groups are able to produce together are better thought of as improvements in frameworks for understanding and coping with an experience that these users have in common. The medical lens of formal psychiatric services is objectionable to many participants, because it does not 'speak to' their experiences in full – and it is sensible to presume that medical professionals are not deliberately excluding interpretive frameworks known to be useful.

This is a case where service users have collaboratively overcome a gap in understanding present within the dominant medical model of voice-hearing, but these insights have not come to be reflected in more formal psychiatric contexts. The significant feature of HVN groups is that individuals' views are *not summarily dismissed* for failing to accord with some pre-existing theory. Participants are treated as equal participants in a discussion that encompasses the nature of voice-hearing, as well as strategies for managing the experiences and associated distress.

Thus, when certain service users come into contact with mental health services, clinicians may lack understanding of their experience without the service user sharing in this lack of understanding. Such a case will amount to contributory injustice.

In line with Dotson's assessment, this is neither to fully condemn nor entirely absolve medical professionals (or, perhaps, medical institutions as a whole) of individual responsibility for the injustice done to those with experiences of hearing voices by ignoring the validity and utility of their interpretive resources. While the majority of medical professionals who actually propagate this injustice[51] are almost certainly ignorant of the alternative frameworks available, it might be thought that this ignorance is *wilful* in the sense that it emerges from a failure of many medical

[51] Naturally, not all do. Many actually participate in, contribute to, and utilise the resources of HVN and similar service-user led organisations. This is, of course, to be strongly encouraged.

professionals to engage with voice-hearers as equal epistemic agents (Pohlhaus, 2012). While they treat various peers within the medical establishment as people who can be relied upon for information and insight, service user experiences of health services (and the contrasting experiences of organisations like HVN) suggest that many do not make an effort to engage with service users in the same fashion. Thus, although the gap in understanding is partially grounded in how the medical professionals are socially situated (both in terms of their job, and, typically, lack of experience of psychosis), it is also partially grounded in an unwillingness to enter a properly interdependent epistemic relationship with service users (Pohlhaus, 2012: 720-723). That is, these clinicians (and others),

> …regard their patients as *objects* of their epistemic enquiry rather than *participants* in an epistemic search for the correct diagnosis and best treatment. (Crichton et al, 2016: 3, *emphasis mine*)

In short, while the injustice-creating limitations of many medical professionals may not be initially their responsibility, it does often reflect a wilful disengagement with certain perspectives that medical professionals are (in part at least) personally responsible for.

6.3.3. Testimonial Smothering in service user experience

Finally, I turn now to identifying plausible cases of testimonial smothering in psychiatric contexts. This is, in a sense, a difficult task, since, by stipulation, testimonial smothering is characterised by a lack of testimony that would be there were it not for the relevant unjust circumstances obtaining. To provide evidence that a case of testimonial smothering has occurred would require either an individual reporting exactly the style of thought and judgment characteristic of it, or, more indirectly, establishing that the kind of external circumstances liable to produce testimonial smothering obtain, and then establishing that some people in such circumstances truncate their testimony. It is this latter strategy that I will pursue here.

The first crucial observation is that, quite generally, when an individual presents to primary care services in a state of acute mental distress, they are exposed to at least a

couple of potential risks. One is obvious; if they do not receive any help, then they are liable to experience extreme distress, perhaps engage in acts of self-injury, or even attempt or complete suicide. This risk is not simply a function of being mentally unwell in the first place; refusal of services (or feeling that clinicians judge them as lacking sincerity or seriousness) can add to service users' common feelings of abandonment or not being deserving of help (Davidson et al 2005). That is, a poor response from psychiatric services may significantly *add* to service users' distress, *especially* since it may already have been difficult for the person to reach out for help.

The second is potentially obvious to people who have, directly or indirectly, experienced the process of committal to an in-patient psychiatric unit. People may go onto such a unit voluntarily or involuntarily (colloquially 'under section'). The details of involuntary committal are not crucial here. What is crucial is that involuntary committal can involve the removal of various legal rights – not just liberty, but also the right to refuse medical treatment. And though there are some safeguards against misuse in place – though service users often report difficulty accessing an advocate to help them enforce correct procedure (CQC 2016) – we can safely take it that involuntary committal is, *ceteris paribus*, a bad thing (naturally, it is designed to be used when all else is *not equal* – that is, when somebody is a significant risk to themselves or others).

The upshot of all of this is that a service user's testimony in a condition where they are reporting to primary psychiatric service experiencing extreme distress has the potential to cause various harms should it be misinterpreted. If a person tries to secure a certain course of treatment for themselves, and they are turned away, then harm may emerge from that refusal. If a person tries to report certain severe symptoms (especially if they have particular aversion to going onto an in-patient ward) then they risk being involuntarily committed, and being deprived of certain rights. Worse, there is significant evidence that even some of those who are 'voluntarily' committed to in-patient units are held against their will, by way of threats of being sectioned should they try to discharge themselves, or simply by way of locked doors – so-called *de facto detention* (Lovell 1995; Gilburt et al 2008; CQC 2014; 2016). Further, a lack of discussion with service users about what medications

or treatments they will receive can amount, effectively, to a lack of informed consent regarding treatment (and hence, a form of coerced treatment). So going onto a ward at all holds the potential for significant harm.

Can a case be made that testimony regarding symptoms made in such a context can be risky in Dotson's sense? This requires that there is a significant chance of the potential harms of the testimony being realised *in virtue of it being less than fully intelligible* to one's audience.

Certainly one can see how such situations could arise. Since people with certain conditions, such as anorexia nervosa, schizophrenia, and bipolar disorder are often thought to display anosognosia[52] (Wu et al 2013; Ghaemi et al 1996; Amador et al 1991), an expression of a genuine, knowledgeable preference not to be admitted to an in-patient unit may instead to be taken as further evidence of a person's incapacity to make sound judgment. Indeed, this may occur even when anosognosia *per se* is not posited; statement of preference not to go onto an in-patient ward may instead be interpreted as a refusal of treatment, or intent to do harm to oneself. Crichton and colleagues actually report such a case as an instance of testimonial injustice. The important aspects of the case are as follows.

> … a young man…was admitted to psychiatric hospital on section 2 despite the fact that he had agreed all along to be admitted and remain in hospital as a voluntary patient. He had been standing near the edge of a high cliff for about an hour until passers-by called the police. The staff involved in his care on admission did not believe that he could be trusted to remain in hospital on a voluntary basis and argued in the tribunal for the maintenance of the section. His community psychiatric nurse attended the tribunal, stating that he should never have been placed on a section, because he had had suicidal thoughts for many years, had gone to the same cliff many times in the past, had been admitted to hospital on several occasions as a voluntary patient, and had misgivings about the stigma attached to being placed on a section.

[52] A symptom whereby somebody falsely believes that they are well when they are in fact unwell, or otherwise lacks insight into the extent of their ill-health.

> All this had been documented in the hospital notes. She conceded that there would always be a risk of self-harm, but that it was a matter of managing the risk without compulsory detention and with the help of his friends and family. (2016: 3)

While this surely *is* a case of testimonial injustice, in the sense that undue lack of credibility was given to the man regarding his intent to stay in hospital voluntarily (see Fricker 2007), it also seems to involve a misunderstanding of his testimony. His testimony is less than fully intelligible to those medical professionals responsible for his care, because they see only the potential for him to access a situation where he will be able to complete suicide. Unaware of his background, they did not interpret his testimony as an expression of genuine concern regarding the stigma of involuntary detention.

A different example involves people diagnosed with borderline personality disorder. People with personality disorder in psychiatric services often report negative responses from staff. Typically these involve attitudes of dismissal, especially in relation to the greater difficulties perceived to be experienced by other patients. For instance, one patient reports,

> Yeah the attitudes can be quite difficult because they can't place you, it's not like I'm a schizophrenic or you've got this very definite problem or perhaps you're just there to be dried out or detoxed or whatever, erm and I think that they think that you're being difficult most of the time, I know they think 'Oh god she's playing up'. (Fallon 2003: 397)

Relatedly, service users diagnosed with personality disorder have reported being "made to feel undeserving of in-patient care." (Fallon 2003: 397). The relevant reactions from staff are responses to particular, often distressed, reports of symptomatology from patients ('playing up'). Nancy Potter notes that real-time expressions of *warranted* anger expressed by BPD patients towards clinicians and other staff are liable to be written off as merely symptomatic of their condition, or of reflecting some past trauma (2009: 40-43), and that communicative intent present in certain acts of self-mutilation may similarly be missed or ignored (2009: 97-98). Thus,

such behaviour is another example of how some service user testimony seems, for a variety of reasons, to be less than fully intelligible to some clinical staff. In part, this is due to a lack of recognition of the legitimacy of their disorder. It can also be due to the extremely emotional nature of the testimony's expression, or the belief (explicit or otherwise) that pathology infects all emotional expression of a BPD patient. And since being made to feel undeserving of care or otherwise dismissed can, naturally, increase someone's experience of distress, this lack of full intelligibility is likely to cause harm.

The upshot here is that if these kinds of cases are *able* to occur, then there is a significant chance of harm being caused to a patient in virtue of their testimony being less than fully intelligible to those professionals responsible for their care. Such testimony is therefore risky, in Dotson's sense.

Moreover, if reactions such as those described above are experienced when the service user does *not* truncate their testimony, then medical professionals clearly are displaying relevant testimonial incompetence on certain topics as well. They are certainly not consistently demonstrating an ability to detect inaccuracies in their own understanding, nor to find the testifier's communicative intent fully comprehensible. Other factors may also contribute to service users' doubting that medical professionals will find testimony about their experiences fully comprehensible.

For instance, an over-reliance on a decontextualized biomedical model of mental illness that does not encourage attention to details of a service user's life relevant to understanding their relationship to their distress (Bracken et al 2012) may reduce service user confidence that their reports will be properly understood. For example, one study participant reports the following about interacting with doctors as somebody who experiences psychosis,

> When I talk to my psychiatrist it's just purely a medical model... but there is so much more to psychosis than just chemicals. (Oakland & Berry 2015: 125)

This report seems to suggest a frustration born of a restrictive view of psychosis held by those responsible for this service user's care. It would be natural to suspect that a service user in this sort of position would be doubtful of their psychiatrist's capacity

to understand reports that don't obviously align with the narrow and inflexible view of psychosis that is usually the topic of conversation.

Moreover, some professionals' propensity to ignore pleas from service users for emotional support[53] (Fallon 2003: 397), may also contribute to service users feeling that medical professionals will not find testimony about their experiences fully comprehensible.

Regardless of the reason, the fact that service users regularly report that they want to be *listened to* by medical professionals (CQC 2016; Tait & Lester 2005; Fallon 2003) would suggest that when service users do report their experiences, responses from staff indicate (to the service user at least) a lack of appreciation or understanding. This counts as evidence that displays of testimonial incompetence may be a regular, perhaps common, phenomenon within psychiatric services. And, as we have seen, it seems likely (assuming that medical professionals are not deliberately, regularly, and systematically mistreating service users) that this incompetence is born of ignorance. Since it is likely to cause harm in the relevant context, this ignorance counts as pernicious, in Dotson's sense. Thus we have all the preconditions of testimonial smothering.

A cynic might note that the evidence I draw upon in the preceding paragraphs does not necessarily establish the conclusion that service users' testimony is being systematically misunderstood. It could instead indicate that clinicians simply do not care about, or choose to ignore, the perfectly intelligible testimony of service users. Such a case would not qualify as displaying testimonial incompetence; clinicians would instead be indicating a disposition to not give due consideration to service user testimony (that they nevertheless perfectly well understand), or even that they are entirely indifferent to such testimony. Thus, technically, such a case would not establish the preconditions for testimonial smothering. I grant that such a case seems possible - indeed it looks like a plausible interpretation of all of the kinds of injustices

[53] Naturally there may well be all sorts of time-constraints and professional barriers explaining this – I do not intend to imply that this is solely the fault of individual clinicians.

identified so far. Do we have any way of pulling these two different kinds of cases apart?

I am not at all confident that individual cases can generally be reliably distinguished along this axis. But it seems unlikely that health services are filled with completely callous individuals. Thus it seems justifiable to suppose that *at least* a significant portion of such cases are indicative of testimonial incompetence rather than sheer cruelty (though, understandably, it may not appear that way to service users, nor be much comfort to them). For that reason alone, it seems worth exploring what the structure and moral force of such cases would be, even if we are unable to find any unarguable cases in the service user or other literatures. We have good reason to believe that some (indeed, many) such cases exist, even if each individual case we can identify is not definitive.

Moreover, this kind of callousness seems like it may be able to play a very similar functional role that testimonial incompetence plays in characterising cases of testimonial smothering. In particular, it seems plausible that interactions characterised not by testimonial incompetence but rather by callousness would be able to generate similar testimony-truncating effects in those subject to it. Imagine that instead of being ignorant of a service user's intended meaning, a clinician is simply indifferent to it. Assume that the service user suspects that this is the case, and further that testifying in that way would consequently involve a significant risk (say, of being detained against their will). It would be natural for the patient to thereby refrain from testifying on that topic at all. There is no testimonial incompetence involved in such a case, but we do get truncated testimony. It is unclear whether such a case might also involve ignorance – perhaps a certain kind of moral ignorance would be necessary for a clinician to callously disregard their patient's testimony in this manner in the first place. Regardless of this, such a case looks very much like testimonial smothering. So if the possibility of such cases is raised as an objection to my overall concern with testimonial smothering in this paper, then it seems to fall flat. Perhaps I should simply offer a small amendment to Dotson's account of testimonial smothering such that it encapsulates these kinds of cases as well. This move does not seem unmotivated. Perhaps there are interesting

theoretical distinctions underlying this discussion, but they do not seem relevant to my purposes here.

To put the point another way, if clinicians *were* systematically ignoring perfectly well-understood pleas from their service users, then clearly something morally dubious (at the very least) would be going on. This moral dubiousness would stem in part from the service users' lack of any significant epistemic role in the psychiatric environment. And one of the many ways this would play out would be in service users systematically truncating their testimony to avoid the potentially harmful effects of presenting testimony that would, in all relevant respects, be ignored anyway. My point is that we needn't assume that such vindictive exclusion is going on in order to think that something morally dubious is occurring. This is revealing, and moves the discussion forward, since many would balk (perhaps fairly, perhaps not) at the thought of this degree of callousness being regularly exhibited by medical professionals. The case being presented here proposes that we needn't assume such an unpalatable premise in order to think there is something seriously amiss with the current nature of service user-clinician interactions in psychiatric contexts, and that the issue has a distinctively epistemic dimension.

So, the right preconditions for testimonial smothering (or at least something similar) seem to exist in psychiatric services. Furthermore, it seems highly probable that patients familiar with those risks involved in testifying about their experience will truncate their testimony in response to it. This is because it seems to be a predictable, sometimes even rational, response to the preconditions of testimonial smothering. Is there any first-hand evidence, however, that such truncation actually takes place? Naturally, such evidence can be hard to come by, as it is by definition evidence of an absence, and service users may naturally be wary about admitting that they withheld information from mental health professionals. Moreover, even the mere risk of testimonial smothering and the existence of its preconditions are clearly serious ethical issues worth combating. Nevertheless, I shall end this section by pointing to a case that can be sensibly interpreted as testimonial smothering.

The mental health charity Mind uses a relevant example in their guide to sectioning. In this case, a patient is scared of being sectioned, and so does not report her self-

harming behaviour to a clinician (Mind 2013a: 17). Say, as is distinctly possible, that such a patient's self-harming behaviour is actually a necessary tension-relieving or emotional regulatory strategy[54] (see Zila & Kiselica 2001; Himber 1994), but is likely to be misinterpreted by the clinician as indicating an increased risk of suicidality (the opposite is in fact true – restriction of her coping mechanism will *increase* the patient's risk of developing suicidal intent). Such a situation is, then, plausibly an instance of testimonial smothering. The service user judges that her testimony is risky (it is likely to be less than fully understood, and carries the attendant risks of section and restriction of a coping mechanism), the clinician plausibly demonstrates testimonial incompetence (many clinical reactions to reports of self-harm are to reflexively treat the action as intrinsically negative, or harmful), and this seems to emerge from pernicious ignorance (ignorance of a service user's relationship with self-harming behaviour, that causes the risk of harm in context – in this case the deprivation of a patient's liberty and/or required coping mechanism). Since Mind use real patient experiences when putting together examples for their guides, this is some evidence that testimonial smothering does actually take place in psychiatric contexts.

6.4. The Moral Force of Epistemic Oppression in Psychiatry

In the previous section, I identified cases of three different kinds of epistemic oppression in Psychiatric contexts, by investigating qualitative reports of service user experiences found in the research literature. In this section I want to argue that the existence of these sorts of injustices have a particular moral force. Specifically, I will argue that these kinds of injustice speak to the need to reform psychiatric practice in various ways. While both hermeneutical and contributory injustice speak to the need to place a much wider range of conceptual resources in the common pool shared by mental health professionals and service users, cases of testimonial smothering speak to the need to alter the power differential that exists between them. What this amounts to exactly will become clear in what follows.

[54] There are numerous other critical functions self-harm may perform for different people rather than being simply symptomatic of mental disorder, including helping to gain a sense of control, escaping traumatic memories, or avoiding feelings of numbness or dissociation (Mind 2013b: 4).

These courses of action will be more or less morally mandated, depending on a) how common these forms of epistemic oppression in Psychiatry are, and b) how strong countervailing moral reasons against these courses of action turn out to be. Identifying some of the moral reasons in play, therefore, is only the first step towards arguing for the need for psychiatric reform. I shall attempt to provide moral reasons one might offer against such reforms as I proceed, but will ultimately argue that they do not defeat the primary moral reasons identified. This cannot, of course, firmly establish that no such defeating reasons exist.

6.4.1. The requirements of hermeneutical justice in Psychiatry

Rectifying the hermeneutical injustice I identify in section 3.1 is perhaps the most simple of the processes I will suggest here. This is not to say that in practice it will be easy, but merely to say that incorporating non-dominant perspectives on the cause, basis, and nature of mental illness into clinical practice, where appropriate, is not in itself something that is obviously difficult to do.

The most obvious intervention is to present to medical students, when they first study Psychiatry, a viewpoint that emphasises that, whatever the benefits of the biomedical/technical perspective when understanding somatic illnesses may be, these benefits do not obviously transfer wholesale to understanding psychiatric illness. In particular, the need to appreciate the complex and multi-faceted role of the environment in both the onset and sustenance of mental distress would need to be emphasised, rather than reflexively treating these conditions as diseases more-or-less like any other. This would not require the teacher, or even the students, to accept alternative models as more appropriate, or to present the biomedical/technical model as wrong or fundamentally wrong-headed when it comes to understanding mental distress. Rather, they would need to be presented as significant perspectives that, despite clear discrepancies in research funding allocation (Lewis-Fernández et al, 2016), are at least as well evidenced as the biomedical one.

Moreover, it would be worth emphasising that such alternative models may improve a service user's understanding of their experiences more than a purely technical presentation, and that this in turn leaves the service user able to make *better informed* choices, given the evidence available, than they would otherwise be able to. The

reason why this kind of direct clinical relevance is crucial to emphasise is because there is a risk otherwise that practicing psychiatrists will deem the social determinants of mental illness as beyond their remit or resources to intervene upon, and hence adopt in their practice a perspective that is, in all but name and minute detail, a restrictive biomedical one. This, of course would not resolve the main problem of some service users being unable to fully understand their own experiences. Indeed, this situation would appear to be worse, since on this picture, clinicians would be making deliberate decisions to withhold potentially relevant information from their service users (albeit while *unaware* of its significance).

6.4.2. The requirements of contributory justice in Psychiatry

Newman et al (2015) note that service user-involvement in mental healthcare across the UK, Ireland and mainland Europe is restricted at best. Though it is often formally a requirement for service providers to significantly involve service users in the planning and delivery of their care, in practice this involvement is very limited, and is often perceived by service users to be tokenistic (2015: 180). As we have seen, this means that the interpretive resources that have developed in service user communities to help them reflect upon, understand, and manage their conditions are in general not being picked up on by medical professionals, or otherwise incorporated into care and recovery planning.

Correcting for this kind of contributory injustice is not easy. At root, it will involve clinicians developing the propensity to, in the terms of Potter (2009), 'give uptake' to service users' communications of their perspectives on their illness and goals for recovery, *regardless* of how initially improbable or bizarre such perspectives seem. This is not an easy propensity to develop. Giving uptake, as Potter points out, involves making "an earnest attempt to understand things from the communicator's point of view" (2009: 140). This does not just involve conscientiously acknowledging and then considering a conventional interpretation or meaning of what the speaker is offering – such conventional interpretations of the perspectives of people who hear voices or exhibit signs of personality disorder are, after all, tainted by stereotypes regarding their deep-set irrationality, unintelligibility, and/or pathology (Potter, 2009: 141; Gray, 2002). Thus, giving uptake in these sorts of cases will involve

sincerely trying to understand things from the perspectives of people who are often thought to have an incoherent or fundamentally unintelligible point of view.

That said, giving uptake is also (thankfully) not a matter of agreeing with the perspective being presented, or sincerely judging that all the resources being used to interpret a particular experience are helpful for that purpose (Potter, 2009: 141). Disagreement is compatible with giving uptake, but in giving uptake we must "take seriously the reasons that person gives for her actions and beliefs" (141). In short, the requirements of contributory justice in Psychiatry are that clinicians are *familiar with* and *take seriously* the many different interpretative frameworks through which service users make sense of their experience.

In this context, to take an interpretive framework seriously may involve several different attitudes. One way is to be open to the possibility that a framework provides genuine insight into the nature of a service users' condition, even where medicalised understandings of the condition suggest that this is *prima facie* unlikely. Another is to be open to the possibility that the framework provides some significant benefit for managing or recovering from distress. The unifying feature of these and other acceptable attitudes is that they treat the service user as an equal participant in a joint inquiry that aims at maximising their wellbeing.

Given the nature of the resources developed within the service user community, this will involve familiarity with perspectives that diverge enormously from the overwhelmingly technical (Bracken et al, 2012) or medical (Oakland & Berry, 2015) perspective that psychiatrists are under a number of professional and personal pressures to adopt. At least, there must be genuine openness amongst healthcare providers to the thought that this perspective may miss important dimensions of some peoples' experiences; dimensions that are relevant to understanding their illness and promoting an appropriate recovery (Kogstad et al, 2011).

All that I have said above is not to promote any specific *alternative model*, but rather to promote the authority of service users in co-negotiating with care providers how *their specific* experiences are best understood, reacted to and managed.

Mere abstract commitment to giving uptake is unlikely to suffice on its own for promoting contributory justice, no matter how strong. Seriously considering the reasons given by an individual for beliefs and actions that a clinician is trained to consider pathological will likely require sustained engagement and interaction with the service user communities from which these alternative perspectives arise.

There are two central reasons for this. Firstly, sincere engagement with these communities will help to reduce a clinician's sense of service users being an 'out-group' to their professional 'in-group', thus diminishing ethnocentric biases[55] (Anderson, 2012) that will otherwise tend to ingrain a technical or medical model to the exclusion of other perspectives. Secondly, such engagement provides an opportunity for clinicians to be exposed to the reasons and justifications for a variety of non-medicalised perspectives on mental distress; a potentially powerful corrective to the restricted view they may have been exposed to in training, and throughout their professional lives. HVN already invites the participation of clinicians, on the understanding that they abide by certain rules within the group, which amounts to operating on an equal footing with, and respecting the diverse views of, service users (HVN 2017). This model could be generalised to promote sincere clinical engagement with other service user communities, though great care would need to be taken with this, as many such communities are, for good reasons relating to experiences of violence, abuse, and malpractice, deeply suspicious of the medical profession.

Some may worry that sincere engagement of this kind may sometimes lead to tension with clinicians' other ethical obligations. Suppose that a service user comes to interpret auditory hallucinations as the voices of guardian angels. Further suppose that this interpretation helps them to cope with what was previously a very distressing experience of voice-hearing – hearing the voices is still frustrating in that it impedes hearing external stimuli, can interfere with sleep and thinking, and so on, but this understanding of the voices' origin significantly decreases their day-to-day distress.

[55] Dispositions to unwarrantedly interpret the behaviour of one social group with reference to the norms and values of our own.

If what I have said above is correct, then the requirement of contributory justice here is that the clinician give uptake to the service user's perspective. While this does not involve believing the interpretation to be accurate, giving uptake to this viewpoint would potentially be incompatible with explaining, in a timely fashion, why it is, in all likelihood, false. If the view's presumed falsity counts as medically pertinent information, then failing to offer this information might violate the requirements of informed consent to treatment (Eyal, 2012). So we have a *prima facie* incompatibility (in certain kinds of situation) between the requirements of contributory justice in Psychiatry on the one hand, and the requirements of informed consent on the other.

I do not share the concern that giving uptake to an idea is incompatible with stridently defending one's position that the idea does not represent reality. It seems perfectly possible for a clinician to sincerely engage with this service user's perspective on their voices as angels, and to observe and reflect upon the benefits this perspective provides to the service user, all the while making it clear that they believe it to be false. What would be crucially important here is that the clinician retained an open-mind throughout regarding its value for the service user, and its potential role in a valid recovery plan.

Furthermore, it is not clear to me that a clinician failing to inform a service user about their beliefs regarding angels counts as withholding pertinent information. If the standard for pertinence primarily relates to protecting a service user from potential harms or negative effects of proposed interventions (Eyal, 2012: §2.1) then, unless temporarily or permanently holding a most likely false belief intrinsically counts as a harm, the information in this case is not obviously pertinent. Nor does the information seem to be of the sort that is necessary to ensure that service users' autonomy is adequately promoted and defended (e.g Beauchamp, 2010), nor that abuse is prevented (e.g. Manson & O'Neill, 2007). Indeed, service user organisations typically promote users' own views precisely as *a way of promoting* autonomy, and preventing or correcting for abuse (Manson & O'Neill 2007; Beauchamp 2010). These are amongst the most common rationales for promoting informed consent.

But let us grant for the sake of argument that the two *prima facie* duties here – giving uptake to and not misleading service users are indeed both operative and

incompatible in cases like this. Should such cases lead us to give up the generality of the obligation to promote contributory justice, or the obligation to avoid misleading service users? Here I must admit to not having a definitive answer.

We should note, however, that misleading service users in this case need not involve lying. It may merely involve the clinician refraining from stating something they know to be the case. If one takes the natural view (Saul 2012: 3) that omitting information is in general less morally questionable than lying, then this minimises the severity of the necessary deception. That said, persuasive arguments both specific to the medical domain and quite general have been made that suggest misleading by any means is typically no better or worse than simply lying (Cox & Fritz 2016; Saul 2012).

All that said, I think I have said enough to sum up the issue as follows; the burden is on anybody who is still concerned about the effect of this practice of giving uptake on informed consent in Psychiatry to provide an argument that a) these two duties are both incompatible and relevantly operative in important contexts and that b) the type of misleading that might go on is sufficiently serious to warrant not giving uptake in the manner I have outlined here.

Another worry is that a clinician may not be able to *take seriously* the suggestion that the service user is putting forward to understand their experiences, not in any way due to the status of the service user, but because of the content of the suggestion itself. The existence of guardian angels is presumably incompatible with the broadly physicalist worldview of many physicians. And, just as it is not possible to choose to believe things that one thinks are false, it is plausibly impossible to choose to take seriously things that one thinks are patently absurd. If so, then a large number of physicians are unable to meet the requirements of contributory justice as I have described them, in a significant class of cases. This might be thought to weaken the obligation to seek out contributory justice in the first place[56].

[56] I thank an anonymous reviewer for the Journal of Medical Ethics for this objection.

I have two replies. Firstly, it is not clear to me that this in any way lessens the general obligation to seek out contributory justice. I agree that nobody can be obliged to do the genuinely impossible. But much of the insight that service users have into their own condition *is* compatible with a broadly physicalist worldview, as we saw in the discussion of the range of views put forward in HVN groups. The existence of fringe cases where physicians are incapable of taking a suggestion seriously says nothing about the *general* obligation to seek contributory justice.

Secondly, to take a service user's perspective seriously need not involve being open to the possibility that it accurately captures the nature of some phenomenon. One could be closed off to *this* possibility, and yet open to the possibility that the perspective in question significantly supports the service user in avoiding or managing distress. One need not even *entertain* the possibility that guardian angels exist to think that the most effective therapeutic strategy *might* involve the service user maintaining such a belief. The *prima facie* absurdity of the belief from the clinician's perspective must not infect their evaluation of its potential therapeutic value.

One final worry is that even if it is *possible* for a clinician to take a service user seriously in the sense relevant to ensuring contributory justice, it may not always be ethical for them to do so. I stipulated in the above example that the service user reports that their understanding helps them to manage distress. Implicit was the assumption that their perspective was not incompatible with effective recovery. But what if a service user interprets their experiences in a way that significantly interferes with recovery? Some examples may include a service user who hears voices that tell them not to trust their psychiatrist/therapist or not to take an otherwise effective medication[5758].

[57] I say 'otherwise effective' to exclude cases where a service user is prescribed ineffective or harmful medication. Having motivations to come off such medication does not interfere with recovery.

[58] I thank another anonymous reviewer for the Journal of Medical Ethics for this objection.

Certainly, if a service user's perspective fails to offer genuine insight into the nature of voice-hearing, *and* impedes their wellbeing, a responsible clinician should not act so as to reinforce it. But this concession does not undermine my argument.

Many perspectives that service users have on their conditions that clinicians are currently dismissive of are *not* of this sort at all. Moreover, ascertaining that a service user's perspective is of this sort *necessitates* engaging with them as an equal participant in a discussion about their recovery, and thus taking the perspective seriously (at least at first). This is because contributory injustice impedes the clinician from fully understanding the role a service user's perspective plays in their understanding and management of their condition. Contributory justice does not require of a clinician that they reinforce *genuinely* harmful views a service user may have regarding their own condition. Rather, a clinician must combat contributory injustice in order to know that a service user's understanding of their condition is genuinely harmful.

6.4.3. The requirements of testimonial autonomy in Psychiatry

I refer to the inverse of testimonial smothering as testimonial autonomy. That is, an individual displays testimonial autonomy when they provide important testimony to an audience in the absence of the pre-conditions of testimonial smothering. I take it that testimonial autonomy is at least as important in mental healthcare as it is in any other context; serious harms are closely associated with the silencing of service users, including impoverished clinical knowledge and worse health outcomes (Simpson & House 2002), in addition to the intrinsic injustice of silencing. Therefore it is critical that psychiatric services are structured so as to enable, or at least so as not to preclude, testimonial autonomy.

I take it that working towards testimonial autonomy entails not merely avoiding cases where people *actually* truncate testimony in response to risk, testimonial incompetence, and pernicious ignorance, but actively working to ensure that these circumstances do not obtain, or obtain as rarely as possible. Merely encouraging people to testify in epistemically hostile circumstances where it is potentially dangerous for them to do so is not a desirable resolution.

I shall only briefly discuss ways we might go about reducing testimonial incompetence and pernicious ignorance here. One method would be to cultivate (by whatever means) a propensity in clinicians to give uptake (Potter 2009) to service user perspectives (for a similar suggestion in a different context, see Miller Tate 2018). Another would be to expand the range of data points and theoretical perspectives on the origin and nature of mental distress that clinicians take seriously. Both of these moves should significantly reduce the kinds of ignorance that typically cause harm in clinical contexts. This is simply because many (perhaps most) of the relevant kinds of ignorance amount to a lack of knowledge either of service users' relationship with their distress and symptoms (including behaviours like self-harm), or the limits of the biomedical model in Psychiatry. Nevertheless, since the notion of pernicious ignorance is so broad, it is unlikely that we are currently in a position to eliminate it entirely (even in principle). There are simply too many different ways one can be ignorant that are liable to cause harm when interacting with a population as vulnerable as those who come into contact with mental health services. That said, we may be able to limit the harm some of this ignorance is able to do in these contexts, a point I shall return to below.

Similar remarks apply to reducing testimonial incompetence. Clinicians will certainly be less prone to failing to demonstrate the ability to understand certain kinds of testimony if they are better positioned to give uptake to service user testimony. Giving serious consideration to descriptions, accounts of, and reasons for particular behaviours or experiences from the perspective of the service user will typically (though not always) indicate a certain ability and disposition to, at the very least, detect and correct for errors in their comprehension. The same goes for expanding the range of understandings of mental illness that clinicians are disposed to draw on in general, beyond the narrow technical model. Once again, however, there are simply too many topics about which clinicians might be ignorant (or otherwise fail to demonstrate their competence with) to be confident that these interventions are, even in principle, capable of eliminating displays of testimonial incompetence entirely.

What about the final precondition of testimonial smothering, then? What steps can we take to make service users' testimony in psychiatric contexts less risky? Degree of safety (the chance that some piece of testimony is likely to be less than fully understood in context) seems likely to be responsive to the interventions discussed above (but also tricky to ensure). Nevertheless, we can minimise the harmful effects of service user testimony in clinical contexts, even if it is liable to be misunderstood.

In section 6.3.3, I gave examples of two kinds of harm that might befall a service user when communicating unsafe pieces of testimony in psychiatric contexts. The first involved being forcibly detained (sectioned) against one's will, and perhaps to one's overall detriment. The second involved something like the converse; being refused care one needs, or being made to feel undeserving of the care one is already receiving. I shall offer suggestions for how to mitigate each of these in turn.

In the first case, there are a number of possible ways the potential harm could be reduced. Firstly, to the extent that the harm of admission to a psychiatric ward (voluntarily or involuntarily) is due to its being stigmatising (Loch 2012; McCarthy 1995), reducing the associated stigma will reduce the associated harm done. This is especially true if reducing the stigma of hospital admission encourages more people to consent to it, since any amount of compulsion is (all else being equal) harmful.

Secondly, and perhaps more pointedly, we could lessen the material harm associated with being sectioned. In the UK, for instance, the most common forms of sectioning permit both deprivation of liberty and the use of force to ensure immediate safety of the service user and others, as well as coercive treatment of the presumed underlying disorder (Katsakou et al 2010). Many service users report coercive treatment as one of the worst aspects of being under section (Tait & Lester 2005; Lovell 1995). Hence, ensuring that permissions for these two practices were kept strictly apart in both legislation and practice (i.e., that obtaining a section did not also typically entail a right to administer treatment coercively) would significantly lower the potential harm of section for many service users.

Thirdly, we could reduce the power of clinicians to obtain a section in the first place, or to enforce one de facto when legal requirements have not been met. This might

involve empowering already statutorily designated limitations on clinicians (including giving social workers the resources to confidently disagree with clinicians' assessments when appropriate), or through altering the law to include more safeguards. While these changes would be controversial (I do not claim otherwise – simply because they would be a route to avoiding the preconditions of testimonial smothering does not demonstrate they are a good idea, all things considered), others would presumably be less so. Measures to reduce the attractiveness of threatening section to keep service users on wards 'voluntarily', or locking doors to achieve a similar effect, should be acceptable to most, if not all observers. This is because such cases are, patently, misuses of power, and therefore morally objectionable in their own right.

The last two of these suggestions both amount to, in general terms, reducing the legal and de facto power that clinicians hold over service users in clinical contexts. I will return shortly to reasons why one may find such reduction an unjustifiable move. One potential misunderstanding should be defused immediately, however. It will not suffice in this case for clinicians to simply stop using their powers irresponsibly, while nevertheless formally maintaining them. This is because the *risk* of such powers being misused and causing harm (especially from the perspective of a distressed service user) will remain, even if clinicians are disposed to use them more responsibly. I take it that the overwhelming majority of clinicians are *already* disposed to use these powers justifiably, if not always entirely appropriately, given the pervasiveness of various forms of pernicious, though non-culpable, ignorance. The point is that it is the powers themselves, not strictly their misuse, which creates the risk that we are looking to reduce.

Regarding the risk of being refused or being made to feel undeserving of care; reducing clinicians' propensity to dismiss the wrong kind of testimony, or the right kind of testimony about the wrong sort of symptoms, can also be worked towards in a variety of ways. As well as training clinicians to respond better to such cases, we could also reduce the difficult circumstances that force clinicians to make difficult judgments regarding which service users most deserve resource-intensive treatment in the first place. Relevant improvements would include properly funding

alternatives to in-patient treatment (including community care), and having more beds available on in-patient wards. Naturally, none of these reforms are novel (or even controversial – at least among service users and healthcare professionals) suggestions. It is worth pointing out, however, that their utility in eliminating the preconditions of testimonial smothering has not before been offered as a reason in their favour.

Returning, however, to suggestions that reduce the power of clinicians, one might object that the power differential we currently have is in place for a good reason. Namely, clinicians need extensive powers to prevent service users from harming themselves or others. This is undoubtedly true in some cases, and I would not try to deny it. But nor do I need to. My claim is just that the requirements of testimonial autonomy must be weighed against these other requirements. The former must, as it were, get into the pool of reasons and actions being considered when we discuss how best to structure clinical psychiatric encounters. This is not to say that they always and forever override the need for clinicians to have extensive powers. That said, it seems as if many of the current powers are poorly scrutinised and, sometimes, irresponsibly used, in ways that unjustifiably increase the risk of testimonial smothering (amongst other injustices). Given this, serious consideration should be given to the thought that the current de facto powers of clinicians go beyond, and in some cases may actively work against, the demands of testimonial autonomy, in a way that is not morally justifiable.

6.5. Conclusion

In this paper, I have argued that epistemic oppression in Psychiatry exists beyond the testimonial injustice highlighted previously (Crichton et al 2016). In particular, I have argued that psychiatric service users are particularly vulnerable to three other notable forms of epistemic oppression – hermeneutical injustice, contributory injustice, and testimonial smothering. Moreover, I have argued that the existence of these injustices place certain kinds of demands on psychiatric education, clinicians themselves, and the formal and informal rules that govern their relationships with, and power over, service users in their care. I do not claim that these observations are exhaustive however, either of the epistemic injustices being done, or of the demands

they may make on appropriate psychiatric practice. Furthermore, it remains to be seen just how much overall reason we have to act to reduce or eliminate the occurrence of these injustices. This will depend both on their prevalence, and the urgency of competing moral requirements. That said, it is imperative that the injustices and attendant moral reasons I have identified here are included in the ongoing discussion regarding the ideal relationship between psychiatrists and those in their care.

6.6. References

Amador, X.F. Strauss, D.H. Yale, S.A.Gorman, J.M. (1991). "Awareness of illness in Schizophrenia". *Schizophrenia Bulletin 17*(1): 113-132

Anderson E. (2012). "Epistemic justice as a virtue of social institutions", *Social Epistemology 26*(2): 163-173.

Biafora, F. (1995). "Cross-cultural perspectives on illness and wellness: Implications for depression". *Journal of Social Distress and the Homeless 4*(2): 105-129

Beauchamp TL. (2010), "Autonomy and consent" in Miller F, Wertheimer A. (eds.) *The ethics of consent: Theory and practice*. Oxford: OUP: 55-78.

Blease, C. Carel, H. & Geraghty, K. "Epistemic injustice in healthcare encounters: Evidence from Chronic Fatigue Syndrome". *Journal of Medical Ethics 43*: 549-557

Bracken P, Thomas P, Timimi S et al. (2012). "Psychiatry beyond the current paradigm". *The British Journal of Psychiatry 201*(6): 430-434.

Bullmore, E. Fletcher, P. & Jones, P.B. (2009). "Why psychiatry can't afford to be neurophobic". *The British Journal of Psychiatry 194*(4): 293-295

Campbell P. (2001). "The role of users of psychiatric services in service development – influence not power". *Psychiatric Bulletin 25*(3): 87-88.

Care Quality Commission (2014). "Monitoring the Mental Health Act in 2012/13". Available at: https://www.cqc.org.uk/sites/default/files/documents/cqc_mentalhealth_2012_13_07_update.pdf

Care Quality Commission (2016). "Monitoring the Mental Health Act in 2015/16". Available at: https://www.cqc.org.uk/sites/default/files/20161122_mhareport1516_web.pd

Cox, C.L. & Fritz, Z. (2016). "Should non-disclosure be considered as morally equivalkent to lies within the doctor-patient relationship?". *Journal of Medical Ethics 42*: 632-635

Crichton P, Carel H, Kidd I.J. (2017). "Epistemic injustice in psychiatry". *Psychiatric Bulletin 41*(2): 65-70.

Cuthbert, B.N. & Insel, T.R. (2013). "Toward the future of psychiatric diagnosis: the seven pillars of RDoC". *BMC Medicine 11*(1): 126

Davidson, L. Borg, M. Marin, I. Topor, A. Mezzina, R. & Sells, D. (2005). "Processes of recovery in serious mental illness: Findings from a multinational study". *American Journal of Psychiatric Rehabilitation 8*(3): 177-201

Davies, W. (2016). "Externalist psychiatry". *Analysis 76*(3): 290-296

Dooley, D. Prause, J. & Ham-Rowbottom, K.A. (2000). "Underemployment and depression: longitudinal relationships". *Journal of Health and Social Behaviour*: 421-436

Dorais, M. (2004). "Dead boys can't dance: Sexual orientation, masculinity, and suicide". McGill: QUP

Dotson K. (2011). "Tracking epistemic violence, tracking practices of silencing". *Hypatia 26*(2): 236-257

Dotson K. (2012), "A cautionary tale: on limiting epistemic oppression", *Frontiers: A journal of women's studies*. 33(1): 24-47.

Drentea, P. & Reynolds, J.R. (2012), "Neither a borrower nor a lender be: the relative importance of debt and SES for mental health among older adults", *Journal of Aging and Health* (24:4), pp.673-695

Eyal N. (2012), "Informed consent", In Zalta EN. (ed.) *The Stanford Encyclopaedia of Philosophy (Fall 2012 Edition)*. Available from: https://plato.stanford.edu/archives/fall2012/entries/informed-consent/ [Accessed 4th January 2018].

Fallon, P. (2003). "Travelling through the system: the lived experience of people with borderline personality disorder in contact with psychiatric services". *Journal of Psychiatric and Mental Health Nursing 10(4)*: 393-401

Fricker M. (2007), *Epistemic injustice: power and the ethics of knowing*. Oxford: OUP

Gilburt, H. Rose, D. & Slade, M. (2008). "The importance of relationships in mental health care: A qualitative study of service users' experiences of psychiatric hospital admission in the UK". *BMC Health Services Research 8*: 92

Ghaemi, S.N. Hebben, N. Stoll, A.L. Pope Jr, H.G. (1996). "Neuropsychological aspects of lack of insight in bipolar disorder: A preliminary report". *Psychiatry Research 65*(2): 113-120

Gray AJ. (2002), "Stigma in psychiatry", *Journal of the Royal Society of Medicine*. 95: 72-76.

Hearing Voices Network (n.d. a), *About Voices and Visions*. Available from: https://www.hearing-voices.org/voices-visions/about/. [Accessed 5th January 2018)

Hearing Voices Network (n.d. b). *Hearing Voices Groups*. Available from: https://www.hearing-voices.org/hearing-voices-groups/ [Accessed 4th January 2018].

Himber, J. (1994). "Blood rituals: Self-cutting in female psychiatric patients". *Psychotherapy: Theory, Research, Practice, Training 31*(4): 620-631

Horn, N. Johnstone, L. & Brooke, S. (2007). "Some service user perspectives on the diagnosis of Borderline Personality Disorder". *Journal of Mental Health 16*(22): 255-269

Insel, T. & Quirion, R. (2005), "Psychiatry as a clinical neuroscience discipline", *Journal of the American Medical Association* (294), pp.2221-2224

Katsakou, C. Bowers, L. Amos, T.Morriss, R. Rose, D. Wykes, T. & Priebe, S. (2010), "Coercion and treatment satisfaction among involuntary patients". *Psychiatric Services 61*(3): 286-292

Kidd I.J., Carel H. (2017), "Epistemic injustice and illness", *Journal of Applied Philosophy*. 34(2): 172-190.

Kirmayer, L.J. & Crafa, D. (2014), "What kind of science for psychiatry?", *Frontiers in Human Neuroscience* (8), p.435

Kogstad R.E., Ekeland T.J., Hummelvoll J.K. (2011), "In defence of a humanistic approach to mental health care: recovery processes investigated with the help of

clients' narratives on turning points and processes of gradual change", *Journal of Psychiatric and Mental Health Nursing*. 18(6): 479-486

Kurs, R. & Grinshpoon, A. (2017), "Vulnerability of individuals with mental disorders to epistemic injustice in both clinical and social domains. *Ethics & Behaviour*. [Preprint]. Available from: doi:10.1080/10508422.2017.1365302.

Lewis-Fernández, R. Rotheram-Borus, M.J. Betts, V.T. Greenman, L. (2016). "Rethinking funding priorities in mental health research". *The British Journal of Psychiatry 208*(6): 507-509

Loch, A.A. (2012), "Stigma and higher rates of psychiatric re-hospitalization: São Paulo public mental health system". *Brazilian Journal of Psychiatry 34*(2): 185-192

Lovell K. (1995), "User satisfaction with in-patient mental health services", *Journal of Psychiatric and Mental Health Nursing*. 2(3): 143-150.

Maitra I. (2010), "The nature of epistemic injustice", *Analytic Philosophy*. 51(4): 195-211.

Manson NC, O'Neill O. (2007), "Rethinking informed consent in bioethics" Cambridge: CUP

McCarthy, J. Prettyman, R. & Friedman, T. (1995), "The stigma of psychiatric in-patient care". *Psychiatric Bulletin 19*(6): 349-351

Miller Tate, A. (2018), "Contributory injustice in psychiatry", *Journal of Medical Ethics* (online first), doi: 10.1136/medethics-2018-104761

Mind (2013a). "Sectioning". Available at: https://www.mind.org.uk/media/19803361/sectioning-2017.pdf

Mind (2013b). "Understanding self-harm" Available at: https://www.mind.org.uk/media/5133002/mind_und_self-harm_singles_4-web.pdf

Newman D, O'Reilly P, Lee SH et al. (2015), "Mental health service users' experiences of mental health care: an integrative literature review", *Journal of Psychiatric and Mental Health Nursing*. 22(3): 171-182.

Oakland L, Berry K. (2015), "Lifting the veil: a qualitative analysis of experiences in Hearing Voices Network groups", *Psychosis*. 7: 119-129.

Pilgrim, D. & Rogers, A. (2014), "A sociology of mental health & illness", New York: Open University Press

Pohlhaus, G. (2012), "Relational knowing and epistemic injustice: Toward a theory of willful hermeneutical ignorance", *Hypatia*. 27(4): 715-735.

Potter NN. (2009), *Mapping the edges and the in-between: a critical analysis of borderline disorder*. Oxford: OUP

Priebe, S. Burns, T. & Craig, T.K.J. (2013). "The future of academic psychiatry may be social". *The British Journal of Psychiatry 202*: 319-320

Rogers, B. & Dunne, E. (2011). "'They told me I had this personality disorder…All of a sudden I was wasting their time': Personality disorder and the inpatient experience". *Journal of Mental Health 20*(3): 226-233

Rutter, D. Manley, C. Weaver, T. Crawford, M.J. Fulop, N. "Patients or partners? Case studies of user involvement in the planning and delivery of adult mental health services in London". *Social Science and Medicine 58*(10): 1973-1984

Saul, J.M. (2012). "Lying, misleading, and what is said: An exploration in Philosophy of Language and in Ethics". Oxford: OUP

Simpson, E.L & House, A.O. (2002). "Involving users in the delivery and evaluation of mental health services: systematic review". *BMJ 325*(7375): 1265

Tait L. & Lester, H. (2005). "Encouraging user involvement in mental health services. *Advances in psychiatric treatment 11*(3): 168-175.

Thornicroft, G. & Tansella, M. (2005). "Growing recognition of the importance of service user involvement in mental health service planning and evaluation". *Epidemiology and Psychiatric Sciences 14*(1): 1-3

Tolmac, J. & Hodes, M. (2004). "Ethnic variation among adolescent psychiatric in-patients with psychotic disorders". *British Journal of Psychiatry 184*: 428-31

Wu, H. Dyck-Lippens, P.J.V. Santegoeds, R. van Kuyck, K. Gabriëls, L. Lin, G. Pan, G. Li, Y. Li, D. Zhan, S. Sun, B. & Nuttin, B. (2013). "Deep-brain stimulation for Anorexia Nervosa". *World Neurosurgery 80*(3-4): S29.e1-S29.e10

Zila, L.M. & Kiselica, M.S. (2001), "Understanding and counselling self-mutilation in female adolescents and young adults". *Journal of Counseling & Development 79*(1): 46-52