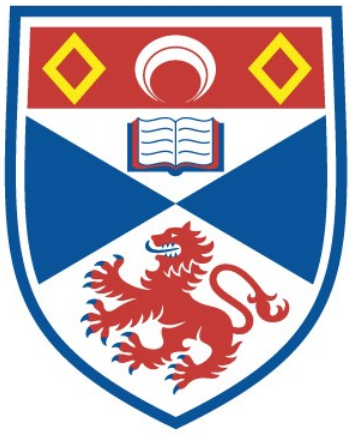
PHILOSOPHICAL PERSPECTIVES ON THE STIGMA OF MENTAL ILLNESS

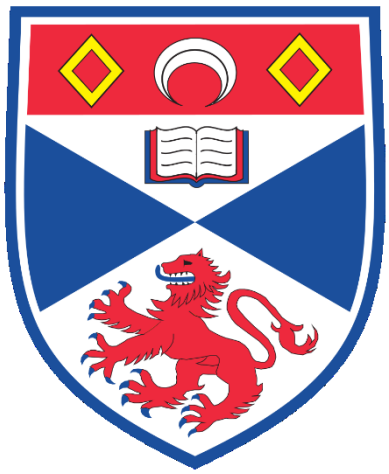
Lisa Rebecca Nowak

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews



2018

Philosophical Perspectives on the Stigma of Mental Illness



Lisa Rebecca Nowak, MA

University of St Andrews

School of Philosophical, Anthropological and Film Studies

March 2018

This dissertation is submitted to the University of St Andrews for the degree of Doctor of Philosophy

1. Candidate's declarations:

I, Lisa Nowak, hereby certify that this thesis, which is approximately 77,500 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in September 2014 and as a candidate for the degree of Doctor of Philosophy in September 2017; the higher study for which this is a record was carried out in the University of St Andrews between 2014 and 2017.

Date: 22-09-2017 signature of candidate:

2. Supervisor's declaration:

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Doctor of Philosophy in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date: 15-9-17 signature of supervisor:

3. Permission for publication: (*to be signed by both candidate and supervisor*)

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

PRINTED COPY
a) No embargo on print copy

ELECTRONIC COPY
a) No embargo on electronic copy

ABSTRACT AND TITLE EMBARGOES

An embargo on the full text copy of your thesis in the electronic and printed formats will be granted automatically in the first instance. This embargo includes the abstract and title except that the title will be used in the graduation booklet.

If you have selected an embargo option indicate below if you wish to allow the thesis abstract and/or title to be published. If you do not complete the section below the title and abstract will remain embargoed along with the text of the thesis.

a)	I agree to the title and abstract being published	YES
b)	I require an embargo on abstract	NO
c)	I require an embargo on title	NO

Date: 15-9-17 signature of candidate: signature of supervisor:

Please note initial embargos can be requested for a maximum of five years. An embargo on a thesis submitted to the Faculty of Science or Medicine is rarely granted for more than two years in the first instance, without good justification. The Library will not lift an embargo before confirming with the student and supervisor that they do not intend to request a continuation. In the absence of an agreed response from both student and supervisor, the Head of School will be consulted. Please note that the total period of an embargo, including any continuation, is not expected to exceed ten years.
Where part of a thesis is to be embargoed, please specify the part and the reason.

ACKNOWLEDGEMENTS

I am hugely grateful to my supervision team at St Andrews: Prof Katherine Hawley and Dr Simon Prosser. To Katherine, thank you so much for all your help. You have supported me so well through my application process all the way until the finished thesis, and I am hugely grateful. To Simon, I want to say a huge thank you for all the discussion and constructive criticism you have offered (and for having to suffer through so many formulations of the first chapter!).

Here I would also like to say a huge thank you to the Scottish Graduate School for the Arts and Humanities for making my PhD study possible at all. I am so grateful for the generous funding you have given me, and for the ongoing support and development you have provided.

Finally, I would like to thank my family and friends. To Mum- even though you have zero interest in philosophy, you have supported me all the way, and have put up with what must be very boring PhD updates. Don't worry, you can just pretend you've read it. Joe- thank you for all the support!

For Mum & Joe

ABSTRACT

This thesis is concerned with philosophical perspectives on the stigma of mental illness, with each chapter exploring different philosophical issues. Chapter one delineates the central concept around which the rest of the work revolves: the stigma of mental illness. It provides an outline of the stigma mechanism, how it applies to mental illness, why it is such a large public health concern and what has been done so far to combat it.

Chapter two is concerned with the application of recent literature in the philosophy of implicit bias to the topic of mental illness. It suggests that we have hitherto been preoccupied with explicit formulations of the stigma mechanism, but argues that there are distinctive issues involved in combatting forms of discrimination in which the participants are not cognisant of their attitudes or actions, and that anti-stigma initiatives for mental illness should take note.

Chapter three applies the philosophical literature concerning the ethics of our epistemic practices to the stigma of mental illness. It contains an analysis of how epistemic injustice- primarily in the forms of testimonial injustice and stereotype threat- affects those with mental illnesses.

The fourth chapter brings in issues in the philosophy of science (particularly the philosophy of psychiatry) to explore the possibility of intervening on the stigma process to halt the stigma of mental illness. The first candidate (preventing labelling) is discounted, and the second (combatting stereotype) is tentatively endorsed.

The fifth chapter is concerned with how language facilitates the stigma of mental illness. It suggests that using generics to talk about mental illness (whether the knowledge structure conveyed is inaccurate or accurate) is deeply problematic. In the former, it conveys insidious forms of social stereotyping. In the latter, it propagates misinformation by presenting the category as a quintessential one.

294 words

Thesis is approximately 78,000 words.

CONTENTS

IV. Broadening Education Strategies
    i. Healthcare Workers
    ii. Policing
    iii. Other Groups & Further Thoughts

V. Further Strategies
    i. Alteration of Context
    ii. Protest
    iii. Hard Work

VI. Conclusion

CHAPTER THREE: Power, Epistemic Injustice and Mental Illness

I. Introduction

II. Fricker on Testimonial Injustice

III. Testimonial Injustice and Mental Illness
    i. The Harms of Testimonial Injustice

IV. Stereotype Threat

V. Stereotype Threat, Doubt and Mental Illness
    i. Mental Illness and Doubt
    ii. Potential for Explaining Behaviour

VI. Combatting Epistemic Injustice
    i. Mental Illness, Epistemic Deficit and the Cooperative Principle

VII. The Role of Advocacy

VIII. Conclusion

CHAPTER FOUR: Combatting Stigma: Labels, Stereotype and Language

I. Introduction

II. The Stigma Mechanism: Possible Interventions

III. Benefits of Labelling
   i. Induction, Prediction and Diagnosis
   ii. Psychiatry as a Scientific Discipline
   iii. Benefits for Patients

IV. Disregarding Stereotype?

V. Conclusion

CHAPTER FIVE: Generics and the Stigma of Mental Illness.

I. Introduction

II. What are Generics?

III. Why are Generics Problematic?
   i. Quintessentialism
   ii. Quintessentialism and Generics

IV. Generics and Mental Illness
   i. Mental Illness as Stable, Serious and Enduring
   ii. Quintessential Similarity, Induction and Prediction
   iii. Entiativity
   iv. Contagion
   v. Drawing Together

V. What should we do when we Hear Generics?
   i. Metalinguistic Blocking
   ii. Going Forward

VI. Conclusion

FINAL THOUGHTS

I. Summary

II. Practical Recommendations

BIBLIOGRAPHY

CHAPTER ONE

Stigma: Mechanism and Harms for Mental Illness

I. Introduction

Mental illness poses one of the most serious threats to worldwide public health. Studies indicate that mental illness is one of the main constituents of the overall disease burden worldwide (Vos et. al 2015). In the UK mental illnesses constitute 28% of the disease burden, whereas heart diseases contribute roughly 16% (The Mental Health Foundation). Research shows that mental illness and related behavioural issues such as anxiety and drug use are the leading drivers of disability worldwide (see Mental Health Foundation 2016 and Lozano et. al 2012). Major Depression is thought to be the second leading cause of all worldwide disability, and it significantly contributes to the number of suicides and incidents of ischemic heart disease (see Whiteford et. al 2013). In 2009, a study regarding adult psychiatric morbidity in England reported that one in four adults, and one in ten children, are likely to have a mental illness issue in any given year (see The NHS Information Centre for Health and Social Care 2009).

It is well known that metal health services in the UK are often under-funded, over-stretched, and suffer from long waiting times. Indeed, only 25% of those with mental illnesses are in

receipt of ongoing treatment, meaning that the remaining 75% are left to struggle with their issues on their own, with access only to the informal support offered by family, friends, colleagues, or other untrained support networks (Mental Health Foundation 2015). However, it is estimated that the provision of the existing services, in conjunction with the costs incurred by mental health problems, costs the UK £70-100 billion each year (The Mental Health Foundation 2015). Indeed, it is estimated that the total global cost of mental health problems is around £1.6 trillion: the largest single source of world economic burden (see Mental Health Foundation 2016).

Clearly then, mental illness contributes hugely to the global burden of disease, and poses a threat to public health worldwide. It poses a social cost, yet also a serious financial one. As a consequence, a huge amount of intellectual and monetary resources have been spent trying to improve mental health and to tackle the challenges associated with mental illness. This work will address some of these challenges, yet it will focus more upon the social, rather than the medical challenge. That is, whilst there are many interesting questions concerning what mental illness is and how it should be treated, I propose to look not at these, but rather, to consider the problem of mental illness stigma, and the related questions of how we can understand it, and what should be done about it. This will be the central topic around which this thesis revolves. Exploring and challenging the stigma of mental illness is an important task, given that stigma generates numerous significant harms to those exposed to it. Indeed, the problem of mental illness stigma is widely recognised within the literature, with Gaebel et. al (2017) noting that:

> the stigma of mental illness is still the main obstacle to the development of mental health services and a heavy burden for all touched by mental illness, people who have them, their families, mental health workers, mental health services and treatment methods. There is no 'end of the story' of fighting stigma in sight.

Indeed, it is because stigma continues to be such a problem that it is a worthy topic of investigation.

i. Structure of the Work

Yet what is stigma? Chapter 1 will focus on answering this question. I will do so by firstly providing some clarity on the stigma process, and its related concepts. This is an important task, particularly given that the term 'stigma' and its related terms of art 'labelling', 'stereotype', 'prejudice' and 'discrimination' are commonly known, yet many are used interchangeably in public discourse, and we may think they all mean much the same thing. However, there are differences to be attended to, and some explanatory value in delineating these concepts. In the following section, I will construct an operational account of the stigma mechanism using two of the most influential accounts offered in the literature. On this operational account, 'stigma' refers to the simultaneous operation of a series of processes. Chapter 1 will also outline why it is that stigma is so problematic for mental illness. I will briefly summarize the literature on mental illness stigma, identifying the myriad harms which are thought to be associated with it. Doing so will help to define the problem further, demonstrating why the stigma of mental illness is a worthy problem for a project such as this to address. Finally, chapter 1 will also outline some of the current strategies utilised to combat mental illness stigma. The reason for this is that later chapters will explicitly refer back to current strategies, and I will analyse them in light of my findings.

With the problem of mental illness stigma defined and characterised, the remainder of the project will be concerned with its examination. Whilst much has been written in social

psychology, politics, and the medical humanities about the problem of mental illness stigma, I propose to take a slightly different route. That is, further chapters of this work will be concerned with how we might apply different tools of philosophical analysis to the problem of mental illness stigma, or how different debates in philosophy can usefully comment upon the root causes of stigma, its manifestations/aspects and potential ways in which we might combat it. Each chapter will take a different perspective on the question of what philosophy has to say about mental illness stigma, although there will inevitably be common themes or crossovers.

Indeed, I believe that philosophical analyses of the stigma of mental illness will be profoundly useful, yet there is somewhat of a gap in the current literature. That is, whilst much work has been done in philosophy on race and gender stigma, far less has been said about stigma and mental illness (although mental illness has been the subject of a great deal of philosophical interest in other areas). This work is intended to apply various areas of philosophical discussion to the topic of mental illness, and to explore what we can learn from doing so. Throughout this thesis I suggest that this application of philosophical concepts to mental illness will yield varied and interesting conclusions, where our treatment of the relevant issues differs for mental illness when compared to efforts to end stigma for gender and race. Simply put, stigma in mental illness cannot be treated in the same way as gender and race, and will require different responses. It is partly for this reason that I think the current project is a worthwhile and significant one. With this said, I will now briefly outline the structure of what is to come.

As noted above, the problem of mental illness stigma is to be the foundation around which this work is based, from which I will then identify various philosophical issues which arise. The second chapter is motivated by recent work in the philosophy of implicit bias: a form of implicit

social cognition which is thought to have a profound effect on behaviour. In chapter 2, I suggest that findings about implicit bias raise a variety of issues for mental illness stigma. Most pressingly, it alerts us to the possibility that various elements of the stigma process (and indeed, stigma itself) may be conducted implicitly, rather than explicitly. That is, we can discriminate or stigmatize even where we are not consciously aware of our doing so. This has several implications for mental illness stigma. For one- features of implicit bias make it such that traditional education strategies for tackling stigma in mental illness may not prove effective. As a result, different kinds of strategies are necessary if we are to combat this social problem.

Chapter 2 will therefore give some detail as to what should be done to combat the stigma of mental illness, given the insights offered by the literature on implicit bias. Particularly, I suggest that certain 'target groups' should be identified and educated about implicit social cognition and bias. These target groups are to be identified according to the risk they pose of enacting implicit stigma. I suggest that this 'risk' is identified according to social power and the likelihood that one's working environment will prime the activation of 'type-1' cognitive processes (which are thought to often bring about biased action). I will conclude chapter 2 by detailing how other kinds of strategies for combatting the stigma of mental illness, aside from education, might be amended.

Chapter 3 applies the philosophical literature concerning the ethics of our epistemic practices to the stigma of mental illness. According to the stigma mechanism as outlined by Link & Phelan (2001), one of the processes which constitutes stigma is status loss. I suggest that one profound form of status loss encountered by the mentally ill is a loss of epistemic status. In some cases, this is illegitimate, and thus the mentally ill face what Fricker (2007) calls

'epistemic injustice'. In chapter 3, I provide an analysis of how epistemic injustice- primarily in the forms of testimonial injustice and stereotype threat- affects those with mental illnesses. I note that the situation for mental illness is complicated by the fact that in some cases, the loss of epistemic status which accompanies being in receipt of or associated with a mental illness label appears justified, yet in others it is not.

For mental illness, establishing whether someone is afforded what Fricker calls an 'undue' deficit in credibility is a complicated matter. For one, it does not seem epistemically prudent to simply ignore stereotypes about mental illness as some are accurate or informative. Yet on another count, it is problematic to simplistically hold the stereotype to be true of all members. In response, I suggest that the stereotype can be used to make judgements of credibility, but it should be done in as minimal a way as possible. Indeed, I suggest that there are certain conditions on the use of stereotype, and that credibility judgements should only be made where necessary, and according to what I have called a 'capacity model'.

Chapter 3 will also discuss another form of epistemic injustice: stereotype threat. I suggest that stereotype threat, and the profound epistemic doubt Goguen (2016) describes as accompanying it, will likely bite particularly hard for mental illness, given the other sources of epistemic doubt to which they are exposed (some legitimate, and others not). Finally, I argue that the phenomenon of stereotype threat may actually explicate some of the behaviours seen in mentally ill people: for instance, the tendency towards social isolation and treatment avoidance.

The fourth chapter returns its attention to the stigma mechanism, and attempts to locate points at which we might hope to intervene on the process in order to prevent mental illness stigma

from occurring. The first suggestion explored by this chapter is that in preventing labelling (the identification and marking of social difference), perhaps we might prevent stigma from occurring. In this sense, the stigma mechanism is interrupted at the outset. I suggest that preventing labelling is a dubious project even where race and gender are concerned: labels can be said to be valuable in that they provide us with a lexicon through which to understand, interpret and talk about experiences of discrimination and social prejudice. Yet, I argue that there are further reasons to think that we will not be able to get rid of labels in mental illness and psychiatry, or that we should not attempt to. Mental illness categories play an indispensable role in both the conceptual and practical viability of psychiatry. For one, many scholars have commented that in the face of psychiatric pluralism, psychiatry is in need of a system of classification furnished with labels if it is to remain a respectable and unified discipline. Further to this, mental illness labels have great clinical utility in that they permit induction, prediction and explanation, whilst also granting many benefits to patients, both in terms of understanding and identity. For these reasons, I suggest that getting rid of mental illness labels in a bid to end stigma is inadvisable, if not pragmatically impossible.

Chapter 4 will also explore another suggestion: that perhaps rather than getting rid of labels, we should get rid of stereotype. I will note that suggestions of this kind have been argued to present a moral-epistemic dilemma: namely, that in attempting to get rid of stereotyping for ethical reasons, we incur an inescapable epistemic cost. I argue that this dilemma is yet more complicated for mental health, given that some of the stereotypes are not ethically problematic in the same way the others are, and we want to retain them. As a result, our strategies for intervening on stereotype must be very nuanced. Indeed, I suggest that this level of nuance may be hard to achieve.

The fifth chapter is concerned with how language facilitates the stigma of mental illness. I argue that the use of generics to convey information about mental illness is deeply problematic, whether the knowledge structure conveyed is 'accurate' or not. Inaccurate generics about mental illness can be problematic in much the same way that Leslie (2013) describes: 'inaccurate' generics can be acceptable to us on the grounds of truth conditional laxity, often where the property described is a 'striking' one. Thus, generics contribute hugely to insidious forms of social stereotyping and stigma as characteristics found in very few category members can be simplistically generalised to other members of that kind.

Yet, we might also use generics to convey 'accurate' stereotypes about mental illness: for instance, 'schizophrenics experience delusions'. I suggest that even though the knowledge structure is not problematic (insofar as it describes an accepted clinical perspective), the use of the generic form is. This is because quintessentializing mental illness kinds can generate a range of problematic beliefs: i.e. those that are inaccurate, appear to deny the possibility of recovery or meaningful change, propagate separatist inclinations and frustrate contact and understanding. As a result, I suggest that to fight stigma, we should refrain from using generics to describe mental illness at all, and instead use quantified utterances. Yet, we must also seek to challenge generic utterances where we find them. I argue that whilst generics appear to evade direct refutation by evidence, one promising mode of challenging them is to engage in what Haslanger (2011) terms 'metalinguistic blocking': where the challenge in question does not claim that the generic is strictly 'false', but rather claims that the use of the generic pragmatically implicates a falsehood which can lead to the cultivation of damaging forms of social reality.

II. What is Stigma?

With the structure of the work laid out, we are now in a position to commence. Given that the stigma of mental illness is the subject and motivation for this project, it is important to get clearer on what stigma is. Stigma is often defined as a 'mark of disgrace'. Those who are the bearers of stigma are often *stigmatized* by the wider public: that is, described or regarded as worthy of disgrace or great disapproval. Combining these two aspects, the Time to Change Campaign (2014, no pagination) describes stigma as "the perception that a certain attribute makes a person unacceptably different from others, leading to prejudice and discrimination against them". An important early discussion of stigma was offered by Goffman (1973) in his influential work *Stigma: Notes on the Management of Spoiled Identity*. Goffman (1973, p.3) utilised the term *stigma* to refer to "an attribute that is deeply discrediting within a social interaction", in which the affected individual is "reduced…from a whole and usual person to a tainted, discounted one". On Goffman's account, stigma is something which is conferred upon an individual or group of individuals by others, not an attribute of the person or persons. Hence, his claim that we must speak "a language of relationships, not attributes" (Goffman 1973, p.3): the stigma is not itself a feature of an individual, but a feature of the way others view them.

The stigma concept has proved profoundly influential. As Link & Phelan (2001) observe, it has been applied to and used to explain common societal reactions to many phenomena, including exotic dancing, AIDS, cancer and most significantly for my purposes, mental illness (see Angermeyer & Matschinger 1994, 2003, Corrigan & Penn 1999 and Phelan et. al 2000). Due in part to the sheer variety of subject matter to which it has been applied, the concept of stigma itself admits of a high degree of variability within the literature: a fuller exploration of which can be found in Stafford & Scott (1986) and Link & Phelan (2001, pp.364-365). I acknowledge

that there are many ways of characterising the concept, and stress that the account offered here is intended to be operational only. However, in order to explore stigma and mental illness it makes a lot of sense to refer to a framework.

In order to lay out a framework through which we can understand stigma, I will outline two accounts of it. This may seem puzzling, but my reasons for this are as follows. The first account belongs to Link & Phelan (2001), and I include it here because their reconceptualization of stigma as a complex process dependent upon political, social and economic powers is perhaps the most influential model of stigma in the literature, perhaps second only to that of Goffman himself. It has been widely cited, and continues to be so to the present day, serving as the basis for many empirical studies concerned with the stigma concept and areas in which it might apply[1]. In a discussion of stigma, the absence of this account would be noticeable indeed. Yet, this account is theoretical, and aims to describe stigma generally rather than as applies to any one discipline. As such, I will also refer to Corrigan's (1998, 2004, 2016) account of the stigma process, which explicitly deals with stigma and mental illness. Using this account will allow me to ground the stigma model, whilst also identifying some of the language commonly associated with the stigma of mental illness, yet which is absent from Link & Phelan's models.

i. Link & Phelan's Theoretical Account of Stigma

Both accounts named above recognise stigma to be the result of a series of social-cognitive processes[2]. Link & Phelan (2001, p.367) argue that stigma should be defined in the relationship of interrelated components:

[1] For instance, Link & Phelan's account has formed the basis for recent work on stigma in regard to the parents of autistic children (Kinnear et. al 2016), HIV and sex work in Western Kenya (Pfeiffer & Maithya 2016) and psychosis in Sub-Saharan African cities (Makanjuola et. al 2016).

> Stigma exists when the following interrelated components converge. In the first component, people distinguish and label human differences. In the second, dominant cultural beliefs link labelled persons to undesirable characteristics- to negative stereotypes. In the third, labelled persons are placed in distinct categories so as to accomplish some degree of separation of "us" from "them". In the fourth, labelled persons experience status loss and discrimination that lead to unequal outcomes. Finally, stigmatization is entirely contingent on access to social, economic and political power that allows the identification of differentness, the construction of stereotypes, the separation of labelled persons into distinct categories, and the full execution of disapproval, rejection, exclusion and discrimination. Thus we apply the term stigma when elements of labelling, stereotyping, separation, status loss and discrimination co-occur in a power situation that allows the components of stigma to unfold.

The first component is intended to emphasise the social selection of human differences. That is, many forms of human variation are not considered to be socially relevant: for instance, what colour car you drive. However, certain kinds of human differences are: whether you are mentally ill, of a particular race or gender, and so on and so forth. If you are different in a manner deemed socially relevant, then this may affect the treatment you receive. However, Link & Phelan stress that whatever constitutes relevant forms of human difference in any context is socially selected. That this is true can be demonstrated in several ways: namely, that drastic oversimplifications are required to create groups (and hence the delineations which must be made are socially enforced rather than naturally given) and that the attributes which are socially relevant vary across time.

The second component involves the link between labelled differences and stereotypes (or the set of undesirable characteristics which constitute the stereotype). Link & Phelan add that investigation into the links between labels and stereotypes has revealed that culturally given categories and stereotypes seem to operate at a preconscious level, providing a shorthand or automatic means of making decisions, freeing the individual up to attend to other concerns. The third feature of stigma involves the separation of "us" from "them". In many cases, the stereotyping conducted in component two can function as a rationale for separation: that is, for believing that the labelled persons are different, perhaps fundamentally so, from those who do not bear it. In the most extreme case, the labelled individual or group can cease to be viewed as human. In some instances, atrocities can be justified by this mode of thinking. Interestingly, Estroff (1989) notes that features of our language may be evidence of concerted efforts to separate "us" from "them". For instance, discredited individuals are often spoken of as if they *are* the very things with which they have been labelled. People talk of 'schizophrenics' rather than 'people with schizophrenia': as a fundamentally other "them", rather than someone like "us" who happens to have a particular condition. Yet, this does not occur for illnesses like cancer or heart disease. Indeed, this will be of note later in this paper.

Link & Phelan's fourth component concerns status loss and discrimination. Precisely, the individual's status is reduced in the eyes of the stigmatizing party by being linked to negative characteristics through the process of stereotype. The stigmatized individual is relegated in one or several of the many status hierarchies that humans construct and maintain. For Link & Phelan, losses in status can lead to tangible forms of inequality. For instance, discrimination occurs where a person comes to have certain beliefs and attitudes about another individual as a result of labelling and stereotyping, which then cause them to treat the labelled individual in

a different manner to those they perceive to be un-labelled: they may refuse to employ a labelled person, or limit their social contact with them.

Where this process occurs frequently or becomes the norm institutional discrimination can occur, in which one group is systematically treated differentially in an institution, setting or social locale. Link & Phelan (2001, p.373) note that whilst status loss and the resulting discriminatory behaviour are often founded upon a prior knowledge of labelling and stereotype, status loss itself can become the grounds for discrimination. That is, where an individual or group becomes aware that another person is of reduced status-for instance, if they observe that person being avoided or marginalised- they may then implement discriminatory behaviour accordingly: for instance, by likewise avoiding them.

Link & Phelan also describe a fifth stage in the stigma process: a stage which notably, is absent in Corrigan's account. This fifth element states that it is necessary that the stigmatizing party be in possession of a suitable degree of power with which to carry out their ends. That is, it is only possible to stigmatize when one has the power to do so. This is best illustrated through the following example (Link & Phelan 2001, pp.375-376). Suppose that patients in a psychiatric unit practise many of the same behaviours outlined above. That is, they identify certain members of staff in their facility as being different from others and tag them with the label "pill pushers". They then form stereotypes around these labels- perhaps that 'pill pushers' are cold, paternalistic and arrogant. Finally, they treat these members of staff differently to others: they discriminate against them by making derogatory comments or jokes about them, or simply avoid contact with them altogether.

However, even if the patients engage in all these behaviours, (which together constitute the previous four stages of the stigma process outlined above) Link & Phelan suggest that the staff do not constitute a stigmatized group: "the patients simply do not possess the social, cultural, economic, and political power to imbue their cognitions about staff with serious discriminatory consequences" (Link & Phelan 2001, p.376). Thus, all the other conditions of the process may be met, but in the absence of power, stigma does not occur. One may designate human differences as socially relevant and treat individuals differently accordingly, yet if one lacks the power to enforce the relevance of these differences on a broader social level, or make the distinction between 'us' and 'them' stick, then stigma does not result. Hence, the stigmatizing party needs to have the social, political and economic power necessary to make the distinctions they draw manifest in the real world, and for this to really *matter*.

ii. Corrigan: Stigma & Mental Illness

This concludes the brief outline of Link & Phelan's account, and I turn now to Corrigan (2004, 2016). His account of the stigma process is relatively similar, although some of the stages differ somewhat, and the power condition outlined above is not specified. Corrigan's first step describes the identification of relevant forms of social difference, much as Link & Phelan's does. Applied to the issue of mental illness, Corrigan (2004, p.614) describes that 'cues' are the first stage of the stigma process. In this first step, the general public infer the presence of mental illness in another individual through the perception and recognition of four cues: "psychiatric symptoms, social-skills deficits, physical appearance and labels" (Corrigan 2000). Explaining the first cue, Corrigan notes that illnesses such as psychosis often manifest

themselves in displays of inappropriate affect or bizarre behaviour, which serve as strong indicators of an underlying illness, and provoke stigmatizing reactions in response (Link et. al 1987, Penn et. al 1994). Furthermore, there is also evidence to suggest that demonstrating poor social skills (Bellack et. al 1990) and an unkempt appearance (Penn et. al 1997) can bring about stigmatizing responses.

Corrigan acknowledges that false positives occur where we mistake eccentricity, rudeness, financial/ personal hardship or being under the influence of substances for mental illness. False negatives may also arise- for instance, where individuals are able to disguise their illnesses. This leads Corrigan to argue that there must be another means of inferring or establishing the presence of mental illness in others. He proposes 'labelling' as a candidate. Labels can generate stigma in one of two ways, generally speaking. Firstly, one can be directly informed that another individual has attracted a certain diagnostic label. In the psychiatric context, this is akin to a psychiatrist, medical professional, or even a member of the public announcing 'Sarah has depression', or perhaps reading it on her medical records. Secondly, one can attain a label through association. For instance, perhaps you see Sarah leaving a psychiatric unit, or you notice a packet of Citalopram in her bathroom. In either case, you infer that Sarah has a mental illness on the basis of her being associated with institutions or treatments designed to treat mental illness.

Corrigan's second stage of stigma is stereotyping. Corrigan (2004, p.615) notes that "stigmas are cues which elicit stereotypes, knowledge structures that the general public learns about a marked social group". Stereotypes, conceived of as 'knowledge structures', are an efficient means by which humankind is able to categorize and process information about social groups. They are composed of collectively agreed upon ideas or beliefs about a given social group, and

are available to, or known by, the majority of the public. Two things might be noted here. Firstly, as 'knowledge structures', stereotypes need not be inherently problematic: indeed, they may reflect useful empirical generalisations. Secondly, stereotypes are 'efficient' in that they function as cognitive shortcuts: people can quickly generate beliefs and impressions of a member of a stereotyped group (see Hamilton & Sherman 1994). Stereotyping of this kind need not be pejorative. For instance, you might see a man in a hi-vis jacket and a helmet, reason that he is a police officer, and therefore infer that he is someone authoritative, whose assistance you might seek if you were in trouble (Corrigan 1998, p.209). However, *stigmas* are stereotypes which are specifically negative in character.

Secondly, being aware of a certain knowledge structure or stereotype is not the same as endorsing it, or believing it's content to be true (see Devine 1989 and Jussim et.al 1995). Indeed, endorsement of negative stereotypes constitutes the third stage of Corrigan's stigma process: prejudice. The development of *prejudice* as an evaluative response to the beliefs associated with the stereotype. Those who are prejudiced are likely to endorse negative stereotypes, for example in the manner of: "that's right; all people with mental illness are violent and incompetent" (Corrigan 2004, p.616). Further to this, they develop negative affective responses as a result of this endorsement: for instance, "I am scared of them" (see Hilton & von Hippel 1996, Krueger 1996).

This cognitive-affective response serves as the foundation for Corrigan's fourth phase of the stigma process: discrimination. Here, the cognitive-affective response associated with prejudice leads to certain discriminatory behavioural reactions (see Fiske 2000). According to Corrigan (2004, p.616), "discriminatory behaviour manifests itself as negative action against the out-group or exclusively positive action for the in-group". He notes that out-group

discrimination most often manifests in the exclusion or avoidance of marginalised groups. For instance, an employer who endorses the stereotype that mentally ill people are incompetent may refuse to hire a candidate with a history of depression. In this sense, the cognitive-affective content of the prejudice stage is manifested in action against the party who attracted the prejudice.

iii. Summary

If we were to compare both these accounts, we might notice some strong similarities. In both, stigma is the result of several inter-connected socio-cognitive processes. Broadly speaking, in the first instance human differences are identified (in Corrigan's case via the identification of cues), designated as socially relevant and *labelled*. Secondly, labelled differences are connected to *stereotypes*: knowledge structures which are "often framed as seemingly fact-based beliefs with a negative evaluative component" (Corrigan 2016). The third stage does, however, differ slightly. For Link & Phelan, the separation of 'us' from 'them' occurs, whereas for Corrigan, stereotypes are endorsed in the form of prejudice. For both, these cognitive elements then produce physical manifestations in the form of *discrimination*: behavioural responses are, as Corrigan (2016) notes, usually in some sense punitive. To this, in line with Link & Phelan's analysis, we might add that stigma occurs only where the stigmatizing party has the power necessary to make their cognitive and behavioural aspects *matter*: that is, for them to be broadly accepted in the culture, or to influence the structure of institutions.

Thus, I hope to have offered a broad account from the literature of the stigma process and its plausible manifestation for mental illness, which I take to be sufficient to inform the rest of this work. Indeed, it is not my purpose here to argue for what I take to be the most plausible variation of the stigma mechanism. Rather, I hope to propose an operational account in terms of which the rest of my work can be understood. I take it that the stigma model is now adequately defined, and turn now to fleshing out why stigma is such a problem for mental illness, and the forms it can take.

III. The Types and Harms of Stigma

There are further conceptual distinctions which may pose useful for this project. Stigma can be further delineated into three distinct- although pragmatically interrelated- kinds. These are usefully distinguished by Corrigan (2016): public stigma, self-stigma, and label avoidance. ***Public stigma*** is what "a naïve public does to the stigmatized group when they endorse the prejudices about that group" (Corrigan 2004, p.616). It occurs where an individual or collection of individuals comes to identify someone as a member of a socially designated group, is aware of and endorses the stereotypes related to that group, and thus practices discriminatory behaviour against its membership. Much of the literature on the stigma of mental illness is concerned with public stigma. The harms of public stigma as regards mental illness are well-documented (see summary below), as are the prevalent stereotypes: amongst the most common are that the mentally ill are dangerous or psychopathic (and therefore should be feared and avoided); that they are rebellious free spirits (who are thus irresponsible or

incompetent, and should not be entrusted with important matters); and that they have childlike perceptions of the world (and must therefore be taken care of)[3].

These stereotypes are often thought to form the rationale for discrimination. As Corrigan (2004, 2016) notes, discrimination is often punitive in nature, and frequently involves the stigmatized party receiving adverse reactions from the stigmatizing party, or being denied their rightful opportunities. On the former point, it has been found that the most common form of public discrimination against the mentally ill is avoidance or withdrawal, often stemming from the belief that the mentally ill are dangerous or unstable (Feldman & Crandall 2007). It is also perhaps one of the most troubling forms of discrimination, given that it can engender a kind of social distance which, in conjunction with stereotype, serves to bring about a further loss of opportunities for the marginalised party.

Stigma can have a hugely detrimental effect in the distribution of life opportunities available to those who are marked with it. Those with mental illness labels are discriminated against in that their access to employment opportunities (Hipes et. al 2016, Krupa et. al 2009, Stuart 2004, Cook 2006), housing (Rüsch et. al 2005), medical care (Corrigan 2004) and independent living (Sartorius 2005, Sartorius & Callard 2012, Link & Phelan 2006) are severely limited. As a result of stigma, social interactions can become strained, and those with mental illness labels are often not granted the support or social networks they desire. Public discrimination of this kind can often engender depressive symptoms (Wright et. al 2000), feelings of shame and inadequacy (Corrigan 1998, 2004, Corrigan et. al 2014), loss of self-esteem (Wright et. al 2000) and can severely reduce the quality of life for those who attract mental illness labels (Link & Phelan 2006, Corrigan et. al 2014). Yet, stigma also creates less obvious harms. There is

[3] See Farina (1998), Hyler et.al (1991), Wahl (1995, 1999), Rüsch et. al (2005).

evidence to suggest that those with mental illness labels are denied access to medical opportunities. Indeed, Corrigan (2004, p.614) notes that research has uncovered two disconcerting trends: firstly, that many mentally ill people never pursue treatment at all, and secondly, that many that do not undertake the full course of treatment recommended.

Why might those who attract mental illness labels fail to engage with treatment in this way? There will be numerous reasons for this, however, one plausible response has been suggested by Corrigan (2004, 2016). He argues that if people engage with treatment and the medical institutions which provide it, they are likely to be marked or labelled as being 'mentally ill': that is, pursuing treatment is one way in which the public may come to identify the individual as a member of a stigmatized group. Hence, in at least some cases, Corrigan theorizes that to avoid being labelled (and being subjected to the harms and diminished social opportunities described above), some mentally ill people may practise the third type of stigma: ***label avoidance.*** Here, individuals attempt to deny their group status by refraining from interacting with the procedures and treatments which would mark them as members of the stigmatized group: they avoid treatment or cease participating in it so that they cannot be identified as being mentally ill. Fear of public reaction becomes an active barrier to treatment, and prevents people who might benefit from treatment from accessing it (Kessler et. al 2001, Abiri et. al 2016). Thus, the fear of public stigma may be one reason why people with mental illnesses avoid being labelled, although there will be others.

In general, label avoidance is particularly prevalent in mental illness. The most likely reason for this is that it is what Goffman referred to as a *x* rather than a *y* mark: in many cases, the mark of shame can be concealed. As such, the dilemma of disclosure arises: should one reveal that one is mentally ill? On the one hand, doing so would expose one to the kinds of public

stigma described above, and thus to the related harms. One's psychological well-being and social opportunities are likely to be severely threatened (Verhaeghe et. al 2008). Yet on the other hand, non-disclosure forces one into a life 'in the closet', which carries with it its own psychological burdens and stresses, and prevents one from attaining treatment which may be beneficial (Pachankis 2007). Either way, the individual is placed under significant mental strain. Thus, it is easy to see how labelling and label avoidances constitute psychological stressors.

Self-stigma occurs where a member of the stigmatized group directs public attitudes towards herself and comes to understand herself in terms of these stereotypes and labels. By internalizing public stigma, the individual comes to stigmatize and discriminate against herself just as others do. Self-stigma is usually analysed in terms of three successive stages, which have been dubbed the 'Three As': aware, agree and apply (Corrigan 2016, Corrigan & Calabrese 2005, Corrigan et. al 2006). Firstly, does the individual have knowledge of the various negative stereotypes surrounding mental illness? That is, does she know about the stereotype that the mentally ill are dangerous and should be feared or avoided? Secondly, does she agree with the content of this stereotype- "yes, mentally ill people are dangerous, and should be avoided"? Thirdly, does she apply this prejudicial attitude to herself? That is, "I have been diagnosed with a mental illness. I must be dangerous as well".

Corrigan (2016) argues that self-stigma can generate several harms, both cognitive and affective. For instance, self-esteem may be seriously damaged where an individual stigmatizes herself: "I am not a good person because I am a danger to others". The diminishment of self-esteem can itself generate further negative consequences for the individual, including the exacerbation of symptoms and a reduction in one's quality of life (Markowitz 2001, Vogel et

al. 2007). Stigma brings with it feelings of shame, and the fear of other people's reactions (Corrigan 2016). Both have been found to contribute to treatment avoidance (Sirey et. al 2001). Further to this, self-stigma poses a serious threat to self-efficacy, or one's belief that one can succeed in certain situations, or to achieve goals: for instance, "it will be impossible for me to ever get married, or to find a partner because I am dangerous, and others will avoid me".

As Corrigan et. al (2015, p.10) note, research has demonstrated that this diminished sense of personal effectiveness has been found to be linked to a failure to attain many goals associated with independent living, and a lack of success in the pursuit of employment and related objectives. For instance, the stigma surrounding mental illness has been found to have a detrimental effect on educational and vocational goals, particularly amongst college students (Garlow et. al 2008, Gibb et. al 2006). The lack of self-belief can generate profound behavioural implications. Most notably, it can create what has often been dubbed the 'why try' effect, in which an individual comes to believe that she is either incapable or unworthy of attaining her personal goals (as she has applied to negative stereotypes surrounding mental illness to herself), and as such, she believes that there is no sense in trying: "why try to make friends? I am incapable of doing so".

Finally, we may add another column to the matrix offered by Corrigan. This column concerns ***structural stigma***, which often occurs through "public and private sector policies that unintentionally restrict opportunities of the minority group" (Sheehan et. al 2017, p.51). Examples of this phenomenon for mental illness may include diminished quality of care or access to care (Link & Phelan 2001). However, we may note that there are also cases in which the opportunities of certain groups are intentionally restricted: for instance, in those cases

where certain states prevent those diagnosed as mentally ill from voting or holding office (Corrigan 2004).

IV. Tackling Stigma

As the above has demonstrated, the stigma of mental illness is a huge problem. Indeed, the stigmatization of the mentally ill has been recognised as one of the greatest challenges facing modern healthcare: a fact which is reflected both in research and in policy making. Indeed, it is because stigma is such a large and complicated problem that it was taken as the motivation for this project. In the material that follows, I will argue that there are many ways in which the application of philosophical tools of analysis to mental illness stigma will be illuminating. Indeed, one of these is that it will provide us with methods with which we can analyse existing anti-stigma initiatives. Recent work in implicit bias, epistemic injustice and the philosophy of language illuminate ways in which our current anti-stigma strategies are likely to fail, prove inadequate, or will require revision. I take this to be significant, particularly given that there is considerable ethical imperative to eradicate stigma.

Thus, at this point it would be useful to provide a brief analysis of anti-stigma campaigns for mental illness to date. For a much fuller analysis of this topic, I direct the reader to the recent edited collection *The Stigma of Mental Health- End of the Story?* (Gaebel et. al 2017). I should also note that I will not provide specific examples of anti-stigma initiatives here. Once again, readers interested in this should refer to the same volume. In the following, I will focus on the *types* of strategies that have been broadly identified to tackle the stigma of mental illness- often

by social psychologists. As many commentators have noted, anti-stigma strategies for mental illness can be divided into three distinct, yet interrelated types: education, contact and protest (Corrigan & Penn 1999, Corrigan 2004, Corrigan 2016, Rüsch et. al 2005, Rüsch & Xu 2017).

i. Education

Education is the most commonly implemented kind of strategy. Whilst it could be conceived of more broadly, most education strategies are concerned with the refutation of false information and the dissemination of correct information. The general idea is that the provision of accurate information and the discrediting of prevalent myths or inaccuracies will cause the public to hold less problematic and more informed beliefs about mental illness and the mentally ill. Education is often concerned with the correction of stereotypes: cognitively efficient knowledge structures, which are acquired in normal human social development, and of which the majority, or many, members of a culture are aware. For instance, 'people with psychosis are dangerous', or 'people with mental illnesses can't hold down jobs'.

The education approach has two prongs: firstly, general education about mental illness and secondly, de-bunking myths about it. The first is concerned with the formulation of accurate beliefs or knowledge structures, whereas the second is concerned with the de-bunking of problematic stereotypes through the presentation of contrary evidence. So, an example of the first strategy may be 'many people with mental illnesses are gainfully employed', whilst an example of the second might be 'it is not the case that people with mental illnesses are particularly dangerous'. Both kinds of strategy can be widely evidenced in anti-stigma campaigns.

In regard to first prong, the charity Mind includes on its website a large amount of information about varying mental illnesses and treatments, along with advice about housing, legal rights and services. Such efforts constitute an attempt to improve public knowledge of mental illness, and are often conducted in conjunction with education strategies designed to improve what Jorm (2011) terms 'mental health literacy': the extent to which the public are aware of how mental illnesses may be prevented, what treatments are available, and when and how to seek them. This kind of project is important as there is still a relative ignorance about mental illness and what it means for those affected by it. This kind of measure can be seen as a preventative one: the provision of an alternative knowledge structure discourages the endorsement of the negative stereotype, and instead encourages the formation of more accurate beliefs about mental illness. That is, telling the public about the low rates of crime amongst the mentally ill will produce a belief that 'the mentally ill are more likely to be the victims of crime' and so discourage the endorsement of the stereotype 'the mentally ill are dangerous'. Thus, education strategies can function by providing alternative and accurate knowledge structures.

However, education also aims to rectify misconceptions. In these cases, education seeks to factually refute the content of these problematic knowledge structures, and therefore to dissuade the endorsement of them. Where this project is successful, the education strategy claims that discriminatory behaviour will be decreased, or will cease entirely. These strategies combat stigma by challenging the problematic myths with facts: by revealing the problematic knowledge structures to have false elements, or to be entirely incorrect. Examples of this kind of approach can be seen throughout many different anti-stigma initiatives. For instance, the Time to Change project (2017) explicitly contrasts common myths about mental illness with

the facts[4]. For instance, the myth 'people with mental illnesses can't work' is countered with 'most of us will work with someone who has had mental health issues'. Thus, the problematic stereotype is shown to be (either partially or entirely) false. Given that people don't generally endorse falsehoods, the idea is that the stereotype is therefore less likely to be accepted by the public. Awareness of stereotype does not develop into prejudice, and come to manifest itself as discrimination: thus, the stigma process is halted.

Education strategies are intuitively appealing. It is intuitive (and reassuring) to think that much of the stigma surrounding mental illness results from misinformation and misconception, and to think that if these errors were to be addressed, stigma would cease. Indeed, the education strategy makes several significant assumptions about the reasons for stigma, and the ways in which people form and revise beliefs. The first is that most people do not practice stigma because they are cruel, or because they feel straightforward dislike or antipathy towards the mentally ill. Rather, stigma occurs because people are misinformed, or have erred in their choice of knowledge structure. The second broad assumption concerns our motivational structures, and the ways in which they are revised. Specifically, the education strategy assumes that we form beliefs and act on the basis of evidence and information. It assumes that if we were presented with evidence that ran contrary to our belief, we would revise it (or at least be willing to), and so act differently: our beliefs and the behaviours based upon them are responsive to evidence. As such, a change in the available body of evidence (either the discrediting of existing evidence, or the introduction of new information) should bring us to revise the belief. Indeed, having beliefs which are evidence-responsive is often understood to be part of what it is to be rational.

[4] See http://www.time-to-change.org.uk/mental-health-statistics-facts

These assumptions- namely, that people mostly act badly because they are misinformed, and that our beliefs are evidence-responsive- form the rationale behind education strategies to combat the stigma of mental illness. As Corrigan (2016) notes, education has been a particularly popular strategy, particularly I think because the assumptions on which it rests reflect our common-sense notions about our motivational structures, and are already plausibly a part of our folk psychology. However, education strategies have been thought to have several limitations. Evidence suggests that education does not always translate into a significant change in behaviour (Corrigan 1998, Corrigan 2004, Rüsch et. al 2005). Furthermore, educational programmes "tend to reach those who already agree with the message" (Rüsch et. al 2005), whilst generally stereotypes have proven to be quite resistant (Corrigan 1998) and the limited benefits of education do not appear to be enduring (Corrigan et. al 2002). As I will explore in chapter 2, traditional education strategies may be of limited use in combatting implicit stigma and implicit bias- due, in part, because recent work has suggested that we sometimes act in biased ways where we have all the information, and where the action goes against our explicitly avowed beliefs.

ii. Contact

Contact strategies seek to promote contact between the public and people with a mental illness label: often those who are in recovery. The thought is that when confronted with a person with a mental illness label who does not meet the expectation set by the stereotype (for instance, someone with a mental illness who holds down a job and contradicts the stereotype 'people with mental illnesses cannot hold down jobs'), an individual may revise her endorsement of

that knowledge structure. The stereotype is challenged as she is presented with an individual who does not behave in the manner the stereotype dictates. Consequently, the individual may come to think that the stereotype is not accurate or useful to her, and thus abandon it or give it less credence. She will hopefully gain new knowledge structures or beliefs which are more positive because of the interaction: for instance, 'people with mental illnesses can hold down a job'. Crucially, these will replace the problematic stereotypes. Contact is often used in conjunction with education strategies. Just as direct refutation dismantles a stereotype, contact with someone with a mental illness can undermine knowledge structures or create new ones.

There are certain factors which affect the success of a contact strategy. It is thought that contact is a particularly effective method of stigma reduction where the following conditions obtain: "equal status among participants, a cooperative interaction as well as institutional support for the contact initiative" (Rüsch et. al 2005, p.536). Further to this, *in vivo* contact has been demonstrably more effective in reducing stigma than contact across other mediums (Corrigan et. al 2012). That is, meeting someone in real life is better than hearing them narrate their story on a charity's website, or seeing a sympathetic depiction of a mentally ill character in a film or on TV.

Trials have suggested that contact strategies have been generally more successful in the reduction of stigma than education-style approaches (Corrigan 2016). This seems intuitively plausible: seeing evidence of something for oneself is likely to be more persuasive than merely being told about it. However, this strategy also has its drawbacks. For instance, research has shown that one must be careful in the selection of contact figure. If the person one encounters does not appear to fulfil any of the expectations set by the stereotype, then research suggests

that the stereotype itself will not be undermined if the contact figure does not demonstrate stereotypical behaviour (Kunda & Oleson 1995). That is, if the contact was mentally ill, yet incredibly beautiful and successful, she may be taken as unusual, and not considered to be a member of the group about which the stereotype speaks. In this sense, she will not disconfirm the stereotype: she will merely be understood as exceptional. As Rüsch et. al (2005, p.536) note, she may be reclassified as 'us' rather than being representative of 'them', and so the initial stereotype remains intact: she is 'the exception that proves the rule' rather than a reason to be generally sceptical of the knowledge structure as a whole.

Thus, if the wrong contact figure is chosen, contact strategies may inadvertently end up corroborating stigma (Rüsch et. al 2005, p.536). Further to this, there is some considerable pragmatic challenge posed by rolling out a contact programme to the public. Although *in vivo* methodologies have been found to be most effective, they would also be the slowest and most difficult to implement. Time must be set aside for real engagement between members of the public and the contact figure. Furthermore, the environment must be right, and it must be possible to get members of the public to attend. One worries that in this sense, contact strategies may also be afflicted by one of the main challenges faced by education strategies: that the programmes only reach those who are already on-board with the message. For instance, if one sincerely believes that the mentally ill are dangerous, one is hardly likely to consent to meet with them, or sanction a contact programme to be implemented at one's child's school. Indeed, it is worth noting that the degree to which individuals will be willing to meet the stereotyped group is likely to vary in accordance with the content of the stereotype.

iii. Protest

These strategies are mainly concerned with challenging negative depictions of mental illness in the media and critiquing the stereotypes upon which they are based. Protest methodologies hope to send a message to both the media and to the public respectively: stop presenting mental illness in such a problematic way, and stop believing it (Corrigan 1998). They differ from the aforementioned methodologies in that they are fundamentally *reactive* rather than *proactive.* That is, they are concerned with rejecting problematic depictions of mental illness, but do not themselves offer up any alternative knowledge structures or information. For instance, they may function by protesting a film which contains a negative stereotypical depiction of a mentally ill character, with the aim of getting the film banned or the depiction modified. However, they do not then offer up more positive knowledge structures or beliefs based upon facts.

Interestingly, Corrigan (2016, no pagination) argues that protest strategies "rely upon an appeal to moral authority". That is, they appeal against what they perceive to be morally problematic depictions, stereotypes, attitudes or behaviours. They 'assume the moral high-ground', with the success of the strategy dependent upon depicting the stigmatizing party's views or behaviour as morally problematic: thereby placing an impetus upon the stigmatizer to change. Protest tends to be a relatively labour-intensive strategy. Constant vigilance is required to monitor the media for problematic characterisations, and then further energy must be expended in the protest and complaint process. Protest is also carried out on a case-by-case basis in many instances, thus adding to the labour required. However, one might hope that this process may have a profound cumulative effect in reducing the presence of problematic depictions, and thus

limiting opportunities for stereotypes to be acquired (specifically, socially-given stereotypes which originate from media depictions).

Whilst there is evidence to suggest that protest strategies have been successful in decreasing the number of problematic representations of mental illness in the media, as Rüsch et. al (2005, p.535) note, there is little known about whether these methodologies have been successful in diminishing prejudice. Indeed, removing a problematic depiction of mental illness may do little to get rid of the stereotype on which it is based. For instance, the knowledge structure may be gained in many other ways. However, research *has* indicated that protest is successful in getting people to suppress their problematic thoughts, or discriminatory behaviour- in part because the public do not act upon their attitudes for fear of reprisal (Rüsch et. al 2005, Corrigan et. al 2005).

However, it is not clear that suppressing thoughts or behaviour tackles prejudice at its root. Indeed, suppression can lead to a 'rebound' effect, in which those asked to suppress problematic thoughts actually end up having *more* of them (Rüsch et. al 2005). Research also shows that that people do not always take respond well to being asked to suppress their thoughts, and offer psychological resistance (i.e. 'don't tell me what to think' (Corrigan et. al 2005, p.184). In some cases, attitudes worsen as a result. In short, whilst protest strategies effectively remove problematic depictions of mental illness from the public sphere and may effectively 'shame' some into suppressing their thoughts, suppression itself can be problematic: it may worsen the situation, and prove to have little or no impact upon prejudice.

V. Conclusion

In the above, I have outlined what stigma is, and the numerous harms it generates for those affected by it. I hope to have demonstrated why the stigma of mental illness is a worthy subject of investigation, and to have outlined my project and the approach I will take: namely, applying areas of philosophical discussion to the subject of the stigma of mental illness. Each chapter will apply different tools of philosophical analysis to this subject. Throughout this thesis, I will argue that the complexity of the subject matter and the prevalence of 'accurate' stereotypes attached to it means that our treatment of mental illness stigma must differ to our approaches for ending the stigma surrounding gender and race.

CHAPTER TWO

Mental Illness, Stigma and Implicit Bias

I. Introduction

In this chapter, I will explore what recent work in implicit bias might be able to tell us about the stigma of mental illness and the success of our efforts to combat it. On the first point, I suggest that implicit bias alerts us to the fact that stigma, and the processes which constitute it, need not always occur explicitly- that is, at the level of conscious awareness. Rather, we can enact stigma, prejudice or discrimination even where we are not aware of our doing so, and indeed, even where doing so is contrary to our sincere avowed beliefs. In what follows, I will outline what implicit bias is and provide an operational account of what it might look like. I suggest that there is reason to suppose that biases such as those identified in race and gender also affect mental illnesses, although I argue that much more research need be done here. Nevertheless, if implicit biases do arise in regard to mental illness (a hypothetical for which there is some support), then this will have consequences for our anti-stigma strategies.

Thus, on the second point, I argue that our current education strategies for ending the stigma of mental illness (usually conceived of as the provision of accurate information and the debunking of inaccuracies or myths) are likely to be of limited use when tackling implicit bias,

due to the characteristics of implicit bias itself. Whilst I do not mean to suggest that education strategies should be discontinued, I argue that a change of tack may be needed in order to target these more elusive causes of discrimination. Education about implicit bias and implicit social cognition is suggested as a suitable candidate, although I stress that there are some ethical considerations to be attended to if this is attempted. I then go on to argue that education programmes of this kind would be best delivered by identifying target groups. I put forward a dual condition for identifying target groups, in which the groups targeted should firstly, have the power and ability to make significant decisions for and about mentally ill people, and secondly, should exist in working or social environments which encourage reliance on 'type-1' fast, automatic processes, and thus introduce a risk of implicit discrimination. I then identify instances of such groups, before closing by making a few further remarks as to how we should go about tackling implicit stigma about mental illness.

II. The Challenge of Implicit Social Cognition

In the introduction, I suggested that new research in implicit social cognition alerts us to the fact that stigma, and the processes which constitute it, need not occur at the level of conscious awareness. Indeed, the stigma mechanism can be fulfilled and carried out even where we are not aware of doing so, or even where doing so contradicts our other explicitly avowed beliefs. This phenomenon has already been identified for mental illness. Indeed, it is often referred to as implicit/ automatic stigma, and is listed alongside other variants such as public stigma, self-stigma and label avoidance Yet, despite being recognised, implicit stigma is seldom the topic of much discussion. Indeed, it is often identified as site upon which further research need be conducted (for instance, as in Sheehan et. al 2017, p. 51). This chapter proposes to do just that.

Just as stigma is the simultaneous operation of the processes outlined in chapter 1, implicit stigma describes where this occurs in a manner beyond conscious detection. That is, where we are not aware that we are stigmatizing, nor of affording others reduced status and so on. The stigma process is made up of a variety of different attitudes, which can be taken together to form the basis of behaviour. In this chapter, I will explore what recent work about implicit attitudes and implicit bias can tell us about the kind of implicit views we might come to hold about mental illness, and how they might guide our behaviour. At the outset, it would be useful to delineate these concepts a little more fully (an even fuller analysis of implicit bias will follow shortly), and their relationships to one another.

Implicit bias is best described as a kind of implicit attitude. In turn, implicit attitudes are most easily explained by contrasting them with explicit attitudes. Explicit attitudes are thought to be those that are under conscious control, whereas implicit attitudes are not: generally, it is thought that implicit attitudes are not consciously controlled, nor indeed are they the kinds of things of which we are consciously aware (Brener et.al 2013). There is a growing body of literature which suggests that most people possess certain implicit attitudes which fall beyond their conscious control (for a review, see Brownstein & Saul 2016). These 'implicit attitudes' often form the basis of implicit biases, with implicit bias being defined as "a term of art referring to relatively unconscious and relatively automatic features of prejudiced judgment and social behaviour" (Brownstein 2016). The relationship between implicit bias and stigma is best described by Goguen (2016, p.232), who observes that one can conceive of stigma as a coin which has two sides: implicit bias and social identity threat. She argues that "implicit bias describes that devaluation from the perspective of the actor of that devaluation" whereas social identity threat "describes devaluation of a social group from the perspective of the target of that devaluation (Goguen 2016, p.233). Thus, implicit attitudes form the basis of implicit

biases, which in turn partially constitute stigma, in conjunction with social identity threat (a variant of which will be explored in chapter 3).

i. Knowledge about Implicit Bias and its Manifestations

With the relationships between the concepts outlined somewhat, it is now possible to more fully explore what implicit bias is. As Brownstein observes, the study of 'implicit social cognition' is a relatively new form of enquiry, and has focussed upon the study of implicit attitudes towards a plethora of topics, including consumer products, alcohol and political values. Yet he observes that some of the most interesting (and concerning) work has investigated implicit attitudes towards socially marginalized or stigmatized groups, for instance, African-Americans, women and the LGBT community. Worryingly, evidence suggests that these implicit biases are likely to be incredibly widespread, and may go some way to explaining the persistent social inequalities which pertain in modern society[5]. It should not surprise the reader to hear that people may have thoughts, feelings and attitudes of which they are not consciously aware, and do not verbally report possessing. As Brownstein (2016) and Brownstein & Saul (2016) rightly note, this intuition has been widely shared.

However, new technologies and techniques have made it possible to measure what Brownstein & Saul (2016. p.3) call 'hidden prejudices' scientifically, and so to witness the manifestation of implicit biases through testing. This is done not by asking the participant to report on her

[5] For a good introduction please see Brownstein and Saul (2016).

own self-evaluation, but by utilising 'indirect'[6] measures, the most well-known of which is the Implicit Association Test or IAT (Greenwald et al. 1998). However, there are many more indirect methods that may be used, which, as Brownstein & Saul (2016, p.6) note, are often derivations of sequential priming. These include the Affect Misattribution Procedure (see Payne et. al 2005) and semantic priming (see Banaji & Hardin 1996), amongst others[7]. It is not my primary interest to expand upon these methods in a great deal of detail, hence here I will focus upon the most commonly utilised of these indirect methods- the IAT- and direct the interested reader to the references below.

In the standard IAT, the subject is confronted with a screen and asked to sort words or pictures into the correct category as quickly as possible, and with as few errors as possible. Subjects 'sort' the words or pictures to the left or right. For instance, in the race IAT, subjects are asked to categorize two racial groups (black or white) and two moral attributes (good or bad) as quickly as possible. From this point "differences in response latency (and sometimes differences in error rates) are then treated as a measure of the association between the target group and the target attitude" (Huebner 2016, p.49), with other indirect measures of implicit attitudes such as the Go-No-go Association task (see Nosek & Banaji 2001) often working in a similar way. Tests contain some concept pairs which are *stereotype consistent:* for instance, some image pairings will be consistent with widespread negative attitudes towards black people, and with relatively more positive attitudes towards white people[8]. Yet, some will also

[6] This work will follow De Houwer et. al (2009) and Brownstein & Saul (2016) in utilising the terms 'direct' and 'indirect' to refer to characteristics of measurement techniques, and to use the terms 'implicit' and 'explicit' to refer to "characteristics of the psychological constructs assessed by those techniques" (Brownstein & Saul 2016, p.4). This is done with several provisos, as laid out by Brownstein & Saul: namely, that the terms 'direct' and 'indirect' can also refer to different kinds of explicit measures, and that the distinction between direct and indirect measures is not an absolute one, but rather, relative.

[7] For a more in-depth review, see Brownstein & Saul (2016), Nosek & Banaji (2001) and Nosek et. al (2011).

[8] For examples of different IAT tests, see 'Project Implicit': a collaborative online exercise founded by three scientists three scientists – Tony Greenwald (University of Washington), Mahzarin Banaji (Harvard University), and Brian Nosek (University of Virginia)- to study implicit social cognition.

be inconsistent with prevalent stereotypes. Most subjects typically respond faster and with fewer errors when presented with stereotype consistent concept pairings, which is thought to demonstrate a stronger association between the two concepts in memory. In contrast, slower or erroneous responses are demonstrative of a weaker association (Nosek et. al 2007).

The IAT has been the subject of several major reviews. Many commentators have been sceptical about the reliability and validity of the test, as we will see later. However, others have argued that the test is reliable, and generally resilient to cases in which subjects intentionally attempt to distort the test (Greenwald et al. 2003, 2009; Nosek et al. 2005, 2007a; Lane et al. 2007). As Brownstein (2016) observes, it can be utilised to predict a range of discriminatory behaviour: indeed, in some cases it is a better measure than self-report. IAT test scores often make for frightening reading: a review conducted by Nosek et. al (2007b) found that of 70,000 subjects taking the race IAT, over 70% of white participants made more errors and were generally slower when presented with stereotype inconsistent concept pairings, suggesting an implicit preference for white over black faces. From this, we may well theorize that that implicit bias is primarily driven by in-group favouritism, however this hypothesis is undermined by the fact that 40% of black participants demonstrate the same preference for white faces (Nosek et. al 2002). This leads Brownstein and Saul (2016, p.6) to remark that "it appears that implicit bias is driven by a combination of in-group favouritism and sensitivity to the value society places on particular groups".

In conjunction with test scores, many commentators argue that implicit biases are also manifested in behaviour. A study conducted by Bertrand et al. (2005) revealed that those with stronger associations between black faces and 'bad' and between white faces and 'good' on the

IAT were more likely to discriminate against 'black-sounding' candidates in the hiring process. Similarly, participants demonstrating a strong 'black-bad' association score on the IAT were found to be more likely to 'shoot' at unarmed black men when engaging in a computer simulation than they were to shoot similarly unarmed white men: thereby demonstrating a pronounced 'shooter bias' (Correll et al. 2002; Glaser & Knowles 2008). Indeed, as Brownstein (2016, no pagination) observes, the IAT is particularly successful in the prediction of non-verbal or 'micro-behaviour' and in cases in which the subject is stressed or under a heavy cognitive load, or where they must make decisions quickly or with incomplete information[9].

As many commentators have noted, the results of IAT tests may go some way to explaining how, despite enormous progression in terms of explicit attitudes towards race and gender in many western nations in recent years, there still exists such persistent inequalities (Banaji & Greenwald 2013). The postulated answer is that many of us are in possession of implicit attitudes which we are often unaware of, or unable to verbalise. These implicit attitudes can conflict with our explicit attitudes- those which we are aware of and can report ourselves as having-, yet they play a substantial role in causing our behaviour. Thus, in the case in race, we might sincerely report ourselves as lacking any bias against black people, yet we may still harbour an implicit bias. As Machery (2016, p.104) puts it:

> When someone sincerely asserts that she does not prefer whites to blacks, she expresses her egalitarian attitudes towards blacks, but this explicit attitude does not exhaust her racial attitudes: she is also likely to have an unconscious negative attitude towards them.

Hence, sincere self-reports of egalitarian attitudes are not enough: we may be in error when reporting what our attitudes are, and ignorant about our possessing implicit attitudes which

[9] See Dovidio et al. (2002); Cortina (2008); Cortina et al. (2011).

guide our behaviour in an unconscious manner. Thus, we may engage in discriminatory behaviour based on our implicit biases, without realizing that there is a problem in the first place, as we take our attitudes towards race to be exhausted by those attitudes we can report ourselves as having. Simply put, the literature on implicit bias suggests that the problem of racial discrimination is a more complex one than we acknowledge, and worryingly, that people with sincerely reported strong egalitarian beliefs may be part of the problem.

Yet, a caveat must be added here. In recent times, there has been some controversy about the IAT in particular. In a recent article in *New York Magazine,* Jesse Singal (2017, no pagination) has suggested that recent scholarly work (see Forscher et. al 2013, Oswald et. al 2013) indicates that the IAT is "a noisy, unreliable measure that correlates far too weakly with any real-world outcomes to be used to predict individuals' behaviour". The IAT, Singal argues, claims to measure implicit bias, but it is not apparent that this claim is true. Bias, if it is to be interesting, or if it is to explain discrimination, must be biased *behaviour*, rather than a cognitive predisposition to think in certain ways. If 'bias' is merely a high IAT score, then we are unlikely to find this troubling. Singal notes that it has repeatedly been claimed that the IAT measures implicit bias in a manner which is relevant to real-world behaviour. Yet, problematically, the IAT has serious issues with both reliability (in terms of measurement error) and validity (which occurs where something actually measures what it purports to). Simply put, recent scholarly work has suggested that in many cases, the IAT is a poor predictor of behaviour. This leads Singal to claim that the current enthusiasm for implicit bias is misplaced, for our knowledge of it is founded upon a psychological test which cannot stand up to scrutiny.

I acknowledge Singal's point here: the IAT has significant limitations. Yet, my project is not seriously undermined in that I suggest that we can still know that implicit bias exists whilst acknowledging that the IAT cannot robustly predict behaviour in all cases. Convincing arguments for this have been offered by a variety of scholars, including Brownstein (2017, no pagination), who acknowledges that "it is clear that general measures of implicit attitudes (e.g., as represented on a race-evaluation IAT) don't predict specific individual behaviour (e.g., biased grading) very well". Yet, he suggests, this should be unsurprising to us, given the challenge posed by predicting behaviour in general, and the fact that "the attitude-behaviour link is highly context- and person-specific" and so "general measures of preferences shouldn't be expected to predict specific behaviours in specific contexts very well" (Brownstein 2017, no pagination). In a similar vein, Frankish (2017, no pagination) argues that whilst "we cannot infer a person's beliefs from their associations between stimuli", this does not mean that we need doubt the existence of implicit bias itself[10]. As such, I take it that Singal is right to warn us of the limitations of the IAT, yet too swift in declaring that implicit bias itself is also suspect. Indeed, I suggest that we have good reason to think that implicit bias exists, and with this in mind, I now turn my attention to sketching out an account of it.

ii. Implicit bias: an account

What is implicit bias exactly? Work in social psychology has often defined implicit bias as occurring outside of the realm of an individual's conscious control (see instance, see Sheehan et. al 2017, p.56). At first glance, this seems relatively unproblematic, and we might think it

[10] For further arguments, see the roundtable discussion at http://philosophyofbrains.com/2017/01/17/how-can-we-measure-implicit-bias-a-brains-blog-roundtable.aspx

intuitive to characterize explicit attitudes as conscious, and implicit attitudes as unconscious. However, these terms of art can be refined further so as to reveal more about the nature of implicit bias. This will be done in this section. Yet, two qualifications must be made. Firstly, the account of implicit bias described here is a minimal one. The empirical investigation into implicit bias is still in its infancy, and much is still being discovered about the character of these biases, what precisely they are, and how they work. The aim of this chapter is to recommend that anti-stigma campaigns for mental illness should take note of the challenge of implicit social cognition, and to offer some preliminary remarks as to how it might inform our anti-stigma campaigns. To this end, it is sufficient for my purposes that the account of implicit bias characterized here is only minimal. Indeed, this is preferable, given that I do not wish any of the claims I make here to rest upon the more controversial or uncertain parts of the literature or research.

Secondly, it is also necessary to outline what the scope of 'implicit bias' is, given that the term has been utilised rather broadly in the literature, and can refer to a wide range of implicit social cognition. This is reflected in Brownstein's (2016, no pagination) definition of implicit bias as "a term of art referring to relatively unconscious and relatively automatic features of prejudiced judgment and social behaviour". Likewise, Saul (2013, p.40) defines implicit biases as "unconscious biases that affect the way we perceive, evaluate, or interact with people from the groups' that our biases target". However, as Holroyd & Sweetman (2016, p.80) rightly note, these 'functional' definitions- in which implicit biases are whatever nonconscious processes influence perception, judgement and behaviour- do not specify which processes precisely constitute implicit bias, and whether we are being confronted with "a singular entity or a range of psychological processes". Indeed, they go on to observe that the term 'implicit bias' has also been utilised to pick out an even wider range of social cognitions, including the unconscious

activation of stereotypes and the resulting affective responses of threat and/or anxiety (stereotype threat).

Whilst these authors recognise the utility and rationale of expansive definitions of implicit bias, they warn that there is considerable heterogeneity within implicit bias, both in terms of the particular associations and their effects on behaviour (see study by Amodio & Devine 2006). As such, they warn against making broad generalizations about implicit bias (see Holroyd & Sweetman 2016, pp.84-92). I think this to be a prudent observation. The account of implicit bias offered here will be in line with the expansive definitions suggested by Brownstein and Saul. I will however, consider stereotype threat (and the broader category of social identity threat) to be distinct from implicit bias. Yet, whilst the account offered below will be expansive, I will take note to avoid the kind of sweeping generalizations about which Holroyd & Sweetman are concerned. Indeed, this will be particularly important, given the relatively high degree of heterogeneity in the category studied. That is, whilst it is certainly true that race and gender categories do not form a unified kind, this is likely even more true for mental illness: a category which is itself divided into many diverse sub-categories of different illnesses, which can be hugely divergent in terms of their etiology and symptomology.

I will now turn to constructing a minimal account of implicit bias. The empirical literature has revealed that we possess many implicit associations between mental constructs such as ‘scientist’, ‘dangerous’, ‘male’, ‘female’, ‘good’ and ‘bad’ etc. Often, these associations pertain between social groups and constructs or roles such as those listed above, and are thought to be generated or strengthened by prominent cultural stereotypes, which themselves link (often undesirable) traits to categories. Where these associations pertain, they may generate biases:

broadly defined as an inclination or prejudice towards or against a group. There are many cases in which biases are likely to be unproblematic (see Saul 2013), yet in this work I will be concerned with harmful biases: those which contribute to discriminatory behaviour. Biases are problematic where we are disposed to "judge others according to a stereotyped conception of their social group (ethnic, gender class and so on)", thus deviating from either a social or rational norm of fairness[11] (Frankish 2016, p.24). Crudely speaking, where one is conscious of one's prejudiced disposition, one's bias is explicit, whereas if one is unconscious of it, the bias is implicit.

But what do we mean in saying that an implicit bias is 'unconscious'? At this point, it will be useful to refer to an example offered by Schwitzgebel (2010, p.532): that of Juliet the implicit racist. Juliet is a philosophy professor working at an American university. She is Caucasian, and after acquainting herself with the literature regarding intelligence and race she has come to believe that there is an equality of intelligence between the races: a viewpoint which coheres with her other egalitarian values and liberal outlook. She can argue for this point with coherence, sincerity and vehemence. Yet, she fails to manifest these egalitarian values in her behaviour:

> And yet Juliet is systematically racist in most of her spontaneous reactions, her unguarded behaviour, and her judgments about particular cases. When she gazes out on class the first day of each term, she can't help but think that some students look brighter than others – and to her, the black students never look bright. When a black student makes an insightful comment or submits an excellent essay, she feels more surprise than she would were a white or Asian student

[11] As the common social norm is to treat all people as equal, this man demonstrates a bias in that he deviates from this.

> to do so, even though her black students make insightful comments and submit excellent essays at the same rate as do the others.
>
> (Schwitzgebel 2010, p.532).

Indeed, her bias is not limited to her interactions with students: when placed on a hiring committee, it will likewise appear to her that the 'black sounding' candidates just don't seem as intellectual as white candidates. To convince her of a black candidate's intellectual prowess, it takes far more evidence than it would have done for a white candidate.

Juliet's case is an interesting one, which is useful for characterizing implicit bias more precisely. The first thing we might extract from Juliet's case concerns the distinction between implicit and explicit bias. As noted earlier, one intuitive way of drawing this distinction is along the conscious/unconscious boundary, with implicit biases being unconscious and explicit biases conscious. However, there is reason to think that this is not an adequate means of distinguishing the two. As Schwitzgebel points out, Juliet could come to find out that she is biased against black students, perhaps by taking one of the indirect tests outlined above, paying attention to the reports of her colleagues, or through examination of her own behaviour. Should this be the case, it seems strange to claim that Juliet is not conscious of her bias. This possibility leads Frankish (2016, p.25) to claim that it would be erroneous to draw the boundary between explicit and implicit biases in terms of the conscious/unconscious distinction: after all, Juliet may know *that* she is biased even if she does not endorse the prejudice, and thus there is a sense in which we might want to say that she is consciously aware of the bias.

Instead, he argues that we might better say that the mental state underpinning Juliet's biases is not *introspectable.* That is, in contrast to her explicitly held beliefs, which she can

straightforwardly report possessing, Juliet cannot *directly* know that she is implicitly biased. Indeed, she becomes aware of the 'biased' mental state only by observing its manifestations on her own behaviour, or by being informed of it by others. Hence, Juliet's self-reports about her implicit attitudes, whilst sincere, are unlikely to tally up with her behaviour. Thus, on this characterization at least, implicit biases are those which are underpinned by mental states which are not introspectable, whereas the mental states upon which explicit beliefs are founded are accessible to introspective access: more similar in nature to our explicitly held or avowed beliefs.

Frankish (2016) acknowledges that even this characterization may be too strong, given that there is evidence to suggest that some aspects of implicit attitudes are introspectable (see Gawronski et. al 2006, Hahn et. al 2014). However, it is sufficient for my purposes to claim that implicit biases are often, although not always, characterised by what Washington and Kelly (2016, p.16) term ***introspective opacity***[12]: in which people who are biased exhibit tendencies "whose presence and influence on thought and behaviour is not easily detectable via introspection". It is worth noting that the fact that many implicit biases are characterised by introspective opacity is consistent with evidence that suggests that subjects are sometimes aware of the cognitive processes being measured on IAT tasks (see De Houwer 2006). In this case, the subjects may well be conscious of what is being measured (this may be inferred by

[12] Washington & Kelly are careful to note that the four 'textbook features' of implicit bias which they list describe implicit racial biases. However, as they note, there is reason to suspect that these features generalize in a straightforward manner to other kinds of implicit bias. Indeed, entries throughout the two volumes on implicit bias and philosophy edited by Brownstein and Saul (2016) appear to refer to the same kinds of characteristics to describe implicit bias, and the 'standard' view characterized in the introductions to the two volumes has much in common with Washington & Kelly's account. As such, I propose to utilise Washington & Kelly's terms to describe implicit bias in the general sense. However, I do so with the qualification that these terms are meant only to introduce a minimal account for the purposes of beginning the investigation. Further research may reveal that some of the characteristics of racial biases are not applicable to the case of mental health. However, in the absence of this empirical data, I think the minimal account a good place to begin the discussion.

looking carefully at the concept pairings), but they may still not have direct introspective access to their own associative tendencies or biases.

Juliet's case suggests another interesting feature of implicit bias. Frankish (2016, p.25) observes that whilst Juliet's bias primarily makes itself known in her unreflective (not consciously controlled) behaviour[13], it may also make itself manifest in her reflective behaviour:

> it affects her conscious judgements, decisions, and feelings, and Juliet may consciously perform actions that display it- for example, consciously disciplining herself to do her grading. That is, implicitly biased actions may be consciously intended, although they are not consciously intended *to be biased*.

Thus, once again, it appears that the distinction between explicit and implicit biases cannot be characterized in terms of the former being conscious and the latter unconscious: there is a sense in which implicit biases interact with, or manifest themselves in, conscious processes. Once again, this leads Frankish to consider that a better way of phrasing precisely why it is that Juliet's bias is nonconscious would be to claim that it is not something which she consciously endorses in her reasoning and decision making. Simply put:

> although Juliet may be conscious of behaving and judging as if there are racial differences in intelligence, she does not consciously think *that* there are racial differences in intelligence. If that thought occurs to her, she rejects it. This is compatible with her having some introspective awareness of her bias, provided she does not endorse it.

[13] It is beyond the scope of this work to provide an argument for the existence of unconscious behaviour, although I acknowledge that to claim that there is such a thing is controversial, and goes against several traditions: some of them philosophical, and others not. I direct the interested reader to Frankish & Evans (2009), and acknowledge in line with Frankish (2016) that whilst such traditions exist, there is also a long tradition of theorizing about nonconscious mentality.

Frankish 2016, pp.25-26.

Hence, on this reckoning, implicit bias affects both our unreflective and (to some extent) reflective behaviour, whereas explicit biases affect only the latter. Furthermore, where the content of a bias is endorsed, the bias is explicit, whereas the bias is implicit if it is not so. The point about endorsement is often raised in the literature. In a similar vein, Washington & Kelly (2016, p.17) claim that implicit bias often involves ***dissociation***. That is, in a single person implicit biases may coexist with near-antithetical attitudes such as commitment to egalitarian values. For instance, Juliet may sincerely hold egalitarian values which she reflectively endorses, yet still demonstrate bias towards black students.

What is also evident from an examination of Juliet's case, and indeed from the studies listed in the section above, is that implicit biases can be ***manifested in behaviour***: worryingly, research suggests that the range of behaviours influenced by implicit biases may be substantive indeed. As Schwitzgebel notes, Juliet's responses are subtle, yet may influence her behaviour in profound ways- for instance, by leading her to mark down black students' essays, or failing to invite black candidates to interview. Indeed, as Washington & Kelly (2016, p.18) observe, whilst implicit biases have been demonstrated to influence decisions made under time pressure (for instance, in the case of shooter bias, or in the IAT), research suggests that they also:

> influence more deliberate, temporally extended decision making, despite confidence that in such cases behaviour is more likely to reflect explicit attitudes and considered views. Examples include what diagnosis or type of health care a medical patient should get (Blair et. al 2011), who should or should not serve on a jury (Haney-López 2000) …

Finally, racial biases are often characterised by what Washington and Kelly (2016, p. 18) term ***recalcitrance***: they are relatively easy to acquire, yet hard to suppress in our judgements and behaviours, and difficult to control. The challenge of ridding oneself of implicit bias is likely made all the more difficult by their introspective opacity: individuals may be unaware that they are biased and that there is a problem to be addressed. To make matters worse, whilst many commentators suggest that there are ways in which we may bring implicit bias under control (see Frankish 2016, Huebner 2016, Madva 2016), doing so is not likely to be an easy process. Indeed, these authors all note that the rectification of implicit bias is likely to consume a great deal of mental resources, whilst being time-consuming in that it requires sustained and focussed attention. If not conducted in the proper manner, attempts to eradicate implicit bias can even serve to reinforce pernicious associations, and so compound the problem. I will say more on this aspect of implicit bias in the following sections.

The above (in conjunction with other findings) has led some commentators to form accounts of the structure of the mind to incorporate some of the recent empirical results. Indeed, it has been thought to be particularly noteworthy that despite the dissociative tendencies of implicit biases, they remain recalcitrant. That is, even where we acknowledge that our behaviour manifests a bias that we find to be discordant with our explicitly held beliefs, these biases remain hugely difficult to rid oneself of. Even if Juliet sincerely believes that race is *not* a good indicator of intelligence, she struggles to rid herself of the inclination that it *is*. As Huebner (2016, p.48) observes, there appears to be a tension here: one which holds between our rational processing and the beliefs they generate and our automatic or reflex-like processes. This tension has motivated many commentators to postulate a dual-system architecture of the mind, in which there are two broad types of cognitive systems through which responses can be generated:

> One (type 1) that is fast, automatic, nonconscious, and undemanding of working memory, and another (type 2) that is slow, controlled, conscious, and demanding of working memory. Type 1 processes are also variously described as associative, parallel, heavily contextualized, heuristic, and biased, and type 2 processes as rule-based, serial, decontextualized, analytical, and normative.
>
> Frankish (2016, p.35)

Type 1 processes generate "attitudes and beliefs that are acquired passively without individuals' awareness and that influence subsequent judgements, decisions and actions without intention or volition" (Dasgupta 2013, p.235). They utilise slow-learning memory systems which are thought to respond to experience and social conditioning- one form of which is thought to be the associations we are exposed to, both in the news, media, and in art. By contrast, type 2 processes form a "slow, controlled, inferential *system* that produces reflectively endorsed beliefs" (Huebner 2016, p.48). As Brownstein and Saul (2016, p.10) observe, type 2 processes possess a memory system which is "capable of one-shot learning in response to explicit tuition".

From this, it is tempting to identify implicit attitudes (and thus implicit biases) as arising from type 1 processes, whereas our explicitly held beliefs are generated by type 2 systems. We might think that our reflective and considered beliefs are the result of a system which is slow, inference-based and controlled, whereas implicit attitudes and biases are generated by type 1 processes with associative and often affective content. Indeed, much research in social psychology is based upon the notion that implicit biases may largely originate and be propagated by associative mechanisms (for details, see Huebner 2016, p.49 and Dasgupta 2013). However, this kind of strict distinction does not accurately reflect empirical data. Indeed, as Huebner (2016, pp.50-51) observes: "recent data have revealed that inferential

reasoning and one-shot learning can sometimes affect implicit attitudes", which "would be surprising if such attitudes were implemented exclusively by associative systems". Thus, whilst many implicit biases may arise from associative systems, they need not arise through these systems *exclusively*. To this we might stress that the vast majority of dual-system theorists allow that implicit biases need not be generated by one mechanism alone. Indeed, many posit the existence of numerous automatic systems, all or some of which may operate in parallel.

Whilst the dual system hypothesis is not the only way of conceiving of the architecture of the mind, it is relatively widely embraced in the literature. Indeed, as Mallon (2016, p.130) observes, this is in line with a general move towards dual-process theories in both psychology and philosophy. Whilst it is worth stressing that implicit biases cannot be simplistically identified with type 1 processes alone, the dual-system model is useful for delineating some of the distinctions between more automatic and reason-based attitudes and occurs in much of the literature about implicit biases. As such, I think it a useful inclusion in the minimal account. However, it is worth noting that this is only a broad sketch of dual-process theory, more specific versions of which have been proposed (for example, see Frankish 2016, Huebner 2016).

III. Implicit Bias and Mental Illness

From this point, we can draw together a brief summary. The term 'implicit bias' is used to refer to a wide range of prejudiced social judgements, which can be made manifest in behaviour and glimpsed through the mechanism of indirect testing. These judgements are often fast, automatic, and bypass the level of conscious awareness. Implicit biases have been argued to be manifested in a wide range of behaviour, and often characterised as being dissociative,

recalcitrant and opaque to introspection: at least in the case of racial biases (Washington & Kelly 2016). We are now in a position to see how this new literature might inform the understanding of mental illness stigma.

At the outset, it is notable that very little work has been conducted on mental illness and implicit bias to this date. The lack of literature on this topic is perhaps unsurprising. For one, research on implicit racial and gender biases is itself very new. Indeed, interest has spiked in the last ten years, which we might note is decades after the civil rights and gender equality movements first made efforts to attack explicit stigma and discrimination. Similarly, we might expect that research into mental illness stigma will take time. Indeed, given that the field of modern psychiatry is itself only roughly 60 years old, and it is only in recent memory that it was even considered problematic to hold stigmatizing attitudes towards the mentally ill, it should not surprise us that mental health lags behind gender and race. Yet, some work has been done, which suggests that such biases do exist in the domain of mental illness, and that said biases can be associated with certain types of behaviour.

i. Empirical Evidence for Implicit Bias in Mental Illness

For the project outlined in this chapter to be interesting, one must establish that there is good reason to suppose that implicit biases are likely to be directed towards mental illness as they have been proved to be towards categories such as race, gender, disability, sexuality and religious or ethnic grouping (for references, see Brownstein 2016). I suggest that there is good reason to suppose this. Indeed, given that implicit biases have been identified which target a plethora of 'socially undesirable' categories, we may have reason to think that they are likely

to also target mental illness as another instance of a category which has been socially maligned and to which there are numerous negative stereotypes attached. Indeed, given that biases have been identified in regard to so many categories, we may think it the most rational response to proceed in any case on the assumption that there are very likely to be analogous biases for mental illness. Indeed, we may even consider it morally irresponsible to fail to pursue this possibility, given the myriad harms the stigma of mental illness produces, and the impact it has upon mental health service users, their families and friends, medical professionals, care workers and so on.

However, we need not proceed on this assumption alone. Indeed, there appears to be a growing recognition within the literature that it would be prudent to move away from measuring mental illness stigma in terms of external attitudinal measures such as self-report, and instead focus upon implicit or unconscious indicators (see Stier & Hinshaw 2007). Further data has been collected, mostly focussing on healthcare workers. When surveyed by Brener et al (2013), 74 mental healthcare workers demonstrated somewhat positive explicit attitudes in regard to their clients with mental illness. They were also asked to report on their affective responses towards their patients, and their desire to help them. When the same practitioners took the IAT test implicit biases were identified and their attitudes towards the patients were found to be somewhat negative, despite the positive explicit self-reports. Brener et. al found that whilst explicit and implicit attitudes predicted negative emotional responses, only implicit attitudes predicted a decrease in the intent to help. They took this to be evidence of the link between implicit attitudes and behavioural intention, and observed that their data reinforced the need to attend to the impact of implicit attitudes towards service users by mental healthcare workers.

A further study has been conducted by Stull et. al (2013), and concerned assertive community treatment (ACT) practitioners. ACT is concerned with the intensive management and coordination of care for patients with severe mental illnesses. Patients are often involuntarily committed, and practitioners have control over the patient's finances and monitor medication consumption. Stull et. al (2013) found that the presence of implicit bias could be utilised to predict a greater degree of support for the more restrictive elements of ACT treatment. That is, those practitioners who displayed a higher level of implicit bias were more likely to agree with the need to control aspects of their patients' lives.

Implicit bias has been linked with the tendency towards over-diagnosis in both clinical professionals and graduate students (Peris et. al 2008). Yet, the same study also found that those with a higher degree of training in mental health demonstrated more positive implicit and explicit attitudes towards those with mental illnesses. However, it is not clear whether this improvement was the result of education or of exposure. Similarly, a study by Kopera et. al (2015) found that when compared to non-professionals (in this case, medical students), mental healthcare professionals reported possessing higher rates of 'approach' emotions, were less likely to discriminate against the mentally ill, and were less restrictive in their attitudes. However, both groups demonstrated ambivalent and negative implicit attitudes towards the mentally ill: implicit attitudes that were not necessarily modified following long-term professional contact with those suffering from mental health issues.

Whilst the majority of the research to this point has focussed on the implicit biases of healthcare workers, there is reason to suspect that these attitudes may be very prevalent. If it is the case that even people who have a great deal of knowledge about mental illness and contact with

those suffering from them demonstrate implicit bias, then it seems plausible that those who lack both will likely demonstrate more, or stronger, biases. Indeed, given that public stigma is still such a concern, and that explicit attitudes are poor in many cases, it would be very surprising if implicit biases were not held by the public against the mentally ill as a form of automatic or implicit stigma.

Further to this, there is evidence that groups outside healthcare demonstrate implicit bias towards the mentally ill. For instance, there is evidence that whilst crisis intervention training (CIT) has been successful in improving the explicit attitudes of police officers towards those with mental illnesses, negative implicit attitudes continue to be held (Mulay et. al 2016). Research concerning implicit bias and mental illness in the wider public is also being carried out. In 2011, Project Implicit Mental Health (PIMH) was launched. It provides a series of IAT tests related to different areas of mental health. Of most relevance to this project is the Mental Illness IAT, which tests whether participants implicitly believe that people with mental illnesses are dangerous. The other tests tend to focus more on the self and one's relation to one's own mental health.

ii. Implicit Bias, Mental Illness and the Limitations of Education Strategies

Yet, although I suggest there is good reason to think that implicit bias will pose as great a problem for anti-stigma campaigns in mental illness as they appear to for race and gender, much more research needs to be done. Indeed, I suggest that this is the first thing that should be gleaned from the literature on implicit bias. We must not only conduct tests to establish where implicit bias is present, but where it is, we must also attempt to establish its character.

As noted in the material above, implicit biases can take many forms. They may track different associations and thus have different content, and they may have different effects on behaviour (if they have any at all). Thus, I suggest that efforts to end the stigma of mental illness should conduct empirical investigations in order to establish which biases exist, what their content is, and how precisely each affects behaviour. Once this is done, we will be able to then ascertain how each bias is to be combatted, which are the most troubling, and which are worth targeting at all (for instance, if one bias has little effect on behaviour, perhaps it is not imperative that we target it).

Yet, when recommending further investigation, we must also be wary of the caution raised by Holroyd & Sweetman: that we should avoid characterizing implicit biases in too general a manner. The concern here is that this will impose a homogeneity on implicit bias which does not reflect the empirical data, and which may as result skew our normative recommendations for alleviating implicit bias. Any investigation into implicit biases and mental illnesses must pay particular attention to this heterogeneity. Indeed, just as it would be a mistake to conceive of gender as a unified kind, the kind 'mental illness' is fractured indeed. The term can be utilised to refer to an incredibly wide range of disorders, many of which vary in severity and symptomology. Indeed, DSM-V classifies the neurodevelopmental disorders such as Autism as mental illnesses, as it does Schizophrenia, Borderline Personality Disorder and Major Depression. These disorders are hugely divergent in terms of their characteristics, and produce very different impacts on functionality.

Yet, the literature on implicit bias may also serve to illustrate other things. For one, just as for gender and race, the problem of mental illness stigma is much broader than we might have

thought. Stigma is not merely an explicit process, but rather, it and its components can occur below the level of conscious awareness. Indeed, stigma appears to be partly, and perhaps largely, constituted by implicit social cognition and implicit biases. As such, if we hope to tackle stigma, it will not be enough to address its explicit variant: we must also find ways of combatting implicit bias. In what follows, I suggest that traditional education strategies (which involve disseminating accurate information whilst contradicting false information) are likely to be successful in addressing explicit stigma (and the processes which constitute it), but will probably be much less successful in addressing implicit stigma and implicit bias. The reason for this is that the characteristics of some implicit biases as described in section II.i make them resistant to traditional education strategies. As such, whilst education strategies providing correct information and denouncing false information are certainly useful, they cannot be our only means of attempting to eradicate stigma. I will now flesh this out further.

Why might education strategies be less successful at tackling implicit bias? It will be useful to recall Juliet's case:

> *Juliet the implicit racist:* after consuming the relevant literature, Juliet comes to hold the sincere and ardent belief that there are no racial differences in intelligence, yet in her behaviour (e.g. whilst marking) she acts as if there were, and it seems to her that 'black-sounding' students and interviewees are never as intelligent as their white counterparts.

What characteristics does her bias have? It is obviously manifested a wide range of behaviours. Furthermore, as Schwitzgebel (2010) observes, Juliet may come to know about her bias, but she cannot report having it immediately in the manner she can with her other beliefs. In Holroyd & Sweetman's terms, her bias is not introspectable. It is also incongruous with her other

explicitly held beliefs and egalitarian attitudes (dissociation) and difficult to her to eradicate through traditional methods, as her strategies for ridding herself of it seem to fail (recalcitrance). We might usefully contrast Juliet with another case:

> *John the explicit racist.* John is a Caucasian professor working at the same university as Juliet. After reviewing the evidence available to him, John concludes that there are racial differences in intelligence. Specifically, he believes that white people are more intelligent than their black counterparts. He sincerely holds this belief, and is prepared to argue for it. He is less likely to call upon black students in class, thinking that they will not be able to answer his question. Whilst university policy dictates that he marks work anonymously, he is very surprised when he finds out that black students perform well. Similarly, when looking at CVs, he chooses applicants with 'white-sounding' names.

We may well feel great alarm that a man such as John could teach at a university, yet this unlikelihood aside, what characteristics does his bias have? Like Juliet's, it is also manifested in a range of behaviour. Further to this, John appears to have *introspective access* to his bias: he knows he is disposed to think that blacks are less intelligent than whites, and can argue for it when asked to. Further to this, his attitudes seem to form a relatively cogent belief system, or at least, his view about racial intelligence is not obviously in conflict with other explicitly avowed beliefs such as a commitment to egalitarian values. In this way, we might say that they are *non-dissociative*. In short, John exhibits an explicit bias: he is consciously aware of his disposition to treat a category in accordance with a stereotypical conception of that kind.

For individuals such as John, education strategies may well prove effective. This is due to another common feature of explicit attitudes or beliefs: that they are reason-governed or

responsive to evidence. Where implicit biases are characterised by recalcitrance, it is thought that our explicit attitudes can be altered through the mechanism of conscious reasoning. That is, reason-governed explicit attitudes (often aligned with type 2 processes) are thought to be responsive to the presentation of new evidence or the discrediting of evidence which was previously given weight. Thus, they are characterised by evidence-sensitivity or *non-recalcitrance*. Recall what was said of the rationale behind education strategies in chapter 1: that it was hoped that social ills such as stigma were the result of misinformation. On this model, we might hope that John's troubling conclusion (that there are racial differences in intelligence) can be explained by his failing to take in a large enough body of evidence (e.g. not doing enough research), or his reliance upon false or relatively weak evidence (e.g. discredited sources or the testimony of a friend respectively).

To combat the explicit discrimination John carries out, education strategies attempt to add accurate information to John's evidence base and discredit some of the (false) evidence he already has[14]. Ideally, changing the evidence base through education should cause John, by a process of rational reflection and reasoning, to form the antithetical belief to the one he previously held and to come away thinking that there are no racial differences in intelligence. Indeed, as explicit beliefs generally form a cogent system (in that they are non-dissociative), this should present him with the choice to accept the new belief and abandon the old one, or to disregard the new belief and retain his old attitudes. The hope here is that the former will happen, although again this may fail. We might note that the decision to take on board the new

[14] To this another related point can be made: it is *relatively* easy to identify John as a target for anti-stigma interventions and to know that he has troubling views we would rather he didn't. For instance, John knows he is biased, and is able to report as such. Yet, of course this ultimately relies upon his disclosing the bias to them: an act which he may be disinclined to do, in such case that his bias contradicts with the values of his society. Nonetheless, if his bias is introspectively available to himself, at least John may become aware that his values do not seem to add up with those of others, which may itself encourage him to investigate whether he is right.

attitude may necessitate the revision of several or many of his existing attitudes. All things going well, the education strategy would hope that the new attitude then became manifest in non-discriminatory behaviour. Thus, the characteristics of explicit bias, particularly evidence sensitivity, make it a suitable candidate for reformation through education.

Yet, conversely, the characteristics of implicit bias make it not so. First off, Juliet is unable to straightforwardly report herself as having the problematic implicit bias: it is not introspectively accessible. Indeed, to her mind she sincerely holds egalitarian beliefs, and so she is unlikely to perceive herself to require education at all. For much the same reason, given that Juliet does appear to espouse genuine egalitarian beliefs, it is unlikely that those around her will think that she has problematic biases, and so would be a good candidate for education strategies. Thus, due to their introspective opacity, implicit biases may create an initial problem in identifying who exactly has them, and so who needs to be educated better.

However, implicit biases also pose a challenge in alteration. Juliet explicitly believes that there are no racial differences in intelligence: a belief she has formed after much thought, and crucially, after reviewing the relevant evidence. Let us suppose that, unlike John, Juliet has already consumed the kind of evidence that the education strategy would disseminate to her: indeed, that she has considered this evidence is the very reason why she has formed the explicit belief that there are no racial differences in intelligence. Yet, her implicit bias remains. Indeed, this is part of the puzzle: Juliet already has access to all the correct information, and so it is not clear what more the education strategy could provide. Juliet already has the very same explicit beliefs the education strategy would aim to induce in her.

Thus, Juliet's problem is clearly not that she lacks enough accurate information. Rather, something else is going wrong: her implicit biases, which conflict with her explicit beliefs, also guide her cognition and behaviour. These biases occur even though Juliet has the correct information, and has formed an egalitarian explicit belief. Using the dual-system hypothesis, we might claim that education strategies may be better at targeting beliefs produced by a type 2 controlled inferential system, whereas implicit biases are more likely to arise through fast, automatic type 1 processes, which rely on association and social conditioning more than reasons and inferences. In this way, Juliet *knows* that there are no racial differences in intelligence, but when she experiences high cognitive load, stress, or she is placed in a situation in which the associations are primed (see Frankish 2016; Huebner 2016), she may be likely to act on her implicit biases rather than on her explicitly held reasoned beliefs. Further, perhaps education strategies are suitable for tackling mental states which are reason-based and evidentially sensitive: something more akin to a belief than a heterogeneous series of fast, automatic processes. Many implicit biases will not meet this specification, and thus changing behaviour may require targeting a heterogeneous system of type 1 processes, for which education may be suitable in *some* cases, yet likely not in others (for example, where the bias is triggered by association). It is likely that a plethora of strategies will be needed to tackle implicit biases. Education alone will not be enough. Indeed, insofar as biases are not sensitive to changes in evidence, they will likely be recalcitrant to such methods.

Thus, what implicit bias reveals to us is that stigma can occur implicitly as well as explicitly. In the case of race, it seems to be that education strategies may be well-suited to combatting explicit biases such as those belonging to John, yet the features of implicit bias make it such that the same strategies may not prove that useful in tackling them. I suggest that this will be a valuable lesson to extrapolate to the case of mental health. I acknowledged that the account of

implicit bias I outlined was a minimal one, and that it was based upon findings which primarily concern race and gender. Indeed, I have highlighted that more research needs to be done about mental illness and implicit bias specifically.

Yet, there is reason to suspect that implicit biases about mental illnesses do exist, and that they might be quite prevalent. If these biases have the same, or at least similar, characteristics to racial biases, then this may well give us warning that education strategies to end the stigma of mental illness are unlikely to be successful in targeting all forms of stigma and its components: indeed, I have suggested that education will likely not be suitable for tackling all instances of implicit bias. Given that public education about mental illness is not as good as it could be, and that many myths still exist, education strategies will obviously continue to be important. Yet, I think it advisable that anti-stigma campaigns for mental illness take note of the new research in race and gender, and consider the likely possibility that great many implicit processes also contribute the stigma of mental illness. As such, we should be alert to the possibility that education cannot be the only strategy we use.

IV. Broadening Education Strategies

But what other strategies should we attempt if we are to tackle implicit stigma in mental illness? I suggest that one useful tactic would be to broaden our existing education strategies. The education programmes I discussed above were concerned with the dissemination of accurate information about mental illness, and the de-bunking of problematic myths or stereotypes. Yet, education can be broader than this. Indeed, I suggest that a plausible first step in combatting implicit stigma in mental illness would be to alert the public that it occurs in the first place.

'Implicit bias' as a term of art is increasingly well-known, yet it is unlikely that the majority of the public is aware of the problem of implicit social cognition in itself, let alone specifically in regard to mental illness. In this way, education could fulfil a vital function- namely, alerting individuals and the broader public that implicit biases exist, and that they may demonstrate them themselves, even where they hold contrary and genuine egalitarian beliefs.

Indeed, I noted earlier that the introspective opacity of implicit biases would likely make it difficult for people to identify themselves as being part of the problem of stigma. Education about implicit bias may well help with this, or at least raise the possibility in the mind of someone like Juliet. Thus, one role for education in anti-stigma campaigns in mental illness would be to inform the public in a broad sense about implicit biases: what they are, how they work, and how they affect behaviour. There may be some initial difficulty in this, particularly given that there is yet much that is not known about implicit bias (particularly in regard to mental health) and what is known is often quite complex, and may prove difficult to disseminate to the general public. Yet, even a stripped-down account would be sufficient to acquaint the public with implicit social cognition, which is the primary task. I take it that informing the public about implicit bias will be valuable, even if not a great deal can be said about the content of implicit biases and mental health (as empirical research into this is still lacking).

Education about implicit social cognition would likely be beneficial to the public at large. In this sense, perhaps it would be suitable to disseminate it in a wide-reaching form such as TV and media. Yet, I suggest that whilst public campaigns are important, there may be some practical and theoretical benefit to identifying particular target groups to whom education of

this kind should be first disseminated, who would benefit most from it, or perhaps that delivering it to them would make the most noticeable dent in tackling implicit stigma. In particular, I suggest that certain individuals or groups may occupy positions of professional or social power such that they can enforce meaningful consequences on those with mental illnesses (and thus, on Link & Phelan's model, are capable of stigma, rather than mere discrimination), and also exist in working and social environments which make it more likely that they will rely on type-1 processes and implicit biases.

The result of this, I suggest, is that certain environments make it more likely that people will discriminate (albeit unintentionally) by relying on implicit processes, and that where this happens, if those people are in a position to make important decisions for and about people with mental illnesses, this implicit and unintentional discrimination may actually produce significant and lasting consequences: qualifying as stigma on Link & Phelan's model. In this way, groups that have power to make big decisions and exist in working and social environments which encourage reliance on type-1 processes are at an increased risk of carrying out implicit stigma. For this reason, I suggest that these groups would particularly benefit from education about implicit social cognition. As we shall see, certain professions and social roles appear to put one in positions such as these. There is a practical benefit to this in that it would be feasible to make education about implicit bias a part of workplace training: something that one must complete if one is to work in that field. In what follows, I will flesh the argument above out somewhat, before identifying groups which do meet the specification outlined above.

It would be useful to say more about the claim that people in positions of social and professional power who make decisions for and about people with mental illness are capable of enforcing

stigma rather than merely discriminating. 'Power' can be defined in many ways, most simply as the ability to act or to perform an action; as a capacity to act; strength; force etc. Power can also be understood as "the possession of control or command over others; authority; and ascendancy (e.g. power over people's minds)" (Cutcliffe & Happell 2009, p.117). Power is thought to be a necessary condition for the propagation of stigma. If we recall Link & Phelan's (2001) model of the stigma process, it was observed that stigma was only possible where the stigmatizing parties had the power to make their discriminatory actions meaningful, and to make them matter in the real world.

Yet, there are many professional and social roles in which individuals occupying them are able, or indeed, are required, to make important decisions for and about people with mental illnesses. In this way, even the decisions that *individuals* make can enforce significant and meaningful consequences on those with mental illnesses. When one extrapolates this further, it is easy to see how particular groups of people could exert a lot of influence over the lives of those with psychiatric disorders. In this sense, individuals or groups have the power to stigmatize, rather than merely discriminating. In this way, any discrimination they do unintentionally carry out is doubly concerning in that the consequences of said discrimination are so significant.

The second claim- that individuals in particular working or professional environments will be more likely to rely on implicit type-1 processes, and thus are more at risk of unintentionally discriminating- is based on the literature on implicit bias. As described by Huebner (2016, p.66), there is evidence that when we are under even minor stress, we tend to "abandon computationally taxing model-based processing, and to rely on computationally cheaper forms

of model-free or Pavlovian processing[15]". In the terms I have used here, we tend to rely on implicit processing when we are stressed. To this, Huebner (2016, p.66) adds that we rely on such implicit processes when we are placed under high cognitive or affective load, or when the associations tracked by the bias are primed or made salient to us. In these situations, we are more likely to rely upon type-1 implicit processes, which crucially, can include implicit bias. Where we are stressed, under heavy cognitive or affective load, or where the associations are primed, we are more likely to act on our biases, and to unintentionally discriminate if those biases are negative ones, or based on negative associations[16]. I suggest that certain working and social environments are likely to put us in these kinds of situation very frequently. As such, these environments may encourage us to rely on type-1 processes, implicit biases, and thus carry out discrimination (even where doing so is against our explicit beliefs).

In this way, I suggest that groups who meet both conditions- that is, who have the power to make significant decisions for and about mentally ill people, and exist in working and social environments which encourage them to rely on implicit biases- are at an increased risk of carrying out stigma implicitly. This is why I think it advisable that they be targeted for education about implicit social cognition. Yet, who are these groups? It seems to be that several can be identified. I will primarily focus here on healthcare workers and the police force, although I will also identify a few others.

[15] See Crockett (2013) and Schwabe & Wolf (2013).
[16] Of course, in some cases our biases will not be problematic. However, here I am interested in those that are.

i. Healthcare Workers[17]

Healthcare workers are undeniably granted power to make decisions for and about people with mental illnesses, and in many cases, these decisions can be very significant. Indeed, part of the challenge posed by mental illness and for caring for people with mental illnesses is that healthcare professionals will be required to make decisions like this very often. In medicine and in psychiatry, the healthcare professional is often thought to be the epistemic superior of the patient: that is, to be an expert on the body and what goes wrong with it (Ho 2014). In this way, they are often expected to make decisions. Ideally, clinical decisions should be made by both clinician and patient, through mutual agreement: healthcare professionals may be experts on the body, clinical skills, treatments and such, but the patient is also an expert on their own lived experience of illness. Yet, this desired relationship of mutuality is exacerbated by many of the concerns particular to psychiatry, and in treating people who are mentally ill. Indeed, in severe mental illnesses, there are cases in which patients' judgements are impaired, and their autonomy must be limited, sometimes severely, in order to ensure their welfare and to protect them.

For instance, where one's illness is severe, one's capacity to care for oneself, make important decisions, and handle one's own living and financial circumstances may be called into question. In milder cases, one's ability to remember information accurately, express oneself, form coherent opinions, look at information in a proportional manner or form cogent belief systems can be (rightly or wrongly) called into question. As a result, healthcare workers are often called

[17] I use this term to refer to any medical professional, not merely doctors or nurses. I do so in order to acknowledge that care teams are very diverse, and contain great many different members of staff.

upon to make important decisions for these patients where they feel their judgement is lacking. The decision to take autonomy away from a patient, particularly where the infringement is severe, is itself an important and significant one. Yet, the consequences for patients can also be profound, once again, particularly where the infringement is severe. Individuals can be sectioned (and thus held against their will), subjected to involuntary treatment (in extreme cases, forcibly medicated) and their personal freedoms limited by mental health professionals where it is deemed appropriate.

Even where the illness is not so severe, healthcare workers can still make decisions about mentally ill people which are significant. For one, even giving someone a diagnostic label is a significant act in that it marks someone out as being mentally ill, and may expose them to the harms associated with mental illness and stigma outlined in chapter 1. Decisions about treatment or medication can be significant. For one, patients can only access certain treatment options if it is granted by a healthcare worker. In short, I suggest that healthcare workers, particularly those working in psychiatry, are often called upon to make significant decisions for and about people with mental illnesses: indeed, this is arguably necessary, given the impact that mental illness can have on cognitive function. In this way, they meet the first part of the dual-condition I have identified for a target group at risk of unintentionally carrying out implicit stigma.

I suggest that healthcare workers also meet the second condition: that they exist in working or social environments that encourage reliance on type-1 processes, and so are at risk of acting on implicit bias and unintentionally discriminating. To reiterate, the features of these environments are that they induce stress, place a heavy cognitive or affective load on us, or

prime the particular associations which are encoded in the bias itself, making us more likely to act upon it. To demonstrate this, it would be useful to look at an example of a healthcare worker who exists in such a working context.

> ***Jane the psychiatrist on a bad day.*** Jane wakes up, gets ready and heads in the work. On her drive in, she hears on the radio that a man suffering from schizophrenia has recently committed a murder. When she gets to work, she gets behind. Many of her appointments run late, and she struggles to find enough time to properly deal with all her patients. Several of her meetings don't go well. One of her patients has recently had a crisis, and is distressed in the meeting. Jane is unsure how to proceed with this patient, as she feels she has exhausted all available medication options. Another patient is aggressive towards Jane, annoyed that he cannot get an appointment for weeks. Jane then learns that funding for some of the local mental health services is being cut, and she is very concerned about the effect this will have on her patients.

Of course, this example is something of a caricature: it is unlikely that quite so many negative elements would all coincide to this degree. Yet, whilst real working environments are unlikely to be quite so extreme in their negative aspects, elements of Jane's day will be familiar to us: we hear something which appears to confirm a stereotype in the news, are late, or under time pressure, do not have enough time to do everything we want to, witness upset or distress, become upset, distressed or stressed ourselves or experience anger or frustration either from others or from ourselves. The characteristics of Jane's working environment cause her to be stressed, pressed for time, under a heavy cognitive load (both rational and affective) and the associative pair 'schizophrenic/dangerous' has been primed through her exposure to certain media earlier in the day.

These are exactly the kind of conditions which empirical research suggests encourage us to rely on a range of heterogeneous type 1 processes: processes which bypass the level of conscious awareness, are often introspectively opaque and which can guide a variety of behaviour, even in cases where the implicit attitude is diametrically opposed to our other explicitly avowed beliefs. As such, Jane's working environment is such that it creates a risk that she might unintentionally discriminate when encountering patients with mental illnesses, or when making decisions for and about them.

Some of the features of Jane's working environment which prime the activation of type-1 processes are likely to come with the territory. For instance, it is likely that working in healthcare in any context carries a risk of affective burden, or of encountering distressing circumstances. Further, when working with people with mental illnesses, it is easy to see how at least one part of associative pairings about mentally ill people will be primed. Yet, I suggest that the modern healthcare system is currently such that many other features of Jane's working environment are likely to be common. Indeed, I suggest that given the pressures on the NHS, and indeed, partly because of the challenges involved in delivering medical care in any circumstance, healthcare workers are often likely to experience situations such as these.

The NHS, and its mental health services, have been stretched close to breaking point in recent times, with many recognising that we are in the midst of a mental health crisis. The availability of treatment has been found to vary hugely across postcodes, with services struggling to keep up with increasing demand. Waiting times for treatments such as cognitive behavioural therapy (CBT), dialectical behaviour therapy (DBT) and to receive appointments at psychiatric assessments are high. Indeed, this struggle to keep up with demand can be seen across our

hospitals and GP practices, and extends to other public services such as local councils and the police.

I suggest that a work environment in which there are not enough resources (either material or human) to meet public need is likely to be incredibly stressful. Employees will be under pressure, both affective and cognitive. As regards the former, it may place a heavy emotional burden on healthcare professionals to see patients in distress, or to be unable to provide the standard of care they would ideally like to. As regards the latter, bring rushed, under time pressure or having to make difficult decisions may also impose a high cognitive burden. I acknowledge that the extent to which this is accurate will depend upon the particular working environment, the individual in question and the tasks which they are charged with.

However, what I do want to suggest is that many working environments in healthcare will be stressful, or will impose various burdens upon us, thus encouraging reliance on type-1 processes and increasing the risk of unintentional and implicit discrimination. This, when taken in conjunction with the fact that healthcare professionals are often called upon to make important decisions for and about those with mental illnesses, means that healthcare professionals are likely at an increased risk of carrying out implicit stigma. Indeed, this is why I think it important, aside from any practical utility in selecting a professional group, that education about implicit bias is directed at them.

ii. Policing

A pressing second example of such a target group is the police force/law enforcement. As I noted in the section on the harms of stigma, the mentally ill are often considered to be violent or dangerous, yet are actually more likely to be the victims of violent crime. Regardless, mentally ill people come into frequent contact with the police (Boyce, et al., 2015), with the latter often called upon to make important decisions for, and more commonly, about, the former. Indeed, the very function of the police force requires that members be granted power over citizens in order to keep the peace. This social power can be manifested in making several important decisions about members of the pubic: to identify them as suspects in a crime, to investigate or question them, to detain them, and even whether to use force on them (where the situation demands it). In this way, just as the police are called upon to make decisions for and about members of the public in general, so too are they asked to make these decisions for those with mental illnesses.

It is worth noting that these decisions often have significant consequences for those affected by them. Indeed, even be considered a suspect may prove hugely damaging to an individual's self-esteem, and may cause difficulty for them in their personal lives. Having one's liberty infringed through detention or curfews (even where justified) is certainly significant. Indeed, where the decision to use force is made, the consequences for the affected individual may be profound, perhaps even culminating in loss of life. It is worth noting that people with mental illnesses appear to feature quite heavily in situations where the decision to use serious force is made. For instance, a 2015 American survey revealed that of the 990 people fatally shot by the police in one year, approximately one quarter were individuals in 'emotional crisis' (Kindy &

Elliot 2015). Thus, police appear to be yet another instance of a group who are called upon the make decisions for and about people with mental illnesses, and whose decisions may bring about significant and serious consequences for those they concern.

Yet, is policing a form of employment which is likely to encourage reliance on type-1 processes and thus put its members at higher risk of carrying out implicit discrimination? That is, is it stressful, does it bring with it high cognitive and affective load, and is there is a risk that relevant associations may be primed? Once again, although this will not always be the case, police work definitely has the potential to have these characteristics: indeed, we may say that it is almost certain to in some forms. For one, the police force is a public service which is much in demand. Officers may be under time and resource pressures similar to those in healthcare, although likely not as pressing. As noted above, stresses like these might encourage reliance on type 1 processes, and so bring with them heightened risk of unknowing discrimination. Further, dealing with the fall out and consequences of crime may also generate high cognitive and affective burdens.

Yet, there is one kind of case that comes to mind which may be relatively common, and which significantly demonstrates the characteristics argued to encourage reliance on type-1 processes: that is, where officers are responding to risky or serious cases in which the person they confront is believed to be dangerous. Here, the officer anticipates danger in that she knows she is putting herself into a potentially risky situation, in which she may come to harm. This anticipation of danger, or of putting oneself in a position of danger, is likely to impose a high cognitive load and be stressful, although of course this might be less so where the officer is experienced, or

has completed specialised training. Yet, even for specialist units, dealing with complex or potentially dangerous situations may bring with it high cognitive load and stress.

To this, we might add that many of the decisions police officers may have to make in situation such as these may be under heavy time constraints (Fridell & Straub 2016). For instance, the decision whether to use force in response to a perceived threat may have to be made very quickly. Indeed, not only this, but the decision would likely have to be made whilst being afflicted with the cognitive and affective burden that would likely accompany a genuine concern that if one does not act, one will come to harm. Indeed, I suggest that these situations are going to be hugely stressful, and will impose many cognitive burdens. In his way, they are likely situations in which we might be strongly encouraged to fall back on type-1 processes, and thus place ourselves at risk of manifesting bias unintentionally.

However, there is a further worry that attending or even hearing about such cases may prime associations and so lead to unintentional discrimination. When engaging a suspect, the mental construct 'mental illness' may be made salient. Indeed, as Corrigan (2000) observes, we may be become aware that an individual is mentally ill through a series of cues: "psychiatric symptoms, social-skills deficits, physical appearance and labels". A police officer responding to a call about an individual appearing agitated, demonstrating an unusual affect or behaving violently may well infer the presence of a mental illness. The same is likely to be true if when approached the individual seems erratic, appears to refer to things that are not there, or claims to hear voices. Indeed, the officer may know in advance that the suspect is in possession of a mental health label. In such cases, the association between the constructs 'mental illness' and 'dangerous' is likely to become salient to that officer: she is likely responding to a criminal

matter which poses a risk of danger, and she is confronted with a suspect who primes the construct 'mental illness'. Indeed, the association may be generally primed by widespread stereotypes about the dangerousness or criminality of people with mental illnesses.

In this way, I suggest that the police force can be identified as a target group at heightened risk of carrying out implicit stigma: that is, members are frequently called upon to make important (and occasionally, life-changing) decisions for and about people with mental illnesses, yet exist in working environments which will demonstrate some of the characteristics which make it more likely that people will rely on type-1 processes, and so carry greater risk of unintentionally discriminating via implicit bias. As such, due to this heightened risk, members of this and related professions would be good candidates for receiving education about implicit bias.

iii. Other Groups & Further Thoughts

Before concluding this section, it is worth noting that very similar arguments can be made for other groups. For instance, perhaps those involved in the prison service may also be at risk of carrying out implicit stigma against people with mental illnesses. Prison officers may be called upon to make decisions about inmates which are similar in nature to those already described for police officers, and further, these decisions may be significant ones for those incarcerated. Further, this form of employment also carries stressors which are analogous to those faced by the police: the awareness that those in your charge have likely committed a violent crime, and so are at risk of being dangerous, for instance. In this way, this group may be at greater risk of carrying out implicit stigma.

Another potential group are social workers and care workers. It is perhaps less clear here that members of this profession will have to make significant decisions for and about people with mental illnesses. Indeed, in many cases, service users will be involved in the creation of their own care plans, and so participate in the decision-making process to some significant extent. Yet, particularly where the patients' needs are significant, it is feasible that care workers will have to make important decisions about treatment. The same is perhaps even more true of social care directors or managers. Further, these kinds of roles are very likely to possess those characteristics which encourage reliance upon type-1 processes: for instance, lack of time and resources, stress, affective burden (particularly where clients are distressed, or where carers are not able to devote the time and attention they want to them). In this way, carers may be at a greater risk of carrying out implicit discrimination, and perhaps implicit stigma where the decisions made effect significant consequences.

Further candidates for target groups may include members of the judiciary, those involved with disability and benefits reviews, and those involved with child and parent cases. In these cases, members of these professions clearly make significant decisions for and about people with mental illnesses: decisions which are likely to have profound financial, social and personal consequences. Whilst decisions such as these are, by design, slow, deliberative and evidence-based processes, I suggest that there remains a risk of bias. Indeed, cases may invoke stress or affective burden (cases where children are involved are particularly likely to do this), and associations can be primed, thus increasing the risk of implicit discrimination occurring. Indeed, a similar argument can be invoked for legislators: people who may make profoundly important decisions for and about people with mental illnesses, yet are likely to be afflicted by bias as we all are. Although, of course, we would hope that legislators will not be subject to time pressure whilst making these decisions.

Other groups can likely be identified, yet I will not discuss them further here. I the above, I suggested that education strategies should be broadened so as to include education about implicit bias and implicit social cognition. I acknowledge that more research must be done before we can say too much about the particular content of implicit biases about mental illness, yet I take it that alerting people that implicit biases exist at all and that we should be wary of them is a valuable first step. When disseminating this kind of education, I suggested that it would be best to identify target groups who would most benefit from it, or to whom delivering the education might reduce harm the most. I identified those groups which are at the greatest risk of unintentionally stigmatizing those with mental illnesses as plausible candidates. The risk of implicit stigma is highest where groups are called upon to make important decisions for and about people with mental illnesses and where their environments encourage reliance on type-1 processes, and thus increases the risk of unintentional discrimination. I identified several groups which fit this description, and would be plausible candidates for the receipt of education about implicit bias.

As noted earlier, there is also a practical benefit to identifying target groups. For one, these groups are often professional groups. In this way, it may prove easier to deliver the training. Employers may be able to require employees to complete the training programme about implicit bias, and it can be delivered in work time. Thus, it is easier to make sure that the training is attended. Holding face-to-face education programmes may also be beneficial in that it will be possible to ensure that the education has been understood, and that it has not been misconstrued. It may also serve to alleviate any distress felt by participants. Implicit biases are, by nature, often introspectively inaccessible, dissociative, and in conflict with our explicitly avowed beliefs. As such, few will be aware that they are biased, nor will they consider that

they are part of the problem of racism, sexism, or prejudice. Being made aware that one may be part of the problem may prove quite challenging or upsetting for some participants.

Indeed, it should be noted that getting participants to conduct an IAT or some indirect measure of implicit attitudes may be profoundly confronting: it may demonstrate to that individual that they do have (potentially problematic) attitudes they are not aware of. In this way, it might be one of the most effective ways of demonstrating to that individual that they are part of the problem. Yet, once again, this may be distressing. Importantly, it may also be misconstrued in that participants may assume that a high score on a race IAT means that they have definitely acted in a racist manner. For these reasons, I suggest that education (particularly that involving the use of indirect measures of implicit attitudes) should be supervised so as to avoid misconceptions, frame test results in an appropriate manner, and to offer support to any participants who are distressed).

Yet, although I suggest that identifying target groups will be hugely useful when delivering education (for the reasons that doing so will be both practically beneficial and will help to reduce harm the most), it is also worth noting that some broad target groups may actually be quite difficult to deliver education to. One example that springs to mind is employers. This group is often called upon to make significant decisions for and about people with mental illnesses- whether to hire them in the first place, or how to treat them once they are hired- and will often encounter many of the situations which encourage reliance on type-1 processes. In this way, they are arguably at risk of carrying out implicit stigma. Yet, the group 'employers' is very diverse, and includes both those at large corporations and small business owners. Whilst the former may impose a requirement that those in hiring positions attend education sessions

about implicit bias, the same may not be possible for the latter. Hence, whilst identifying groups can be useful when delivering education, it does have certain limitations.

V. Further Strategies

In this chapter, I have demonstrated that, unfortunately, stigma has far more complex and deeply-seated roots than mere misinformation and lack of understanding. Just as the mechanisms underlying stigma are more varied and complex than is often appreciated, so too should it be little surprise at this point that eradicating, mediating or preventing the expression of implicit bias is likely to be no simple matter. Indeed, as Holroyd & Sweetman (2016) have observed, implicit bias is a term which has been utilised in a very expansive manner, and can refer to a huge array of heterogeneous processes. Just as we should be wary of imposing a homogeneity of explanation on such a disparate phenomenon, so too should we wary of imposing a homogeneity of solution. It is true that some forms of what is traditionally referred to as implicit bias may be receptive to education strategies (simplistically conceived of as the process of correcting false information and disseminating accurate information). However, what is almost certainly true is that many different strategies will be needed in order to meet the myriad ways in which implicit bias might be encoded, retrieved and manifested. To close the chapter, I will briefly comment on some other strategies apart from education we may adopt when attempting to combat implicit bias towards mental illness.

i. Alteration of Context

One broad tactic which might be implemented is modification of context or environment. Many commentators have noted that the expression of implicit bias is hugely context dependent. On the matter, Huebner (2016, p.66-67) has this to say:

> Even in the best of cases, where we are not under stress, and where we are not facing an increase in cognitive and affective load, the extent to which we rely upon a particular association, or the extent to which our decisions are guided by a particular value, will be sensitive to a wide range of situational factors that guide the online construction of action-guiding behaviours. As the brain attempts to find a way to sift through all of the representations that are relevant to our ongoing behaviour, it must rely on a wide variety of pressures and contextual cues that arise in a particular situation….Precisely how such computations are carried out will depend on the context in which a decision must be made…both the strength and accessibility of implicit biases will vary across contexts.

Hence, through altering the context or environment, we may be able to reduce bias. Several scholars have argued for responses broadly fitting this remit, although their responses are usually specified to the particular account of the nature of implicit bias that they endorse. Huebner (2016) believes that implicit biases are not implemented solely through associative mechanisms. Instead, he argues for a computational account in which both inferential and associative systems guide behaviour in the cases in which it is normally said that implicit bias occurs.

Huebner suggests that his computational model of implicit bias can explicate why some interventions are likely to be fruitful, and others not. For one, he argues that problematic

implicit biases arise under certain conditions, and no matter how vigilant or attentive we are about these biases, we will fail to escape them unless we alter the conditions under which they are expressed. In particular, he echoes Dasgupta (2013) in claiming that it is not enough that we attempt to eradicate negative implicit attitudes: rather, we must also strive to develop more egalitarian attitudes. These egalitarian attitudes can only be fostered in a world in which egalitarian policies are implemented. Given that we do not inhabit such a world, Huebner (2016,p.72) argues that we must construct our own niche, and "manipulate *our world* in ways that make it represent new things for us". This process of world-building will involve the rejection of many dominant social norms, and thus to an extent, the cultivation of a type of imagination which takes advantage of the human capability to transform the attitudes, beliefs and behaviours of others.

In terms of my argument, one strategy which could be attempted in order to reduce the likelihood of implicit stigma being brought about would be to alter the working environments of the target groups. To recall, I argued that working climates are likely to encourage reliance on type-1 processes where they generate stress, high cognitive and affective load, and where they prime relevant associations. To minimise the risk of implicitly biased behaviour by target groups, we might try to alter these working environments by giving employees adequate time and resources to make important decisions, and attempt to reduce stress by cutting down on working hours, granting breaks or providing a calmer working environment (perhaps by cutting down on noise, or even playing calm music). Here, the hope would be that in doing so the environment will be calmer, and so less likely to encourage reliance on type-1 processes.

This suggestion may prove valuable, yet may be difficult to practically implement. For instance, it would obviously be desirable for healthcare workers to be granted more time and resources, yet there is no feasible solution at present as to how this might be attained. Allowing for more breaks, hiring more staff and providing more facilities all cost money, and in an institution already under such strain, it is difficult to see how this might be financed. Indeed, as of 20th February 2017, it was estimated that NHS trusts had actually overspent to the tune of approximately £900m (Triggle 2017). Regardless, alteration of context may be a valuable means of combatting implicit bias, and so perhaps further research might be devoted to exploring how this might be achieved on a limited budget.

ii. Protest

There is another way in which we might hope to alter context in order to reduce implicit bias. Empirical research has demonstrated that we are more likely to rely upon type-1 processes where relevant associations are primed. Thus, another way in which we might attempt to halt the expression of bias is by combatting the prevalence of troubling associations about mental illness, and so reduce the circumstances under which they are activated or made salient. One way in which this might be done is by challenging problematic depictions of mental illness in art and media.

The media has often been accused of depicting mental illness in a problematic light (see Backer 1985). Indeed, Thornton & Wahl (1996) have argued that media coverage of violent crimes committed by those with mental health difficulties is a major factor in negative public attitudes towards the mentally ill. They note that headlines often sensationalise the issue, accentuate the

violent or disturbing features of the crime and reinforce public fears about mental illness. Indeed, Stuart (2006, p.99) observes that "studies consistently show that both entertainment and news media provide overwhelmingly dramatic and distorted images of mental illness that emphasise dangerousness, criminality and unpredictability". Film and television are no better, with media analyses revealing that common stereotypes about the dangerousness, child-like nature and incompetence of mentally ill people are prevalent (Corrigan & Watson 2002).

I noted in chapter 1 that protest strategies are often criticised for being purely reactive, and requiring a great deal of time, attention and effort (Corrigan 2004, 2016). Yet, it strikes me that challenging problematic depictions of mental illness may be hugely valuable when attempting to eradicate implicit bias. Indeed, protest may actually prove to be preventative. If the associativist hypothesis is correct, and implicit biases are formed through association between concept pairings, then breaking down the associations (or the strength with which they are held) will likely prevent bias. Protesting problematic depictions of mental illness can be successful both in removing that particular depiction, and in discouraging the production and dissemination of other materials which contain problematic associations about the mentally ill and mental illness. By reducing the visibility of problematic associations, protest may also reduce the circumstances under which they are activated or made salient, thereby preventing reliance on type-1 processes.

If a strategy such as this were to be pursued, it may prove useful to target particular associative pairs about mental illness. I take this to mean two things. In the first case, we should seek to remove only those depictions of mental illness which are problematic. It would be unwise to prevent the media or news referring to mental illness at all. Indeed, a great deal of anecdotal

evidence suggests that reporting about mental illness and storylines featuring mentally ill characters can have some profound benefits (see Drickey 1990). Where the portrayal is suitably complex, avoids simplistic stereotyping and is accurate, the inclusion of mentally ill characters in various forms of media can serve to make the subject of mental illness easier to approach. Indeed, according to a survey conducted by the charity Mind in April 2016[18], "news reports and soap and drama storylines about mental health are having a huge impact on audiences", with:

> More than half (52%) of people who have seen a storyline involving a character with mental health problems say it helped to improve their understanding of mental health problems, while one in four (23%) have been inspired to start a conversation about mental health after seeing a story about mental health in the news.

Thus, when conducting protest strategies, we must be careful what we target. We will need to acknowledge that mental illness is a complex subject, and that many conditions can be deeply distressing to those who suffer from them. Depictions in which mentally ill characters experience stigma may not necessarily be problematic in themselves: indeed, we may think that we should allow depictions such as these in order to highlight the difficulties faced by many mental health service users. We must therefore allow accurate, affecting or difficult depictions, and challenge only those which are inaccurate or rely on stereotype.

There is also a second sense in which I believe protest may be targeted, although I acknowledge that this second suggestion is more tentative. It strikes me that certain associations about mental illness may be more problematic than others, and so should be the particular focus of protest strategies. There are two ways in which this might be true. In one sense, they may be more

[18] See Mind (2016).

problematic in that, when primed, certain associations may warrant a stronger or more negative response than others. In another sense, particular associations may be more troubling than others in that they are gained more *quickly* and are held more *strongly*. Some theoretical framework must be introduced at this stage in order to explain why this may be the case. These suggestions were initially inspired by recent work by Sarah Jane Leslie (2013) in the philosophy of language.

Leslie focuses her discussion on instances in which shocking or extreme behaviour on the part of individuals is simplistically generalised to a broader group. She gives the example of generalizations such as 'mosquitos carry the West Nile virus', which she believes are part of a smaller sub-group of generic utterances: striking property generics. Leslie observes that this generic utterance is acceptable in conversation and not obviously false, despite the fact that less than 1% of mosquitos carry the virus. This, she argues, is because the property affixed to the category (carrying the West Nile virus) is a striking or dangerous one: one which it would obviously be in our best interests to avoid. This, she observes, can shed some light on instances of social stereotyping. Where a social kind (or at least, a perceived one) is associated with a property which is particularly dangerous or undesirable, any generic utterance which predicates that property to the kind will likely be thought to be true. Hence, because we typically have an interest in avoiding danger, we will likely judge the generic 'schizophrenics attack people' to be true: crucially, even where only very few members of this perceived category actually instantiate the property.

I will say more on this in chapter 5. However, at this point, I want to make a tentative argument: that Leslie's work may suggest that certain associations or knowledge structures about mental

illness may be more troubling than others. In one sense, I noted that certain associative pairs may, when they play a role in bringing about action, be more likely to lead to more troubling or harmful behaviour. Consider two associations we might make with the broad category 'mental illness': mental illness/childlike and mental illness/dangerous. As noted in chapter 1, both are common stereotypes about mental illness. However, an individual with a high associative score between mental illness/childlike may behave in a rather different manner than another individual with a high score for mental illness/ dangerous. Why is this? It seems that on Leslie's terms, dangerousness is an obvious example of a striking property, whereas being childlike is less so. It is certainly in one's best interests to avoid someone or something one believes to be dangerous in all scenarios. If one believes another person to be childlike, one may limit their involvement with important tasks, or perhaps one may feel protective over them. However, whilst one may fail to trust them in certain matters, one is unlikely to feel the need to withdraw from them completely.

Obviously, both forms of treatment may bring about serious harms and loss of opportunities for the person to whom the negative property is predicated. However, the behaviour one might engage in if one sincerely believed the mentally ill person to have the striking property 'dangerous' appears to me to be more extreme than the behaviour warranted by believing them to be childlike: not only would one not trust them with certain tasks, but indeed, one is likely not to trust them in any capacity. Many forms of discrimination are harmful, and the end goal of any anti-stigma strategy should be to eradicate it entirely in all the forms in which it presents itself. However, it would perhaps be prudent to target those associative pairs or knowledge structures which bring about the *most* negative behaviour towards the mentally ill in the first instance. Perhaps breaking down the strength with which these associations are held will bring about the greatest reduction in discriminatory behaviour most quickly. In this way, perhaps

tackling the association between mental illness and dangerousness may take precedence over combatting the association between poor mental health and being childlike.

Leslie's work may also suggest that breaking down striking property knowledge structures may be challenging. Indeed, I noted that the second sense in which certain associations may be more problematic than others is the relative speed and strength with which they may be held. I suggest that as a knowledge structure containing a striking property, the associative pair 'mental illness/dangerous' make take very few exposures to form quite a strong cognitive bias (even where it is not manifest). Type 1 processes are normally described as utilising slow-learning memory systems which are thought to respond to experience and social conditioning (Huebner 2016). However, perhaps associative pairings which contain a striking element will require far fewer exposures to be held. That is, because it is important that we avoid dangerous things, it may take fewer exposures to an associative pairing with a striking property (such as 'dangerous') for us to register a high score on an indirect measure of implicit attitudes than it does a neutral associative pairing.

This would be in line with Leslie's recognition that certain generics are deemed acceptable with fewer confirmations than others: as Brownstein (2016) observes, Leslie (forthcoming) recognises that "it takes far fewer instances of murder for one to be considered a murderer than it does instances of anxiety to be considered a worrier". Leslie takes this to mean that striking property generics can explain some social stereotypes quite well (e.g. 'Mosquitos carry the West Nile Virus), but not others (e.g. 'Chinese people are good at maths').

If Leslie is correct, striking properties can have profound effects on our cognition. For one, generics expressing social stereotypes with striking aspects are acceptable even where very little of the category's membership exhibit the property. Perhaps striking properties may well have other effects on our cognition: one tentative suggestion is that where a striking property is attributed to an individual (and generalized to an essential kind), we will take particular note of this (even at an implicit level), and thus we may come to hold this association quite strongly even after few exposures. In this way, we may develop strong cognitive biases quickly when parts of the associative pairing are striking. Similarly, it may take more exposures to a non-striking associative pairing for it to 'stick' with us.

If this is correct, then it may have further implications. For instance, certain associative pairs about mental illness may be less easily acquired if they do not have striking elements. That is, because it is less important that we avoid those who we believe to be childlike, we may acquire these kinds of associations more slowly, and hold them in a weaker manner where we do. By contrast, knowledge structures containing striking constructs may be acquired quickly, and held strongly. This may suggest a need to target certain depictions before others when carrying out protest strategies. Once again, this is not to claim that knowledge structures such as 'mental illness/ childlike' are not problematic. However, combatting more damaging associative pairs may take precedence. In particular, those knowledge structures which are held strongly and acquired after relatively few exposures may be the most pressing targets of protest. If relatively few depictions of dangerous mentally ill characters in the media engender strong associations quickly, then it may be prudent to direct more effort into protesting these depictions, rather than depictions of the mentally ill as childlike (if it is true that these associations are formed more slowly in virtue of them being non-striking properties).

Thus, although protest strategies are undeniably resource-intensive, they may well be particularly well placed to tackle implicit bias. If the associativist hypothesis is correct, then the media and arts may be one way in which "socially salient attitudes are encoded automatically and stored associatively" (Huebner 2016, p.49). Problematic depictions of mental illness create associations which may guide behaviour. By removing these depictions, protest strategies may reduce the circumstances under which associations are created, activated or made salient, this reducing reliance upon type-1 processes, and decreasing the risk of unintentional discrimination.

iii. Hard Work

Finally, I would like to draw attention to a characteristic which will likely characterise all attempts to get rid of implicit bias for mental illness, regardless of the particular shape the strategy takes: namely, that all strategies to combat implicit bias are likely to be complicated, and will require a great deal of effort in order to achieve success. In chapter 1 I stated that education strategies to combat the stigma of mental illness have tended to assume that social ills such as stigma are the result of misinformation: when misunderstandings are cleared up, malign practices such as discrimination will cease, and in a relatively straightforward way. Yet, it is unlikely that eradicating many forms of implicit bias will be a simple matter of being given more information.

Indeed, it seems clear that it is not going to be possible to be passive in this matter: in the vast majority of cases, being fed accurate information may not itself be enough to alter our implicit social cognition. Rather, many commentators have noted that getting rid of implicit bias is

likely to require a great deal of effort, both on behalf of the individual and in terms of the collective endeavours of the society in which the individual resides. As regards the former, bias can occur even where one sincerely holds explicit egalitarian attitudes, and worse, such biases are often recalcitrant to normal methods of altering behaviour (Frankish 2016, p.38).

This is not to say that there is no hope of getting rid of biases. However, as one argument offered by Frankish (2016) suggests, eradicating implicit bias may be an incredibly complex matter: effecting change may prove incredibly difficult. Frankish (2016) argues in favour of a more moderate picture of mental architecture in which the two systems (type 1 and type 2) both contribute to cognition, but are not wholly independent. He notes that implicit biases (which he conceives as biased type 1 judgements) can, in certain circumstances, be overridden by non-biased type 2 judgements: in particular, where the individual in question has sufficient 'metacognitive motivation'. Yet, the conditions for override are strict (see Frankish 2016, p.38), and so the override will often fail.

Frankish's argument is built upon his own views about the architecture of the mind- an account which is far more detailed than the minimal account I have offered here. Yet his claim that implicit bias may prove incredibly difficult to control might still usefully inform future anti-stigma campaigns for mental illness. In a broader sense, it seems likely that the success of strategies against implicit bias may depend in some significant way on the ability and motivation of individuals to put in the hard work required and to devote resources to it. This may perhaps be somewhat concerning for anti-stigma initiatives: how do you motivate people to strictly monitor their own cognitive processes, and encourage them to continue to do so? Indeed, we might note that influencing individuals to carry out one charitable donation may

provide challenging enough, let alone a near-constant process of reflection and modification. Whatever shape our strategies to combat implicit bias about mental illness take, they will require sustained and focussed effort if they are to succeed.

iv. Conclusion

This chapter has been concerned with the ways in which recent philosophical attention paid to implicit bias can inform our strategies to end the stigma of mental illness. In particular, I have demonstrated that the problem of stigma is much larger and more complicated than we might expect, and that as such, we may need to modify existing strategies and create new ones in order to meet the challenge of implicit stigma and implicit bias. I began by constructing a minimal account of implicit bias based upon findings in gender and race. I suggested that initiatives to end the stigma of mental illness should take heed of this new research, and devote attention to investigating the content and effect on behaviour of implicit biases about mental illness (given that we have good reason to suppose they exist).

From this point, I also suggested that traditional education strategies (conceived of as the dissemination of accurate information and the refutation of inaccurate information) may prove unsuccessful in combatting implicit bias due to the characteristics it possesses. As such, we should broaden our education strategies to include education about implicit bias itself (albeit, whilst little can be said at present regarding the content of mental illness biases). I suggested that these education programmes should be disseminated in the first instance to target groups-identified as those at greatest risk of carrying out implicit stigma. Finally, I have closed the

chapter by exploring a few other strategies which may prove fruitful when attempting to eradicate implicit mental illness bias.

CHAPTER THREE

Power, Epistemic Injustice and Mental Illness

I. Introduction

The previous chapter focussed upon a discussion of how new findings about implicit bias might usefully inform our anti-stigma campaigns for mental illness, and how implicit bias might contribute to stigma. In this chapter, I propose to examine the stigma of mental illness from another perspective by exploring one way in which the mentally ill experience status loss: an element of the stigma mechanism as described by Link & Phelan (2001). In particular, this chapter will focus upon the loss of epistemic status which can so often accompany a diagnosis, or perhaps even just the suspicion, of mental illness. The undue diminishment of one's epistemic agency and credibility is known as 'epistemic injustice': the phenomenon by which people are distinctively wronged in their capacity as *knowers* (Fricker 2007).

In what follows I will outline two accounts of epistemic injustice- one belonging to Fricker and one belonging to Goguen (2016)- before exploring how both theories may be applied to the issue of mental illness, and what kinds of harms might be generated. Following this, I will argue that rectifying epistemic injustice for mental illness is a far more complicated matter than it is for gender and race. Briefly stated, the reason for this is that stereotypes of irrationality,

whilst entirely fictitious when applied to gender and race categories, are at least partially true of at least some mental disorders, and at least some of those suffering from them. I will conclude by arguing that resolving epistemic injustice for mental illness will involve a balancing act between avoiding *a priori* attributions of epistemic deficiency on the basis of stereotype whilst recognising that the symptomology of certain mental illnesses will, in fact, create epistemic limitations for those suffering with them. I suggest that advocacy may be a suitable means of attaining this balanced response, and so one tenable method of combatting epistemic injustice for mental illness.

II. Fricker on Testimonial Injustice

Fricker (2007, p.1) argues that it is possible to be wronged specifically in one's capacity as a knower. She focusses on cases in which certain forms of social stereotyping (mostly gender and race) cause a hearer to afford a speaker belonging a stereotyped group less credibility than they would a member of a non-stereotyped group: for example, where a hearer dismisses a woman's testimony because she is a woman, but accepts a man's testimony because he is a man. The hearer's prejudice leads him or her to "make an unduly deflated judgement of the speaker's credibility, perhaps missing out on knowledge as a result" (Fricker 2007, p.17). Fricker (2007, p.2) begins her analysis by noting that social power "is a capacity we have as social agents to influence how things go in the social world", which can operate either passively or actively. For Fricker, power can only be exercised within an appropriate context, with this context being defined as a functioning social world furnished with shared institutions, meanings and expectations. This 'social alignment' is the foundation of all power relationships,

and without it, they cannot exist. Fricker's conception of social power is neutral: the exercise of power need not, strictly speaking, be a bad thing for anyone. However, it can be problematic.

Indeed, Fricker holds that operations of social power are always intended to effect social control. Some operations of power depend to a significant degree upon "shared imaginative conceptions of social identity": for example, shared imaginative conceptions of what it is to be a woman, to be a man, to be black, to be white (Fricker 2007, p.7). These operations of power are instances of (either active or passive) *identity power*, which depend upon what Fricker (2007, p.8) calls 'imaginative social co-ordination' in that they are possible only where both parties are in possession of the same conceptions of what it is to fit one identity category or another, "where such conceptions amount to stereotypes". Crucially, identity power can be exercised even where one does not endorse the relevant stereotype, and normally works in tandem with other forms of social power, such as class systems.

What is a 'stereotype'? In chapter 1 I defined stereotypes as 'knowledge structures' (Corrigan 2016, Link & Phelan 2001). The kinds of things that count as stereotypes can be vastly different, and vary in their inherent ethical quality. For instance, some stereotypes can be characterised as useful generalisations based on empirical experience (e.g. I may form the generalisation that people wearing police uniforms work for the police, and so when confronted with an individual wearing such an outfit, I will assume that she works for the police, not that she is on her way to a party). Using these generalisations is pragmatically and epistemically advantageous to us as often it is not possible to access a wide enough set of data in order to form a judgement, and so we apply generalisations to make often reliable (yet not infallible) inferences on the basis of associated characteristics or on category membership.

Generalisations thus allow us to simplify our experience, rendering it manageable and permitting us the usage of certain 'shortcuts' in understanding.

Fricker does not deny the utility of knowledge structures based on empirical generalisations. She notes that the attribution of credibility is a complex matter. The judgement admits of degrees and is often made quickly. Further, it occurs below the level of conscious awareness and is likely heavily tied up with social and institutional practices. Despite this complexity, credibility assessments are an important part of our daily lives. Marsh (2011, p.281) observes that amongst the myriad ways in which we might trust someone, one salient example is the trust we might have that another person is telling the truth. When we choose to act on another person's testimony this commits us to investing our own resources: for instance, time and effort.

Marsh argues that because taking someone's word on something does impose certain costs on us, it is necessary for us to have some way of establishing credibility, and of ranking potential informants on the basis of it. We need to know who we can trust, and who we cannot, before we commit our own resources to plans or actions which are based upon the testimony of others. Yet, both Marsh and Fricker recognise that information pertinent to establishing credibility is often hard to come by, and so we usually defer to generalisations. For instance, Fricker (2007, p.17) observes that due to deficits in available data, there exists a "need for hearers to use social stereotypes as heuristics in their spontaneous assessments of their interlocutor's credibility".

Yet, Fricker holds that doing so is sometimes ethically irresponsible. Not all stereotypes are generalisations based on empirical data. Rather, some are based upon prejudice[19]. Prejudice is central to the understanding of testimonial injustice and what Fricker calls its 'central cases'. Testimonial injustice, as the name implies, is injustice carried out in our practices of testimony and testimonial acceptance. It occurs where a hearer's prejudice causes them to afford a speaker's testimony less weight or credibility. Fricker (2007, p.10) observes that prejudice can interfere with testimonial practices in one of two ways: it may cause the hearer to afford the speaker a credibility excess, or a credibility deficit. Excess is generally advantageous, whereas deficit is generally the opposite. Fricker (2007, p.12) argues that the injustice in credibility deficit should not be characterised in distributive terms (i.e. in terms of one's not being afforded a fair share of some good- in this case, credibility), but rather as a distinctly epistemic kind of injustice "in which someone is wronged specifically in her capacity as a knower": they are denied proper respect *qua* subject of knowledge. Thus, she argues that excess credibility attributions are not *generally* cases of epistemic injustice (see Fricker 2007, pp.13-14.

Fricker goes on to note that testimonial injustice may not always arise from prejudice alone. Indeed, it can also result from innocent errors in which we are mistaken about the relative expertise and motives of a speaker. Fricker believes that these cases are both epistemically and ethically blameless, and constitute only a weak form of testimonial injustice. She offers the same evaluations of mistakes which are ethically blameless but epistemically irresponsible (for instance, when one is mistaken about a speaker's level of expertise due to careless research). Fricker (2007, p.15) observes that what makes testimonial injustice ethically problematic is the

[19] Fricker's usage of this term appears to differ from that used to describe the stigma process in social psychology. Where Corrigan (2016) and other authors describe prejudice as endorsement of stereotype, Fricker seems to have in mind that 'prejudice' is a term used to describe preconceived opinions which are not empirically backed up, nor based upon reason. In what follows, unless specified otherwise I will be referring to Fricker's understanding of the term.

hearer's assessment being derived from some "ethical poison", where the poison in question is prejudice. Here, she is not objecting to the necessity of stereotype (conceived of as a 'knowledge structure') *per se* in our epistemic practices. Rather, she argues that what is distinctive about testimonial injustice is *prejudice* (conceived as pernicious, unjustified or unsubstantiated knowledge structures).

Some cases of testimonial injustice, whilst undeniably problematic and harmful, may be quite localised in character. Fricker (2007, p.20) uses the example of a panel of scientists who possess a dogmatic prejudice against certain research methods. Any potential contributor who submits material composed according to these methodologies may well be victims of prejudicial credibility deficits. However, "the prejudice in question (against a certain scientific method) does not render the subject vulnerable to any other kinds of injustice (legal, economic, political)", and so Fricker (2007, p.21) proposes that the testimonial injustice is incidental. However, there are myriad forms of social injustice which may be connected via the mechanism of common prejudice with testimonial injustice. In these cases, Fricker (2007, p.21) argues that the testimonial injustice is systematic and occurs "by those prejudices which 'track' the subject through different domains of social activity- economic, educational, professional, sexual, legal, political, religious, and so on". Being exposed to testimonial injustice also exposes one to the risk or actualisation of other kinds of social injustice. Fricker (2007, p.21) proposes to call cases such as these 'identity prejudice' in that she believes that the only prejudices which track individuals across multiple social domains are those relating to identity, where "the influence of identity prejudice in a hearer's credibility judgement is an operation of identity power". Fricker holds that identity-prejudicial credibility deficits of this kind constitute central cases of testimonial injustice, where the most severe forms of testimonial injustice are both persistent (multiply occurring) and systematic in character.

To illustrate this, it would be useful to look at one classic example described by Fricker (2007) and Marsh (2011): a central case of testimonial injustice. A young black man leaves his house late at night and heads for a 24-hour shop. On the way, he is stopped by a police officer, who enquires why he is out so late. The man replies that he is going to the shop to buy cold medicine for his grandmother, who is sick. When assessing this reply, the police officer comes to rely upon the stereotype that young black men do not tell the police the unvarnished truth (Marsh 2011, p.282). As a result, the police officer does not believe the young man's story, thus subjecting him to testimonial injustice. Here, the use of prejudiced stereotypes causes credibility disadvantages which "undermine, insult, or otherwise withhold a proper respect for the speaker qua subject of knowledge" (Fricker 2007, p.20). The youth suffers a credibility deficit: his testimony is considered less valuable than it would be if he were not a member of that category, and he is less likely to have his testimony believed, or for others to think he is telling the truth. This, as Kidd & Carel (2016, p. 6) note, leads to a loss of testimonial authority, "especially in relation to other socially and epistemically dominant groups who might enjoy a corresponding credibility excess".

This example usefully highlights the harms of epistemic injustice. There are some consequences of not being believed that are pragmatic in nature. Perhaps the youth is wrongfully suspected of a crime, or he misses out on a job. Perhaps a teacher does not believe that he is actually sick when he misses school, or a doctor does not believe he feels as much pain as he says he does. Other harms of testimonial injustice concern its potentially exposing those afflicted to other forms of social injustice. However, as Hawley (2011) observes, Fricker also describes cases in which the harms generated are specifically epistemic in character. Fricker (2007, pp.47-48) observes that suffering a credibility deficit will, particularly if it

occurs frequently, profoundly affect one's confidence in one's own epistemic capacity, and one's expectations regarding the reception of one's testimony.

If one is subjected to testimonial injustice, one may come to expect that others will not value your testimony, or one will expect not to be believed at all. This may be damaging to one's self esteem, particularly given that human conceptions of rationality, dignity, identity and agency are all inextricably linked with our epistemic statuses and practices: namely, how we interpret and convey our own experience to others and how others gain information from us. Where our credibility as epistemic agents is undermined, so too may we be undermined (both from the perspectives of others and in our own) as rational, dignified beings with the ability to act when required. Our very sense of identity may come under fire. As Hawley (2011, p.2) describes Fricker's view: "wronging someone as a giver of knowledge—by perpetrating testimonial injustice—amounts to wronging that person as a knower, as a reasoner, and thus as a human being".

III. Testimonial Injustice and Mental Illness

With Fricker's framework laid out, we are now in a position to see how the mentally ill may be exposed to testimonial injustice (for other such analyses, see Kidd & Carell 2016 and Sanati & Kyratsous 2015). In particular, I argue that the mentally ill are likely involved in what Fricker calls 'central cases' of testimonial injustice: those of identity-prejudicial credibility deficit. For one, the context of social alignment and practical coordination with other social agents which Fricker holds to be a pre-requisite for the existence of power relations and the exercise of power is undoubtedly in place. Just as there are shared conceptions in the social imagination about

what it is to be a woman, or to be black, or to be gay, so too there are undeniably shared conceptions of what it is to be mentally ill. In chapter 1 I noted that the most common stereotypes suggest that the mentally ill are dangerous, irrational, impulsive and child-like (see Corrigan 2004, Corrigan 2016, Rüsch et. al 2005). Indeed, much of the treatment people with mental illnesses receive may plausibly be based upon these shared imaginative conceptions of what it is to be mentally ill (perhaps these are encoded as implicit, and sometimes explicit, biases). Some power operations will significantly depend on such shared imaginative conceptions, and so are instances of identity power.

I suggest that as in Fricker's work, identity power can be mobilised against the mentally ill even where not intended. Fricker (2007, p.7) illustrates this by referring to an example from *The Talented Mr Ripley* in which Greenleaf silences Marge's suspicions by claiming "Marge, there's female intuition, and then there are facts'. Here Greenleaf exercises identity power over Marge: he silences her by referring to shared conceptions of what it is to be a man (i.e. rational, calm), and what it is to be a woman (i.e. intuitive, emotional). Precisely, he uses his identity power over her (which he holds in virtue of being a man) to influence her actions. Fricker notes that this may not be malicious. Perhaps he intends to calm her down and encourage her to look at matters objectively. Nonetheless, Fricker argues that this is an exercise of identity power. Interestingly, even if neither party endorsed the stereotype 'women are irrational' (or something like that), they nonetheless demonstrate imaginative social coordination in that they are both aware of it.

We can easily see how an analogous situation might occur with mental illness. There is undeniably a shared imaginative social conception linking mental illness with epistemic

deficiency. The mentally ill are often associated with failures in understanding (e.g. the neurocognitive and neurodevelopmental disorders), irrationality (e.g. to varying degrees phobias, obsessive compulsive disorder, psychosis, schizophrenia) and in some cases, deliberately lying to others (e.g. some of the personality disorders). The use of the term 'irrationality' may be quite broad here. Some disorders, particularly schizophrenia and psychosis can involve the subject reporting the existence of beliefs which are bizarre, false or inconsistent. In other cases, subjects may experience hallucinations (defined as experience in the absence of an appropriate stimuli): things which are 'not there'. In other cases, the beliefs held by subjects may appear to be almost 'magical' in character. One prominent example may be in obsessive compulsive disorder, in which sufferers commonly perceive there to be a link between repeated rituals and success in their lives, or perhaps between the ritual and the prevention of bad things from happening.

I will say a little more on this later in this chapter. Indeed, the fact that several severe mental illnesses are characterised by certain epistemic deficits means that when contemplating epistemic injustice, our treatment of the issues will differ for mental illness than it does for categories such as gender and race, in which there are no inherent links between the category and epistemic deficiency. However, at this point I want to establish that there *are* cases of testimonial injustice carried out against the mentally ill: cases in which judgements are formed on the basis of stereotypes which are not empirical generalisations, but rather, instances of prejudice.

Let's look at two potential examples of how identity power can be used to afford credibility deficits to the mentally ill, leading to testimonial injustice. Suppose Mary meets Sue, a woman

with depression. Sue tells Mary that the resident's meeting is at 6 o'clock tonight rather than 7 o'clock. Mary accesses a prejudiced stereotype with the content 'mentally ill people are liars', and from there reasons 'depression is a kind of mental illness, and so people with depression must be liars too'[20]. She does not believe Sue, and turns up at 7 o'clock. Here Mary has obviously committed an instance of testimonial injustice: she has used an incorrect stereotype to afford Sue a deficiency of credibility, and as a result, she has both missed out on knowledge and has disrespected Sue as a giver of knowledge.

However, there is another kind of case: one which is complicated by the fact that the imaginatively shared stereotypes may have some truth to them. Jade has recently discovered that her friend Betty has attracted a diagnosis of Borderline Personality Disorder (BPD). Betty tells Jade that she and her husband have had a fight yesterday in which he was in the wrong. Whilst listening to Betty, Jade accesses the imaginatively shared stereotype that people with BPD can be manipulative, and sometimes distort the truth. As a result, Jade does not believe Betty's story. Perhaps this is a conscious process, and she reasons in this way, or perhaps it is unconscious, and it merely seems to her that Betty is less credible. Jade might reply by saying something like: "Now Betty, are you sure that's what happened? You know how your illness can make you perceive things strangely". Perhaps she may simply change the subject or gently suggest that Betty was not blameless in the matter. Betty drops the subject.

This case bears strong similarities to that of Marge and Greenleaf. Jade silences Betty by exercising the identity power both she (and arguably, Betty's husband) have over Betty in virtue of them being 'sane' and her being 'mentally ill'. In the case where she directly

[20] This reasoning process may be implicit or explicit.

references the illness, the exercise of power is more overt. Jade's act may be well-intended or even unintentional. Perhaps she is genuinely trying to calm Betty down, much as Greenleaf does. Or perhaps Jade takes it as obvious that Betty will not report events accurately, and so in expressing disbelief she does not perceive herself to be doing anything wrong: she is merely expressing a truth about Betty's condition, and not consciously trying to exert social control over her at all.

The case in which Jade suggests that Betty may be wrong or changes the subject is more subtle. Yet the same thing is happening. Jade has accessed shared conceptions of BPD sufferers as liable to speak falsities, and does not afford Betty's testimony due weight. Betty may still be silenced by the exercise of identity power even if she knows she does not lie[21]. Indeed, in her friend's response, Betty is made aware (to varying degrees) that Jade has accessed the stereotype. Given that it silences her, she must also share in this imaginative conception of BPD.

This is a clear case in which identity power can be used against someone with a mental illness. However, what precisely is wrong with what Jade does? Here I will focus on exploring whether she does anything ethically wrong rather than whether she is at fault in an epistemic sense, given that Fricker observes that injustice requires *ethical* culpability or prejudice. It is true that

[21] An interesting point can be made here. According to DSM classification, diagnostic labels are afforded where an individual demonstrates a set number of operationally defined criteria, where that number is jointly sufficient for the application of the diagnostic label, but none are individually necessary. As such, people with the same illnesses might have different symptoms, and just because someone has a certain label, this does not guarantee that they have particular symptoms. Members of the public may be unaware of this: they may assume that someone with BPD has all the traits associated with the diagnostic label. Thus, Jade might assume that because Betty has BPD, she lies. Hence, whilst identity power can be exercised even where neither Jade nor Betty hold the stereotype to be true, there is a danger that common misunderstandings of the way in which psychiatric classification works may mean that members of the public (and even sufferers) may mistakenly believe that people in receipt of that diagnostic label will instantiate all the properties associated with that label.

the stereotype 'people with BPD aren't trustworthy' has some empirical basis: it is a feature of BPD that those suffering from it can sometimes be manipulative, and fail to tell the whole truth. This is a different kind of case, in some ways, from Marge and Greenleaf: unlike gender, there is something about mental illness (in some cases) which warrants affording diminished credibility. In some cases, the use of stereotype and of identity power seems somewhat justified. Maybe we can afford Betty's testimony less credibility, given that the she is a member of a group for which the stereotype might be true? Is this truly an injustice, or is it a reasonable reaction?

Let us turn to Fricker's analysis once again. Fricker acknowledges that when engaging in spontaneous testimonial exchange, there exists a need to defer to stereotypes as heuristics for establishing credibility. This can be unproblematic, but causes difficulty when "the stereotype embodies a prejudice that works against the speaker", and leads to an 'unduly deflated judgement of the speaker's credibility" (Fricker 2007, p.9). Undoubtedly, the stereotype 'people with BPD are untrustworthy' works against the speaker: being untrustworthy is not generally seen as being a good thing (on that matter, neither are many things stereotypically associated with mental illness, for instance, having delusions, being dangerous, being hard to be around, being anxious etc.). However, whether it works against the speaker or not, we might want to say that the stereotype still retains some accuracy. Just because a knowledge structure works against the speaker does not itself mean that it is problematic.

The key word here is '*unduly*'. Is Jade guilty of affording Betty an *undue* deficit in credibility? I suggest that this is a matter of context. For instance, if Jade knew from past experience that Betty lied frequently, then her conduct would not be as problematic: indeed, it would arguably

be epistemically irresponsible to uncritically believe everything Betty said. However, in the example I have offered, this was not specified. Indeed, should Jade have no reason to suspect Betty of lying other than her being a member of a category stereotypically associated with lying, then this is a problem. What is ethically wrong with Jade's response is that she utilises shared conceptions of the illness to guide her interactions with Betty, rather than direct personal experience of her. In general, credibility should be afforded to an individual on the basis of her conduct, rather than according to a broad generalisation. Jade has done Betty an injustice, and the deficit in credibility is undue, because she has simplistically assumed the stereotype to be true rather than considering her prior experience[22]. Forming a judgement on the basis of stereotype alone seems unjust, even where the stereotype is generally accurate of people with BPD. Indeed, even if Jade was aware that Betty did in fact conform with the stereotype, it would remain unethical to form a judgement based on stereotype alone: Jade would have to be aware of evidence for Betty's possessing the characteristics outlined in the stereotype, and form a judgement according to this.

We might well pause at this point. Perhaps it is problematic if *Jade* acts according to shared conceptions of Betty's illness: she presumably knows Betty, and has some personal experience of her prior conduct on which she can base judgements of credibility. But what about cases which truly are spontaneous- where we rely on stereotypes as heuristics for making credibility judgements precisely because there is a lack of alternative evidence we might consult? This seems to be a more difficult case, as in these instances, all we have is the stereotype to guide our credibility attributions, and, as I have conceded, in some cases the stereotype may have

[22] We might claim that because Jade is Betty's *friend*, her basing judgements on stereotype is doubly wrong, and violates some kind of norm of friendship. Indeed, it could be claimed that the fact that Betty and Jade are friends might have particular epistemic consequences: perhaps the fact of their friendship means that Jade ought to trust Betty more, or continue to trust her even where she would not trust someone else? Both are interesting claims, but I will not explore them here.

some truth to it. In these cases, I argue that the ethical failing lies in granting the stereotype *too* much weight in making one's attribution of credibility. Crudely put, if one knows that a stereotype has some empirical basis, this does not warrant one acting as if it was a universal truth: it might give one reason to *suspect* that someone's credibility is diminished, but it does not give one reason to be *certain* that it is. What does this mean, in practical terms?

I argue that where action is not immediately required, one should hold off making credibility assessments about people with mental illnesses until one has more information. That is, to wait until one has an ample body of information concerning that individual with which to make one's judgement. This will allow one to make the judgement on the basis of that individual's characteristics, rather than shared imaginative conceptions of what it is to have that illness. For instance, if Jade meets Betty for the first time, even if she is aware that Betty has BPD I suggest that it would be ethically problematic to apply this stereotype to Betty (or indeed, to make a credibility assessment based on stereotype). The reason for this is that Jade does not need to make a credibility assessment, nor to act. Thus, her application of the stereotype leads to an undue judgement of Betty's testimony in that it is not only unproven that Betty actually is untrustworthy, but also, Jade has made this judgement where she didn't even need to. Thus, where credibility judgements and action are not immediately required, one should hold off making judgements of credibility for as a long as possible, or at least until one has adequate evidence upon which to base that assessment.

It is worth anticipating a likely objection here. In holding off from making an attribution of credibility until further evidence is gathered, are we not subjecting the mentally ill to differential treatment? Indeed, we tend to assume that people are generally credible. Holding

off from making a judgement about someone's credibility because they have a mental illness is certainly a departure from this, and so arguably itself a form of epistemic injustice. Yet, whilst I acknowledge that holding off making a credibility assessment does constitute differential treatment, the response need not be undue, and so does not constitute an ethically problematic act. As I have explored, there are some features of at least some mental illnesses which may well properly affect credibility assessments.

Indeed, given that there is some link between mental illness and epistemic capability, it would be epistemically imprudent if we were to ignore some stereotypes completely. Yet, it is of course ethically irresponsible to allow a stereotype to colour one's assessments of individuals and groups in an unguarded, unnecessary or enduring manner. My position- that we ought to hold off making an assessment until it is necessary and we have adequate information with which to do it- occupies a middle ground, intended to balance avoiding the epistemic costs of disregarding potentially useful information with the simultaneous avoidance of the ethical pitfalls of stereotype. Thus, holding off making an assessment because someone is mentally ill does constitute differential treatment, but this is not necessarily undue, and may prove the safest course of action on balance (indeed, it is sometimes safest for the mentally ill person themselves). In this sense, it need not itself be a form of epistemic injustice.

However, I acknowledge that it will be not always be possible to hold off making an assessment: there will be situations in which inaction is not feasible. Perhaps when encountering someone with mental illness, sometimes one must make a credibility judgement as a decision needs to be made and resources invested. These kinds of situation are unlikely to come around that often. Where they do, I argue that the stereotype can be used to guide

judgement and action, but that one should be very careful how one applies the stereotype in order to avoid committing an injustice. This is likely to be a complicated matter, but a few conditions can be formulated which might go some way to ensuring that the use of stereotype is not undue. Firstly, the situation must be one in which one *has* to make a spontaneous credibility assessment, and where one has *no other* information to go off apart from the stereotype. Secondly, one must have good reason to suppose that the stereotype is accurate, or has some empirical basis. Finally, the stereotype can only guide *that particular* credibility assessment and the resulting action. That is, one cannot use stereotype to make general assessments of credibility: rather, one must apply it on a case-by-case basis.

To this I add a general suggestion: that attributions of credibility should function somewhat like attributions of capacity. For instance, the 2005 Mental Capacity Act suggests that capacity is not an all or nothing measure: one does not simply have capacity, or not. Rather, when attempting to establish capacity, one should ask 'capacity for what?'. Someone may lack capacity to decide whether they have a baby, but be perfectly capable of deciding whether they want to meet a friend for lunch. If we accept that we need to make credibility attributions because of the need to invest resources (as Marsh argues), then to my mind, it is sensible to enquire 'how many resources?', 'for what purpose?'. When we decide whether or not to trust someone, we don't have to either trust them completely or not at all. As Fricker observes, credibility can be afforded as a matter of degree.

To summarize: where we are in a situation in which a credibility assessment must be made (to carry out an action) and we truly have a poverty of other relevant evidence, the application of stereotype may not necessarily be undue. However, it must only inform as minimal an

assessment and judgement as possible: one cannot use it to decide whether someone is credible full stop, but rather, one can only use it to decide whether someone is credible in that moment, relative to the resources one must invest. Furthermore, the stereotype can inform the judgement, but it should not determine it. That is, when making a credibility assessment, the stereotype can factor into the decision, but one should not assume that it is straightforwardly true. Finally, one must be prepared to revise one's assessment in the light of new evidence (either positive or negative): i.e. when the assessment is no longer spontaneous[23]. To fail to do this is, I argue, an ethically irresponsible use of stereotype, and thus an instance of testimonial injustice.

Let me illustrate this with an example. Suppose Jade and Betty had not previously met, but that Jade was aware that Betty had attracted a diagnosis of BPD. They meet in the gym and say hello to each other. Here Jade does not need to make an assessment of Betty's credibility: she does not need to invest any of her own resources into a course of action informed by Betty's testimony. In this case, it would be completely undue of her to apply the stereotype of untrustworthiness and afford Betty diminished credibility.

However, imagine that they meet again and Jade is called upon to make an assessment. Suppose Betty tells Jade that she and her husband have fought, and that he was in the wrong. Here, it appears that Jade does need to assess Betty's credibility as she is called upon to respond to this

[23] Indeed, when making credibility judgements about mentally ill people, the capacity model demonstrates why it is important that we be willing to revise such judgements on a regular basis. Indeed, one may lack capacity to decide whether to have a baby when one is in the grip of a severe psychotic episode, but when one is in established recovery, one may have capacity to make the same decision. Analogously, one may plausibly lack credibility when one suffers from BPD and one's symptoms are not managed, but this may not be true when one is in receipt of successful treatment. Thus, it will be particularly important for mental illness that credibility judgements continue to be revised to allow for progress and recovery. Indeed, we must be careful that credibility attributions for the mentally ill are not taken to be unduly enduring.

information in some way. I suggest that here the best course of action is for Jade to hold off on making wide-reaching attributions of credibility when making this decision. Instead, she should assess Betty's credibility in this case, and be open to revising whatever judgement she reaches in the future.

Further to this, even though there is no other information available to her, Jade should not allow the stereotype to completely determine her judgement. She may well consider the *possibility* that Betty is lying, but she should not assume she is. In her reasoning, Jade may well ask herself 'what will I have to do if I accept her testimony?', 'what do I have to lose if I take her word for it?'. This may well depend upon the nature of the argument Betty and her husband had. If Betty is accusing her husband of leaving his underwear on the floor, Jade is committed to very little: perhaps some advice or sympathy. In this case, the most reasonable response would likely be for Jade to take Betty's word on it and assume that she is credible. However, if Betty is accusing her husband of abuse, then Jade is committed to much more. Given that Jade must invest more resources here if she gives Betty credence, she may be more cautious in judging whether Betty's testimony is credible or not. She may attempt to delay making a credibility assessment until she has more information. Or perhaps given the lack of other information, she may wish to at least consider the stereotype (given that her use of stereotype is subject to the restrictions outlined above). If done correctly, neither approach need constitute an injustice.

It is worth acknowledging one final limitation of the response I have offered. I have spoken of our responsibility to refrain from carrying out testimonial injustice, and have suggested some conditions which, if met, might help us to avoid awarding undue deficits in credibility. Yet this may be challenging, given that we may well be unaware that we are making undue assessments

at all. Where we are aware, modifying our behaviour may still prove difficult. Attributions of credibility are, like implicit biases, complex, and often bypass the level of conscious awareness. Yet, just as this should not resign us to implicit bias, so to can we take measures to combat epistemic injustice occurring through unduly diminished credibility. Perhaps, like implicit biases, some forms of credibility attributions are introspectively inaccessible. If this is true, the same treatment as implicit biases may be warranted: an education programme describing how we sometimes aren't aware how we judge credibility. Further to this, perhaps it is the case that introspection is warranted. That is, we likely have a responsibility to reflect upon and monitor our credibility assessments, ensuring that they are only made where necessary, and taking measures to ensure that our judgements are regularly revised. Indeed, we have this responsibility in virtue of the harms associated with testimonial injustice.

i. The Harms of Testimonial Injustice

Monitoring how we make credibility assessments will likely be a complicated matter, but there is considerable ethical impetus for doing so. Indeed, we must be very careful about how we use stereotypes to inform judgements of testimonial (or more broadly, epistemic) credibility. The reason for this is that, on Fricker's model, exposing individuals to epistemic injustice (particularly those forms of injustice based upon a shared conception of social identity which can be tracked across multiple domains) brings about serious harms (both pragmatic and epistemic) for affected individuals and further exposes them to other kinds of social injustice. Mental illness is certainly an example of an identity prejudice: the shared imaginative conceptions of mental illness track sufferers across multiple domains such as employment, housing, law and into their social and political lives. As outlined in chapter 1, people with

mental illnesses can suffer from myriad forms of social injustice. Thus, if one commits a testimonial injustice against someone with a mental illness, it is incredibly unlikely to be a one-off instance of injustice which, whilst harmful, is localised.

Indeed, just like gender and race, testimonial injustice for mental illness is connected via a common prejudice with other types of injustice, and so should be properly thought of as *systematic*, to use Fricker's terms. Indeed, mental illness constitutes a case in which speakers may be exposed to testimonial injustices which are both persistent and systematic, and so fall into the category of testimonial injustices which Fricker deems to be most severe. As such, just as we have an ethical responsibility to prevent ourselves from committing testimonial injustices against women and black people, so too do we have a responsibility to prevent this from happening to the mentally ill. In part, this is because one form of injustice will lead to another.

However, we can also analyse our responsibility to refrain from committing testimonial injustice in terms of harm prevention. Many of the harms of testimonial injustice have already been discussed in my analysis of Fricker's work. Some, like in the case of the black youth, will be pragmatic. However, there are pragmatic harms of testimonial injustice which are distinctive to mental illness, and arise in the domains of psychiatry and healthcare.

Of course, not all instances in which we dismiss the testimony of someone with a mental illness will constitute testimonial injustice. Indeed, in many cases the assessment of diminished credibility will be warranted due to the symptomology of the condition, and based upon sound clinical or personal judgement. However, establishing credibility is a complex process, and one which may go wrong for several reasons. Testimonial injustice against those with mental illnesses undoubtedly occurs, and where it does, it may be very harmful. For instance, some

authors have suggested that committing testimonial injustice against those who are ill, particularly those who are mentally ill, may impede treatment. Sanati & Kyratsous (2015) have argued that patients with delusions are sometimes subjected to testimonial injustice due to the difficulties in distinguishing delusions from other forms of epistemic irrationality. This may be concerning, given that psychiatric evaluation is mostly conducted by a psychiatrist listening to patient testimony and making evaluations accordingly.

More generally, we might argue that trust is a necessary part of the therapeutic relationship, and that where testimonial injustice is committed (and indeed, epistemic injustice more generally), trust will be severely violated and the relationship may break down, jeopardising the quality of care. On this point, Lakeman (2010) describes a case in which he was administered with a psychotropic drug to which he had an unusually strong reaction. Upon reporting this, he was met with incredulity and scepticism: his testimony was not accepted. This may obviously be dangerous in some cases, and at the very least will damage the therapeutic relationship. Patients who experience testimonial injustice may lack confidence, and might fail to participate in the planning and delivery of their own care. Where abuses, or simply poor standards of care arise, patients may not feel able to criticise them, fearing they will not be taken to be credible. This, combined with the credibility excesses which Fricker describes can be granted to doctors, may place much of the decision-making responsibility with the healthcare professional, thereby increasing the risk of implicit stigma as I described in chapter 2 by augmenting the opportunities for healthcare workers to make important decisions for and about people with mental illnesses.

A final point can be made here. Whilst instances of testimonial injustice towards the mentally ill undoubtedly arise, not all cases in which an individual does not base her actions on the testimony of someone with mental health issues constitute an injustice. Yet, we must be careful that we do not commit further epistemic injustices when considering the testimony of people with mental illnesses. This can be demonstrated in the example below.

A medical professional may listen to a patient's testimony (for instance, he claims that there are spiders all over the walls), yet may not give him credence in that she does not base her own actions on what he reports. Here the medical professional uses her knowledge of the condition and experience of the patient to judge that the patient's utterance is not credible, and so she decides that she does not need to hire an exterminator or something similar. However, even though she does not afford him credence in the sense of her basing her actions on his testimony, this does not mean that she should disregard his testimony altogether. It may not be a suitable basis upon which to base her own actions, but it is certainly valuable: both clinically and to the patient.

Indeed, dismissing this patient's testimony altogether runs the risk of losing valuable insight into his condition, state of mind and beliefs. Refusing to acknowledge his belief at all may be very distressing for the patient, and may deny him a form of self-expression. As Bortolotti (2017) notes, merely because a belief is irrational does not mean that it lacks value, or should be dismissed. Therefore, although the medical professional may be right to think that his testimony is not a solid foundation upon which to base her own actions, if she is to dismiss it altogether then she does him another kind of epistemic injustice: she appears to claim that his beliefs are completely nonsensical, or are not worthy of attention being paid to them. The

utterance may not be reliable evidence that there are spiders on the walls, but it is a valuable report of the patient's beliefs at that time. As such, whilst she need not act on the former, she should take the utterance seriously as an instance of the latter, and take appropriate action (acknowledging that he believes there are spiders, and taking action to make him feel more comfortable, perhaps).

To summarize this section briefly: not only is it likely that those suffering from mental illness are subjected to testimonial injustice, but indeed, they plausibly suffer identity-prejudicial credibility deficits in a severe, persistent and systematic manner. I have argued there are epistemic costs to ignoring some stereotypes when making credibility assessments, yet also ethical costs when one applies stereotype unduly[24]. As a middle ground, I have argued that one should hold off making a judgement where possible, and where this is not possible, one should monitor one's use of stereotype. I have suggested some ways in which we might do this. Caution is necessary when making credibility judgements about mentally ill people as we risk adding to the burden of those who are already exposed to other forms of social injustice if we carry out epistemic injustice against them.

IV. Stereotype Threat

Fricker argues that testimonial injustice can expose individuals to other forms of social injustice. In this section, I want to explore a related point: that testimonial injustice can be

[24] Of course, there may not be any single best way of trading off these epistemic and ethical dangers.

compounded and exacerbated by other forms of epistemic injustice. To this end, I will now discuss a phenomenon which is not itself necessarily a form of epistemic injustice, but which can lead to epistemic injustice: stereotype threat. Mallon (2016, p.2) describes stereotype threat as: "the threat each of us faces in a situation in which our behaviour or performance might be interpreted as confirming a stereotype about a group to which we belong". It is a psychological threat: one fears either confirming the stereotype through one's actions, or being evaluated in terms of that stereotype.

The presence of stereotype threat can be empirically investigated by making the subject aware of their group affiliation (a process known as 'priming'), and then measuring her performance in a task in which the group to which she belongs is stereotypically associated with performing poorly. As Mallon (2016, p.2) notes, even very subtle primes can bring about drastically reduced standards of performance across various tasks. Famously, Steele & Aronson (1995) demonstrated that African-American students tended to perform more poorly than white students on SAT tests where they were lead to believe that the test measured intellectual ability. However, where the purpose of the test was not indicated, African-American students performed as well as their white counterparts. As Link & Phelan (2001, p.374) note:

> this research tells us that the existence of a stereotype and the administration of a test of "ability" can lead to an invalid assessment of the academic potential of African-American students and thereby to discrimination against such students on the basis of a seemingly "objective" test.

In this case, no one in the immediate vicinity of the stereotyped group need have acted in a discriminatory manner. Instead, "the discrimination lies anterior to the immediate situation and rests instead in the formation and sustenance of stereotypes and lay theories" (Link & Phelan

2001, p.374). The harmful consequences are generated by the mere existence of stereotype, and by the stereotyped group's knowledge of them.

It has been well-documented that stereotype threat can hinder performance across a variety of tasks. Yet, Goguen (2016, p.216) has argued that stereotype threat also generates epistemic harms which effect "the ways in which we engage with the world as actual and potential knowers". She suggests that certain kinds of self-doubt constitute an epistemic injustice, and that a broader conception of stereotype threat (namely, as having effects other than underperformance) is needed to appreciate this. Framing underperformance as the lone direct effect of stereotype is arbitrarily narrow, and risks obscuring aspects of the phenomenon which are psychologically and philosophically interesting. Indeed, she suggests that "stereotype threat is not reducible to underperformance" (Goguen 2016, p.220). As a broader definition, Goguen (2016, p.217) observes that stereotype threat occurs:

> when an individual becomes aware, consciously or unconsciously, that their behaviour in a specific social arena, or 'domain', could render salient a negative stereotype about them or their social group. At worst, their behaviour could be interpreted as confirming the stereotype.

Indeed, she observes that whilst underperformance is certainly the most well-established effect of stereotype threat, recent literature has revealed that many other exist[25]. Goguen suggests that stereotype threat is often accompanied by 'disengagement' and 'domain avoidance'. She characterizes disengagement as "a reduction of motivation to participate or succeed in a domain, such as a general sphere of life (intellectual pursuits), an academic subject (mathematics), a hobby (video games), or even a particular social scene (nerd culture)" (Goguen 2016, p.218). In domain avoidance, individuals avoid a particular domain or distance

[25] For a summary, see Shapiro & Aaronson (2013, p.97).

themselves from it (in either a physical or a social sense). Goguen suggests that these responses can all be framed as coping mechanisms which are implemented in response to certain kinds of psychological threat.

In particular, Goguen suggests that underperformance, domain avoidance and disengagement can be understood as strategies designed to protect self-worth from threats of devaluation. Disengagement represents an effort to lower one's self-esteem in the stereotyped domain, thus protecting one's self-worth. Avoidance represents an effort to prevent threats to self-worth becoming actualised or manifested. Underperformance represents efforts to avoid a lowering of self-worth by "proving the stereotype wrong, at least in this instance" (Goguen 2016, p.220). Goguen illustrates this by using an example of women taking a difficult mathematics class: a domain which carries a high risk of stereotype threat. She posits that three different women might respond differently to this threat.

The first might drop the class and choose to major in something else. This is domain avoidance: she protects her self-worth by avoiding opportunities in which the stereotype could be confirmed. The second woman attends class, but does so in a very half-hearted manner. She sits at the back and does not really engage. In doing so, Goguen argues that she deflects any implication that her performance in the class has anything to do with her being a woman, and suggests to those around her that if she does perform poorly, it is because she is not interested. This is a case of disengagement. The third woman attempts to escape the reach of the stereotype by showing that it cannot be applied to her. She goes to each class, studies rigorously and puts in a lot of effort. Despite this, Goguen suggests that if the test is difficult her attempts will likely prove fruitless, given that a variety of cognitive mechanisms have already been put into

play. As a result, the student experiences underperformance. All three phenomena are, Goguen (2016, p.221) argues, "reactions to threats of devaluation that are triggered by the presence and possible salience of a negative stereotype".

She then suggests that stereotype threat can lead to profound epistemic costs in that it impacts how we come to gain, keep and share knowledge[26]. It can undermine or erode our sense of personhood itself, which she takes to be partially constituted by our perception of ourselves as reliable knowers: indeed, stereotype threat may have severe ramifications for "the very foundations of our epistemic lives" (Goguen 2016, p.222). Goguen argues that these harms stem from doubt. We often encounter doubt in our lives. Certain doubts may trouble us more than others. For instance, after arriving at work, I may suddenly come to doubt that I have locked the front door. This is certainly annoying, but unlikely to be particularly worrying: it is a common phenomenon.

However, other kinds of doubt can have serious ramifications for our sense of security in our own epistemic capabilities. Indeed, Goguen observes that were one to have serious and sincere doubt about one's having two hands, this would have serious repercussions for one's epistemic life, and would likely cause one to doubt many of the other things we usually take ourselves to 'know': that we are reliable perceivers, for one. Goguen terms this phenomenon 'epistemic spillover', which concerns doubt arising in relation to our "networks of belief and knowledge" (Goguen 2016, p.224). Epistemic spillover occurs where doubt in one domain spills over and affects our certainty about beliefs we hold in other domains.

[26] For other commentaries on this subject, see Gendler (2011), McKinnon (2014).

Goguen observes (2016, p.224) that most beliefs and uncertainties will bring about some sort of epistemic spillover, as ideas usually imply or entail other ideas. Yet, some cases of spillover are more concerning than others: for instance, if I were to learn that I had Alzheimer's disease, one instance of doubt could lead to epistemic spillover of a profound kind. Indeed, Goguen argues that stereotype threat can cause epistemic spillover which deeply affects our epistemic lives[27] and our sense of our identity (to which it is inextricably linked). In the worst cases, this may impact upon our very sense of ourselves as rational and full persons. In her 2016 paper, Goguen offers one particularly profound example of how this might play out: a character in a television show who is in an abusive marriage. This character finds that her marriage has greatly undermined her general capacity to engage with the world as a reliable knower. The character experiences the tension and contradiction between how she might reasonably expect her husband to treat her (i.e. that as her husband, he should love her) and her actual experience (i.e. that he beats her and treats her in an unloving manner). She is, Goguen argues, so epistemically destabilised by this dissonance that she comes to doubt whether she can have secure knowledge of anything at all. Her epistemic credibility and very sense of identity is undermined by epistemic spillover (although in this case the spillover is not caused by stereotype threat).

This character's experience can be linked to Fricker's work on epistemic injustice: specifically, 'hermeneutical injustice'. Hermeneutical injustice occurs when an individual or group is not able to "understand, interpret or render intelligible a significant experience" (Goguen 2016, p.227). This can occur due to a lack of conceptual or linguistic machinery (e.g. victims of postpartum depression before the term was introduced or the phenomenon acknowledged),

[27] For instance, studies conducted by Gates & Steele (2009) and Walton & Cohen (2007) suggest that stereotype threat can lead to epistemic uncertainty of a global kind.

where an individual's manner or affect in speaking is not generally accepted as credible and where cultural norms or presumptions obtain.

Further, epistemic uncertainty may be profoundly exacerbated by existing social stigmas and marginalization. For instance, women have long been marginalized and denied voice to their experiences, and there is much cultural presumption about women (i.e. a culture might 'know' that women feel a certain way about their babies, and that sadness and anxiety are not part of this). Hence, the dissonance they experience often originates in and is propagated by stigma, stereotype and marginalisation. Goguen cites Fricker's (2007, p.163) remark that:

> when you […] seem to be the only one to feel the dissonance between received understanding and your own intimated sense of a given experience, it tends to knock your faith in your ability to make sense of the world". Thus, social stigma can engender self-doubt which spills over in an epistemic way to other domains.

Goguen argues that stigma delivered via the mechanism of stereotype will do much the same thing. Consider Steele's example of black students facing stereotype threat. Goguen notes that feeling like one is not doing well in one's academic pursuits may take on added poignancy when combined with the experience of racism. A black student may be aware of the threat that she will be evaluated according to a stereotype of academic inferiority, but she may also come to fear that her race will limit her in many, if not all, arenas of human potential and achievement (Goguen 2016, p.228). Whilst white students may also experience stereotype threat, it is unlikely to bring about the same degree of epistemic spillover as it might for black students: the reason being that the kinds of stereotypes to which they are exposed do not insinuate that this is just part of a more fundamental flaw. Unfortunately, for black students as a racial group, there exist damaging cultural suspicions of sub-human status, and so when one area of doubt

is triggered, the epistemic spillover can extend to a more global uncertainty about their ability to succeed at all.

Epistemic doubt can spill over to a more general worry about agency and humanity, given that rationality is often held to be a fundamental part of what it is to be human. Where one is suspected of being irrational, one's doubts are more susceptible to epistemic spillover because rationality is foundational to so many aspects of our social and intellectual lives: where someone has doubts about their ability to be a rational person, it is easy to see how this may lead to them fearing that they are deficient in a plethora of ways. Goguen concludes by observing that there are many social groups which are suspected of irrationality, and so at risk of epistemic spillover of this kind.

V. Stereotype Threat, Doubt and Mental Illness

Goguen's account is a plausible one. If she is right, then this may have interesting implications for mental illness. There are strong reasons to suspect that the mentally ill experience stereotype threat. Henry et. al (2010) have found that stereotype threat leads to social difficulties in those with schizophrenia. Furthermore, Foy (2013) presents a variety of evidence which suggests that being in possession of a mental health label affects performance across a variety of tasks. Quinn et. al (2004) found that having to disclose that one had a mental illness negatively affected performance on the GRE.

As Goguen observes, stereotype threat concerning irrationality is perhaps the most troubling catalyst for epistemic injustice and devaluation, given that rationality is inextricably linked to so many areas of our lives, including our own sense of humanity. I suggest that the mentally ill are perhaps some of the most likely candidates to be affected by stereotype threats regarding irrationality. Stereotypes about irrationality and mental illness are prevalent, and may be held at an implicit and explicit level by both the public and those to whom the stereotypes apply. Furthermore, there are many situations under which individuals might risk being evaluated in terms of these stereotypes. Indeed, given that rationality is so central to daily life, there exist myriad tasks under which one risks being evaluated by this stereotype. One's rationality (or suspicions of a lack thereof) is likely to be made salient on a very frequent basis, and so there are many opportunities under which those with mental illnesses may suffer stereotype threat. Compounding the matter is the fact that arguably, at least some of these stereotypes have some element of truth to them. Indeed, certain mental illnesses produce symptoms which seem to be associated with irrationality. For instance, some severe mental illnesses give rise to delusions, hallucinations, disordered thinking and an unusual or disordered manner of speaking. Further to this, having certain conditions may make it difficult for sufferers to perform well in tasks traditionally though to be diagnostic of rational capacity. For instance, conditions such as depression or anxiety may be so affecting as to make it impossible for sufferers to concentrate enough to be able to perform well in difficult tasks. Anxiety in particular may generate a large degree of cognitive interference which may make performance incredibly difficult.

Indeed, the consideration that the stereotypes of irrationality regarding mental illness are not wholly inaccurate leads me to make one tentative suggestion: that the perceived or actual truth or falsity of a stereotype may affect the frequency or degree to which stereotype threat is felt by those to whom the stereotype applies. It has been demonstrated that, despite the lack of any

evidence to suggest that there are racial differences in intelligence, black people tend to underperform in a variety of tasks thought to be diagnostic of intellectual ability. Here, stereotype threat generates underperformance even though the stereotype that black people are less intelligent is false. Yet, whilst gender and race have been erroneously associated with irrationality through falsity and prejudice, there are some features of at least some mental illness which give the stereotypes (understood as a knowledge structure) an aspect of truth.

That is, if we understand rationality to consist in being a reliable perceiver, being able to express oneself in a clear and ordered manner and having beliefs which are intelligible, true and consistent, then it is easy to see how some individuals with severe mental illnesses may fall short of this standard[28]. Given that the stereotype under which people with mental illnesses risk being evaluated is true, in at least some cases, perhaps this may have interesting implications: perhaps experiencing stereotype threats about irrationality will be very common amongst people with mental illnesses, or perhaps they might feel it particularly strongly, leading to a more pronounced reaction (for instance, a more profound degree of underperformance, domain avoidance or disengagement)? Perhaps it is the case that irrationality stereotype threats about for mental illness will prove harder to tackle than those for race and gender? I suggest this somewhat tentatively, but this is certainly an avenue of research worth exploring.

[28] Once again, this does not mean that we can understand the stereotype to be wholly and simplistically true or accurate. Some people with severe mental illnesses may truly be 'irrational' yet others will be capable of some rational thought, whilst others (either because they have formed adequate coping strategies or because their condition is not particularly severe) may be as rational as those not in receipt of a mental health label.

Another interesting point arises from applying Goguen's work to the topic of mental illness. Rationality is central to conceptions of humanity and agency. Where one has reason to doubt one's rationality, one may come to doubt great many other things as a result of epistemic spillover. I suggest that because the stereotypes of irrationality and mental illness are so prevalent (and in some cases, accurate) and because there are so many instances in which one risks confirming the stereotype, that the mentally ill will be particularly susceptible to epistemic doubt and global uncertainty. Doubt will arise frequently, and may be profoundly troubling for the mentally ill, and the epistemic spillover may be great indeed. The reason for this is that stereotype threat for mental illness is exacerbated by two other factors: the presence of other forms of epistemic injustice, and the myriad examples taken from daily life in which the mentally ill may be given cause to doubt.

On the first point, I noted in section III that the mentally ill experience testimonial injustice, which itself may engender doubt and epistemic spillover. We might add to this that those with psychiatric disorders also likely experience hermeneutical injustice. As psychiatry is still such a young science, throughout history and up to the modern day the language available to people with mental illnesses to describe their experiences was, and still is, relatively impoverished when compared to that available to most people. Indeed, the terms used to describe many mental illnesses are relatively new, whilst it is entirely likely that many disorders have not yet been identified, or adequately distinguished from similar disorders. The language has simply not been available.

This has not been aided by the culture of silence and shame which typically surrounds mental health (particularly in certain cultures). The language required to make experiences of illness

intelligible may not be available, and furthermore, people may not feel like they even should interpret or disseminate their experiences. In this sense, they experience hermeneutical injustice. It is likely that these forms of epistemic injustice will exacerbate one another, each mutually supporting the other in causing the person suffering from mental health issues to doubt their own rationality, humanity and epistemic status. This may then go on to seriously affect them in their daily lives, leading to a degree of epistemic spillover which is particularly high. Indeed, this is yet further compounded by the fact that the mentally ill experience social stigma, marginalization and cultural presumption, much as Goguen argues women do. Where stereotype threat is felt by the mentally ill, it will generate profound epistemic harms. Yet, even worse, these epistemic harms (including doubt) will be exacerbated by other forms of epistemic injustice and stigma, rendering the situation ever more harmful.

i. Mental Illness and Doubt

Epistemic spillover may be increased or exacerbated for those with mental illnesses due to other sources of doubt they may encounter. Many things may bring someone with a mental illness to wonder, as Du Bois' (1903) voice does in *The Souls of Black Folks*, 'what if they are right? Maybe I think that I am rational, but I am not'. For instance, certain features of some mental illnesses will inherently lead to doubt. If one experiences hallucinations or has bizarre beliefs, one's rationality will immediately be called into question when a medical professional or even acquaintances report that they are not able to see what one does, or that one's beliefs seem strange, bizarre or unsubstantiated. This would certainly lead to profound epistemic spillover. However, it should be noted that this epistemic doubt, whilst undeniably worrying

for those afflicted by it, may not be unwarranted. Indeed, at least some mental illnesses appear to engender doubt as a result of their symptomologies. As such, just because someone is given reason to doubt their epistemic capacities does not entail that they have been subjected to an epistemic injustice.

However, doubt of one's rational capacities may arise even in cases where one's rationality is not inherently affected by the symptomology of a mental disorder. Recall the case of Betty, the BPD sufferer who tells her friend Jade that she and her husband have quarrelled, and that he was in the wrong. Those with Betty's mental health label are not usually associated with having defects in rational ability. Yet, for reasons outlined in an earlier section, Jade subjects Betty to testimonial injustice and does not take Betty's testimony seriously. In perceiving that Jade does not believe her, Betty may plausibly doubt herself. She may think 'Jade is my friend, but she doesn't seem to be giving me the help or support I would usually expect. Perhaps she's right and I'm not thinking clearly, or maybe I have overreacted'. In perceiving that she has been afforded diminished credibility by her social peers, Betty may come to doubt herself rather than attributing Jade's response to testimonial injustice.

Doubt may also occur where mental illness is minimized or dismissed. Imagine a university student, Luke, who suffers from depression. A few days before a test Luke finds that he is very unwell. He informs the school and they arrange for him to take the test later. The next day, he speaks to a friend, who says 'I can't believe you made that up just to get out of taking the test. Everyone knows that you can't be too sad to go to school'. Here Luke may come to doubt his evaluation of himself, and his own judgement about what he is feeling. He may come to think that he was wrong to think he was too ill (he may think 'what if I've been deceiving myself

and really I'm just being lazy and useless?') and may also doubt his belief that mental illness does exculpate him from certain things as it would with physical illness (maybe he might think 'if you can't be too sad to go to school then clearly I just can't cope with the things other people can'). Here, social misunderstandings about mental illness and the profound effects it can have on those suffering with it might cause Luke to experience profound and potentially wide-reaching epistemic doubt.

There may be many more cases in which those will mental illness will experience epistemic doubt, but I will not list them here. However, what is common to all these cases is that mentally ill people may be given cause to doubt their rational capacities in other ways, and these further sources of doubt will likely exacerbate the epistemic spillover experienced from stereotype threat, or make stereotype threat more likely to occur. The high prevalence of both doubt and stereotype threat for those with mental illnesses will, I suggest, make it likely that they experience global uncertainty of a particularly damaging kind, perhaps leading to insecurity about their own personhood or identity, as Goguen describes.

ii. Potential for Explaining Behaviour

The final point I want to make is slightly unrelated. I suggest that stereotype threat may be able to at least partially explain some common patterns and behaviour demonstrated by mental health service users. One obvious example where stereotype threat may have some explanatory power is where people in receipt of diagnostic labels underperform in tasks understood to be diagnostic of rational ability. However, there are a few possible manifestations of stereotype threat which may be less obvious, and more interesting. In Goguen's analysis, she notes that

whilst underperformance is the most well-documented effect of stereotype threat, her expansive definition can include other phenomena such as disengagement and avoidance. This may have interesting implications for the subject of mental illness, and may go some way to providing at least a partial explanation for certain behavioural phenomena that have been observed by social psychologists, anthropologists and medical professionals.

Indeed, I noted in chapter 1 that many commentators have been puzzled by the fact that many people with a mental illness do not complete treatment, whilst many never seek treatment at all. In answer, it has been suggested that this might be partially explained by stigma: individuals practice label avoidance to prevent themselves from being exposed to the pitfalls associated with being in receipt of a diagnostic label. However, Goguen's analysis hints that there may be a related, yet more specific explanation. I suggest that failure to seek or complete treatment might be a form of domain avoidance on behalf of the person with mental health issues. The female student who switches from Mathematics to English *limits* the domains under which she risks being evaluated by negative stereotypes. Perhaps someone who is of poor mental health and suspects that she may have a mental illness does not seek treatment because she is aware that in attending places and services associated with mental illness and receiving a diagnosis, she exposes herself to the label and a variety of attached stereotypes, and also hugely *expands* those domains under which she may would experience stereotype threat associated with mental illness. As such, she may avoid seeking treatment partly because by avoiding one domain entirely, she drastically reduces the number of instances in which she may be evaluated according to common stereotypes about mental illness.

Here, treatment avoidance can be construed an effort to limit the domains under which she would be exposed to stereotype threat about mental illness. If she does not receive the label in the first place (by engaging with mental health services), then she cannot be evaluated according to the stereotypes associated with it. In many ways, this seems like a rational response to the reality of stereotype threat and existing cultural attitudes to mental illness. In some cases, perhaps it is an explicitly reasoned response, yet in others, it may simply be the result of implicit processes. Failure to complete treatment could be analysed in a similar way: in distancing oneself from the domains of psychiatric health, one may hope to sever the association between oneself and mental health services in the hope that one limits the circumstances under which one might be evaluated in accordance with associated stereotypes.

Where one fears stereotype threat about one's rationality, one might plausibly engage in other kinds of behaviour. For instance, rationality is obviously integral to decision-making: particularly with making sound, proportional and sensible decisions. Decision-making and information processing are a large part of forming treatment plans. In an ideal scenario, the medical professional and the patient come to a shared understanding of what is wrong with the patient, the treatment pathways available, and will cooperatively decide what is to be done. However, if one is in receipt of a diagnostic label, one may fear that any input one gives or decisions one makes will be judged according to the stereotype of irrationality. Perhaps when Laura meets her psychiatrist she does not understand some of the things said to her, or she does not agree with her doctor's suggestions regarding treatment. But she does not say anything, reasoning: "if I ask him to simplify that he'll just think I'm stupid, or if I question his judgement he'll think I'm ridiculous. After all, he's the professional, so why should I question him?'.

Here Laura practices avoidance behaviour in response to stereotype threat. Her rationale is that by not saying anything, none of her contributions can be compared to the stereotype of irrationality. Yet, in doing so Laura misses out on engaging with her own treatment, and may end up having treatment she doesn't want. Further to this, by practicing avoidance behaviour she leaves her psychiatrist to make all the decisions. She thereby disempowers herself and removes herself from the decision-making process (which, as I have explored in chapter 2, is problematic in terms of implicit stigma). Here avoidance behaviour, whilst an understandable response to the potential devaluation accompanying stereotype threat, can lead to problematic or non-ideal outcomes. In a similar vein, Laura may use almost the same reasoning, but disengage from the appointment instead. She may pretend to be blasé or uninterested, in the hope that if she says something 'irrational' then this will be attributed to her failing to pay attention, or her not caring, rather than her having a mental illness. In this case she avoids devaluation by disengaging, yet likewise, this may also have less than optimal outcomes in terms of the quality of her treatment and recovery.

Stereotype threat may also partially explicate other phenomena and behaviour observed in mentally ill people. In chapter 1 I noted that people with mental health issues often tended to be socially isolated, had difficulty forming and maintaining friendships, and suffered from low self-confidence and self-efficacy. Social interactions involve rationality to some extent, and many people would feel uncomfortable engaging in them if they had reason to think that the other party will feel negatively about them. On the first point, when considering whether to go to talk to a stranger at drawing class, David may be aware that there are stereotypes about the mentally ill being irrational, and he be aware that social interactions are one arena in which he risks being judged according to this stereotype. He may worry 'what if they know that I am mentally ill and when I start talking they think I don't make sense, or I sound weird? I had

better not talk to them'. Here David's awareness of the stereotype and concerns about confirming it lead him to avoid activities or domains in which he fears others might think him irrational. His fear of confirming the stereotype leads him to practise domain avoidance, avoid social interactions, and as a result he risks becoming socially isolated, or missing out on opportunities to make new friends. At its most extreme, he may struggle to establish a support network at all.

A similar thing may happen, but with a different stereotype: namely, that people with mental illnesses are dangerous. Suppose that Sean has psychosis. He lives in a small town, and worries that people know that he attends psychiatric services. Sean is lonely and wants to make some more friends. However, he is aware of the stereotype, and worries that people may think it true of him. When he considers starting up a conversation with a stranger, he thinks 'what if they know that I have psychosis? If they know that I do and I start walking towards them they'll be terrified'. Sean fears being evaluated according to the stereotype of dangerousness, and so reduces his chances of this happening by avoiding social interactions (i.e. he engages in domain avoidance). Doing so will limit his social opportunities, and so diminish his quality of life. At worst, this may lead to damaging behaviour such as isolation or complete withdrawal.

However, avoidance is not the only strategy he may employ here. Suppose he goes along to a football match and one of the other players starts talking to him. Sean may have the same fears about his conduct confirming the stereotype, but instead of avoiding the domain, he disengages. Just as the female student in the Mathematics class doesn't try to do well (so that if she fails, it will be perceived that she doesn't want to do well, rather than because she is a woman), perhaps Sean doesn't try to do well in the social interaction. He may really want to make friends, but

when someone attempts to talk to him, fear of stereotype threat leads him to come across as disinterested, standoffish or even rude. Here he might hope to give the impression that his failure to make friends is because he didn't want to, not because people won't like him because of his illness.

Again, there may be many more examples. However, in the interests of brevity it is sufficient to say that stereotype threat may at least partially explicate certain kinds of behaviour. If this is true, then this might have profound implications for anti-stigma initiatives. For instance, perhaps domain avoidance might be avoided if mental health services were delivered more discretely? Furthermore, if I am right about cases such as Sean's, then people in receipt of mental health labels associated with dangerousness or criminality may be very reluctant to participate in contact strategies (wherein it is hoped that stereotypes, prejudice and discrimination will be eroded or destroyed through direct contact with actual mentally ill people).

The material above, if correct, may well prove particularly interesting to people who work in mental health, but do not have a background in philosophy. That is, the psychological/philosophical concepts outlined in Goguen's work (and other work on stereotype threat) provide a framework through which they might understand phenomena which have been observed in mental healthcare (e.g. treatment avoidance, social isolation). I suggest that this might be an interesting subject for future interdisciplinary research, given that I have offered only a brief analysis of how work on stereotype threat might inform and explain phenomena observed whilst caring for people with mental illnesses.

To recap briefly: in this section I have outlined Goguen's theory of social identity threat, doubt and epistemic spillover. I have argued that mental illness constitutes a case in which sufferers are at a high risk of being exposed to the most damaging kinds of stereotype threat: those which concern one's rational capacities. I have suggested that the stereotypes regarding the rational deficits of mental illness are common, and that there are many opportunities in which one risks being evaluated according to them. As such, people with mental health issues may be afflicted by stereotype threat on an extremely regular basis. This will bring with it a high degree of doubt and epistemic spillover, likely culminating in sufferers questioning their humanity itself. I have suggested that this dire situation will be exacerbated by the fact that the mentally ill experience other forms of epistemic injustice, and are also exposed to many further sources of doubt in their daily lives. Finally, I have argued that Goguen's suggestion that stereotype threat can prompt disengagement and avoidance as well as underperformance may show that some of the behaviour seen in people with mental health problems can be analysed as a response to forms of stereotype threat.

VI. Combatting Epistemic Injustice

In the work above, I have demonstrated that Fricker and Goguen's work can be applied to the issue of mental illness, and that the mentally ill are at high risk of suffering epistemic injustice via both mechanisms. To conclude, I will return to a point that was raised in both these sections: namely, that the legitimacy of applying stereotypes to mental illness is a more complicated matter than it is for gender and race. Whilst some stereotypes are obviously false, others may have some truth to them. Inappropriate application of stereotype is one way in which someone

can be wronged in their capacity as a knower (and submitted to epistemic injustice). Indeed, this is at the heart of both of the epistemic injustice mechanism outlined by Fricker and Goguen.

On both accounts, epistemic injustice is partly constituted by the undue or unfair application of stereotype, which affects both hearer's perceptions of the speaker and the speaker's perception of herself, and generates profound pragmatic and epistemic harms. Epistemic injustice is a huge problem in mental health, and contributes hugely to the stigma of mental illness. In virtue of the harms (both pragmatic and epistemic) it generates, and its links to stigma in mental health more broadly, it is clear that in order to fight stigma, we must also tackle epistemic injustice. I do not intend to offer my own account of how this might be done here, but will instead direct the reader to other work[29]. However, I will close this chapter by commenting upon how the strategy for mental health may differ from the path we may want to take for race and gender as regards tackling stereotypes of irrationality. That is, I will outline what key differences there will be between the strategy for mental illness and those for race and gender.

I have already noted that the stereotypes concerning the rational deficiencies of women and people of colour are entirely false: there is no empirical evidence that can be offered to substantiate these knowledge structures. As such, when attempting to put right the epistemic injustice that has been done to these groups, we must restore epistemic peerhood. That is, righting epistemic injustice will involve the recognition that the stereotypes driving instances of injustice were erroneous: that groups which had been afforded diminished credibility should not have been so. As the stereotypes of rational deficit are false, tackling epistemic injustice in these cases essentially involves granting marginalised groups the epistemic status they deserve,

[29] See Fricker (2007), Goguen (2016),

and should have been granted. Simply put, it will involve the recognition that groups which have traditionally been thought of as epistemically deficient are, in fact, epistemic peers. This will be a relatively simple process (conceptually, at least), in that we just need to discredit the stereotypes altogether.

Yet, the matter will not be as simple for mental illness. It is true that applying the stereotype of irrationality may constitute a prejudiced and erroneous assumption when applied to some category members, but when applied to others it may be a useful heuristic which captures a great deal of their experience[30]. Of course, this is true of the stereotype of rational deficiency commonly affixed to women: it will be false when applied to most women, but may be true of others. Indeed, some women will be irrational, just as some mentally ill people will be. However, this seems to be the root of the difference between the permissibility of using of stereotypes as heuristics for groups like women and for the mentally ill. Even where one had no other information and an assessment must be made, it would not be ethically acceptable to use a stereotype concerning the rational deficits of women when making one's assessment. Why is this? Simply put, because the stereotype is erroneous. Whilst some individual women may be irrational, there is no proven link between the category 'women' and irrationality. Indeed, very few category members will possess the trait of irrationality, given that it is not a concomitant feature of being a woman. As such, applying this stereotype would not be appropriate: it is unlikely to be of much epistemic use, and liable to result in an ethically problematic and unwarranted reduction in credibility or epistemic status.

[30] Indeed, it is partially because the stereotype will not be true of all mentally ill people that I recommended that one should hold off on making assessments of credibility until one has enough information to make an informed judgement. Where this is not possible, and a spontaneous assessment is forced, I suggested that we might be justified in utilising the stereotype in a one-off case to assess credibility or epistemic status at that present moment (much as capacity might be decided).

However, I suggest that there is a difference here for mental illness. As with the category 'women', not all, or perhaps not even most mentally ill people will be irrational. However, there are stronger links between the category 'mental illness' and irrationality than there are between 'women' and irrationality. Indeed, the category features of some severe mental illnesses as outlined in DSM-V (APA 2013) are precisely the kinds of things which are likely to impact upon rational functioning. For instance, as outlined above, we might plausibly think that suffering from delusions, hallucinations and disordered thought (to name but a few) will frustrate our attempts to form rational decisions, and suffering from hallucinations is likely good reason to have one's testimony doubted. As Bortolotti (2009) notes, at least some severe mental illnesses make it difficult to achieve epistemic rationality (that what you believe is based on the evidence you have) and agential rationality (where your actions are consistent with your beliefs[31]). In this way, certain mental illnesses are undeniably associated with deficits in rational capacity. Given the strong link between rationality and epistemic capability, at least some mental illnesses will also lead to defects in the latter (although the degree to which this is the case for individuals may vary hugely).

Given this, combatting epistemic injustice in mental illness will not be a simple matter of establishing epistemic peerhood: at least some severe mental illnesses are categorically linked with epistemic deficit. Thus, when it comes to mental illness, epistemic injustice must be combatted whilst acknowledging the very real epistemic limitations which can accompany some conditions. In what follows I will offer an account of how this might be done. However, before doing this it is worth going into more depth about the ways in which at least some mental illnesses may seriously and routinely lead to epistemic limitations.

[31] However, these failings are commonplace, and not only found in people with mental illnesses.

Many of these have been widely discussed (see Bortolotti 2009, 2017). However, the kind of limitations I am talking about can be aptly demonstrated by bringing in some conceptual machinery from the philosophy of language. In what follows, I will demonstrate that there are mechanisms whereby certain mental illnesses make it such that their sufferers genuinely do have epistemic limitations, and may struggle to both understand the meaning of a speaker, and to express their own meaning with accuracy. In order to show this, I will use Grice's (2010) influential account of how language is used to convey meaning[32]. I have chosen to use Grice's account because of the close relationship between grasping speaker meaning, expressing one's own meaning, and being a proficient epistemic agent. Indeed, if one fails to establish speaker meaning and accurately express one's own meaning, it seems straightforward to say that this person does have significant epistemic limitations/ deficiencies. In what follows, I will use Grice's model and demonstrate that certain mental illnesses are likely to interfere with it in certain ways, or make it nearly impossible for sufferers to abide by it. Thus, I will show mechanisms whereby certain mental illnesses genuinely frustrate our models of arriving at and conveying meaning, thus making it such that sufferers genuinely possess epistemic limitations/deficiencies.

i. Mental Illness, Epistemic Deficit and the Cooperative Principle

Plausibly, one's capacity as a knower and as an epistemic agent are profoundly linked to one's ability to understand: to recognise and process what it being communicated to you.

[32] In using Grice's structure I do not intend to commit myself to saying that his account is the correct one, or that other accounts of how meaning is delivered by conversation could not be offered. It is however, relatively commonly accepted, and serves to illustrate my point. Indeed, I hold that the same point could be made utilising many different accounts of speaker meaning.

Communication can be written, but in many cases, it is verbal. Through conversation we exchange knowledge and ascertain meaning. Grice's account outlines the principles we jointly adhere to in order to do this. He distinguishes between conventional meaning (i.e. what the words constituting a sentence individually mean, and what the literal meaning of the sentence is) and implicature (i.e. what is not literally expressed in the conversation, but is implied through subversion of conversational maxims). Grice (2010, p.173) holds that conversation is governed by a cooperative principle:

> Our talk exchanges do not normally consist of a succession of disconnected remarks, and would not be rational if they did. They are characteristically, to some degree at least, cooperative efforts; and each participant recognizes in them, to some extent, a common purpose or set of purposes, or at least a mutually accepted direction…We might the formulate a rough general principle which participants will be expected (ceteris paribus) to observe, namely: Make your conversational contribution such is as required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged. One might label this the *Cooperative Principle*.

From this point, Grice distinguishes four categories (each specified into maxims and sub-maxims) which, if abided by, will generally produce results which accord with the cooperative principle. The first category is 'quantity'. Under this category fall two further maxims: that one should "make your contribution as informative as is required (for the current purposes of your exchange)" and secondly, that one's contribution should not be more informative than required by the current exchange (Grice 2010, p.174). The second category is 'quality', under which falls the super-maxim "try to make your contribution one that is true", and two more specific maxims: "do not say what you believe to be false" and "do not say that for which you lack adequate evidence" (Grice 2010, p.174). The third category, 'relation', has one maxim: "be

relevant" (Grice 2010, p.174). Grice observes that relevance may seem deceptively simple, but conceals may perplexing questions about how relevance may shift over the course of an exchange, how focused it is etc. Finally, the category 'manner' concerns how an utterance should be made (in contrast with the first three maxims, which govern content). Here the super-maxim is "be perspicuous" and the maxims "avoid obscurity of expression", "avoid ambiguity", "be brief (avoid unnecessary prolixity)" and "be orderly" (Grice 2010, p.174).

Grice illustrates each of the maxims by means of analogy to transactions which are not speech acts. For instance, when helping me to build a car, if I tell you that I need 4 screws, I expect you to give me that number, and not too many or too little. To give me 2 or 6 would be to violate the maxim of quantity. When helping me to bake a cake, if I ask you for sugar I do not expect to be given salt. If you were to give me salt, your contribution would not be genuine, and so would violate the maxim of quality. When we come to mix the cake, I expect you to had me the relevant utensil: a whisk would be appropriate, but not a book. Of course, the book may well be a relevant contribution later in the evening, but at that point your contribution violates the maxim of relation. Finally, in any transaction, a partner fulfilling the maxim of manner would be able to make it clear what his contribution was, and to execute it in a sensible way.

Grice suggests that these conversational maxims have particular kinds of conversational implicature attached to them, some of which are generated where participants in a talk interaction fail to abide by some of the maxims. We might fail to fulfil a maxim by simply violating it, opting out (e.g. by refusing to cooperate by refusing to say any more on the matter), if we face a clash (e.g. Grice gives the example of a clash between quantity and quality, in which you need to say more, but cannot provide adequate evidence for it) or we flout it. The

flouting case is different from violation. In the former case, the failure to meet the maxim is more blatant, and it remains possible that someone can flout the maxim whilst continuing to observe the cooperative principle. Grice (2010, p.176) offers an account of how this can take place:

> A man who, by (in, when) saying (or making as if to say) that *p* has implicated that *q*, may be said to have conversationally implicated that *q*, provided that (1) he is presumed to be observing the conversational maxims, or at least the cooperative principle; (2) the supposition that he is aware that, or thinks that, *q* is required to make his saying or making as if to say *p* (or doing so in *those* terms) consistent with this presumption; and (3) the speaker thinks (and would expect the hearer to think that the speaker thinks) that it is within the competence of the hearer to work out, or grasp intuitively, that the supposition mentioned in (2) *is* required.

Grice notes that conversational implicature must be the kind of thing for which an argument can be offered. For instance, even if you intuitively grasp what someone is actually trying to say when they say something which literally means something else, you must be able to explain how you reasoned this out if it is to count as an instance of conversational implicature. To work out if conversational implicature is present, Grice (2010, p.176) suggests that hearers will reply upon the following data:

> (1) The conventional meanings of the words used, together with the identity of any references that may be involved; (2) the Cooperative Principle and its maxims; (3) the context, linguistic or otherwise, of the utterance; (4) other items of background knowledge; and (5) the fact (or supposed fact) that all relevant items falling under the previous headings are available to both participants and both participants know or assume this to be the case.

One thing that is immediately striking about Grice's account is how complex the process of conveying, disseminating and interpreting meaning are. Indeed, it may appear to some readers to be an incredibly complex way of describing an activity in which we all take part in on a daily basis. Yet if Grice is right, this might have interesting implications for mental illness. Indeed, I suggest that Grice's account can usefully illustrate ways in which people with severe mental illnesses may experience difficulty in understanding. This is significant as one's success as an epistemic agent (as a knower, as one capable of disseminating knowledge, and of acting according to knowledge) is plausibly intimately linked to one's ability to understand, and to successfully communicate what one understands to other parties. I will also point out places in which the symptomology of certain illnesses may frustrate one's ability to get across one's own meaning.

To recap briefly: Grice (2010, p.173) specifies that meaning (including that derived from implicature) is arrived at via a process of reasoning, and through recognition and fulfilment of the cooperative principle and its maxims. It is a complex process, and one which Grice holds may often go wrong, even where one has no deficits in understanding. As such, I suggest that the symptomology of some mental illnesses will make it difficult for some sufferers to extract speaker meaning (and thus arrive at an understanding) via the mechanism outlined above. Grice Yet, certain mental illnesses may have features which may make it difficult for suffers to recognise the need for the cooperative principle, or to abide by its maxims. In particular, the symptomology may cause difficulties for those with mental health difficulties in establishing the speaker-meaning of their conversational partner, and also in getting their own meaning across in an exchange. Both, I suggest, frustrate shared understanding.

I will now offer some plausible examples of this, starting with what is known in DSM-V as the Neurodevelopmental disorders[33]. Intellectual Developmental Disorder can lead to deficits in reasoning, judgment, communication and social participation (APA 2013, p.33). Language Disorder is characterised by "persistent difficulties in the acquisition and use of language across modalities (i.e. spoken, written, sign language or other)" due to minimal vocabulary, limited sentence structure and a reduced ability to "use vocabulary and connect sentences to explain or describe a topic or series of events or have a conversation" (APA, p.42). This leads to difficulties in carrying out effective communication and participating in social life. Social (Pragmatic) Communication Disorder can lead to "persistent difficulties in the social use of verbal and nonverbal communication" where sufferers struggle to communicate in a manner "appropriate for the social context" and experience "an impairment in the ability to change communication to match context or the needs of the listener" (APA 2013, p.45). It is also characterised by "difficulties following rules for conversation and storytelling" and "difficulties understanding what is not explicitly stated" (APA 2013, p.46). Finally, Autism Spectrum Disorder leads to "persistent difficulties in social communication", deficits in "social-emotional reciprocity", "failure of normal back-and-forth conversation" and difficulties adjusting behaviour to meet different social contexts (APA 2013, p.50).

There are many ways in which disorders of this kind may cause problems for cooperative conversational exchange on Grice's model. Where the symptomology of the mental illness

[33] There has been a great deal of controversy around the inclusion of the Neurodevelopmental disorders in DSM-V. That is, some have suggested that they are not 'mental illnesses' at all. In describing how these conditions may produce deficits in understanding in conversational exchange, I do not mean to make any comment as to whether these conditions are classified appropriately in DSM-V: rather, I am simply using the expansive definition to demonstrate the scope of epistemic limitations which may accompany some purported examples of mental illnesses.

produces general deficits in reasoning, this may plausibly inhibit one's ability to recognise that the cooperative principle is required, or limit one's ability to act in accordance with it. One may be unable to work out what others want to achieve in a conversational exchange, and thus one may struggle to identify what kind of response is warranted. As such, it may be difficult to abide by maxims of quality or relevance as one does not know what the desired response is, or how detailed it need be. Where the symptomology of the disorder renders participation in conversation difficult, it will be virtually impossible to extract meaning from exchanges and disseminate one's own meaning effectively. Once again, it may be difficult to establish what the goals of the interaction are, and one may struggle to get across what one wants to talk about.

Disorders which inhibit one's ability to change one's behaviour or make contributions relevant to the context are likely to frustrate the maxim of relevance. Generally speaking, mental illnesses with symptomologies which hinder one's ability to communicate clearly, follow rules or orders for conversational exchange or express oneself in an ordered way may be liable to cause one to fail to meet maxims of manner (in particular, being orderly and avoiding ambiguity of expression). Indeed, these difficulties occur even where conversation is literal. Yet, as Grice notes, a great deal of meaning can be conveyed through non-verbal cues and the flouting of conversational maxims. The symptomology of the Neurodevelopmental Disorders will likely make the recognition of both of these incredibly difficult. Thus, plausibly these disorders may lead to deficits in sufferer's abilities to grasp speaker meaning and to convey their own meaning.

The Schizophrenia Spectrum and Other Psychotic Disorders may prove to be another example of conditions which frustrate the mechanism of cooperative conversational exchange offered

by Grice. Brief Psychotic Disorder can be characterised by the presence of delusions[34], hallucinations, "disorganized speech (e.g. frequent derailment or incoherence)", and disorganized behaviour (APA 2013, p.94). Schizophreniform Disorder can be characterised by much the same, with the inclusion of negative symptoms such as diminished emotional expression or avolition (APA 2013, p.96). Schizophrenia itself involves delusions, hallucinations, disorganised speech, disorganised behaviour and negative symptoms (APA 2013, p.99). Psychosis is also characterised by delusions, irrational beliefs, hallucinations and incoherence.

Schizophrenia spectrum disorders have interesting implications for the maxim of quality when one makes a conversational contribution referencing one's delusion or hallucination. Delusions are understood as fixed beliefs which do not change even in the light of conflicting evidence, whilst hallucinations are perceptions that occur without an appropriate external stimulus. Grice notes that quality demands that one does not say what one believes to be false, nor that for which one lacks adequate evidence. Disorders with delusions may strongly predispose one to violate the maxim of quality: plausibly, one may not have adequate evidence for what one asserts, and indeed, one can be presented with evidence to the contrary. If one has no insight (if one is not aware that one is ill, or that one's contribution is false), then perhaps no violation occurs. However, perhaps one does violate the maxim of quality if one continues to assert it even after being given conflicting evidence (for instance, if you are told that your perceptions are not reliable, or you are shown evidence which appears to conflict with your belief).

[34] For an interesting exploration of the philosophy of delusions and other irrational beliefs, see Bortolotti (2009).

Hallucinations may provide a similar case. Indeed, we usually take it that being in receipt of a stimulus and experiencing a perception-like experience is sufficient evidence that we did, in fact, perceive something. If I hear a voice, I may simply assume that there was an appropriate stimulus. Indeed, it is difficult to see how I might become aware that there was no appropriate external stimulus accompanying my perception, short of someone telling me they didn't hear anything. Hence, reporting a hallucination in a conversational exchange may only violate the maxim of quality if one has insight, or reason to believe that there was no such stimulus[35]. In this way, in reporting delusions and hallucinations, one may fail to adhere to the maxims of the cooperative principle. However, we may think that this is not problematic: after all, psychiatric assessments are conducted through listening to what the patient reports, and if patients do not report what is happening to them (whether this would constitute a violation of the cooperative principle or not), then it is difficult for psychiatrists to be able to establish what is going on.

However, there are other ways in which Schizophrenia Spectrum disorders may cause problems in conversational understanding which are more concerning. For instance, disorganised speech and incoherence may violate the maxim of manner. One's expression may be ambiguous, obscure or disordered, making it difficult to get one's meaning across and making it challenging for one's meaning to be understood. In the most extreme cases, manner may be affected to such a degree that the utterance is unintelligible, which may then also lead one to break the maxim of quantity or quality: the utterance one makes may be so disordered that it appears untrue or rambling. The presence of negative symptoms may also affect the conversational exchange and the establishment of shared understanding. Sufferers may not have access to a range of nonverbal cues to express their own meaning, and may be unable to recognise it in others, thus

[35] A similar argument could perhaps be made for the quasi-magical thinking often demonstrated by some sufferers of Obsessive-compulsive Disorder.

curtailing the degree to which they can access speaker meaning. Thus, once again, cooperative conversational exchange may be limited by the symptomology of these disorders.

Other disorders may also frustrate linguistic exchange. It can be noted that disorders such as Attention-Deficit/ Hyperactivity Disorder and Bipolar Disorder (amongst many others) can be characterised as severely affecting concentration and the ability to participate in sustained tasks. In the second case, sufferers may experience racing thoughts or flit from one activity to the next. Where sufferers find concentration challenging, this may affect the maxim of quantity. For instance, ADHD is often characterised by a reluctance to engage in difficult or sustained tasks. This may lead to very short responses being offered (limiting one's ability to convey one's own meaning), or with impatience or avoidance when listening to others (limiting one's chances of grasping the full speaker meaning). Where one's thoughts tend to race, one may violate the maxim of quantity in saying too much, or perhaps one may run the risk of violating the maxim of relevance in that one flits from topic to topic. Furthermore, it may prove challenging to express racing thoughts in an ordered manner, and so one's meaning may be difficult to grasp. The presence of racing thoughts in one's own head may also be quite distracting, and may prevent one from concentrating on working out the speaker's meaning.

The anxiety disorders may also make arriving at shared understanding through cooperative exchange challenging. Social Anxiety Disorder, for instance, is characterised by fear of being scrutinized or negatively evaluated by others, persistent intense fear and anxiety in social situations and avoidance of social situations (APA 2013, p.202). Generalized Anxiety Disorder involves excessive worry and anxiety, which leads to difficulty in concentrating, sleep disturbance and fatigue (APA 2013, p.222). Conditions of this nature may predispose sufferers

to avoid social interactions entirely. Where they do engage in conversational transactions, sufferers may be so worried about being negatively evaluated that they curtail their contributions.

For instance, anxiety may cause them to not say enough (violating the maxim of quantity), to fail to report their symptoms accurately for fear or seeming silly (violating the maxim of quality) or they may ramble, stutter or be ambiguous in their expression (violating the maxim of manner). Further to this, anxiety may be so arresting as to make it difficult for those with it to concentrate on working out speaker meaning. Indeed, any condition which causes upsetting or intrusive thoughts may have a negative impact on concentration and motivation to work out speaker meaning. Indeed, unless properly managed, all mental illnesses can be hugely distressing, and may cause significant pain. Pain and distress may bring with them a large degree of cognitive interference, which may make concentrating on and completing even minor tasks quite difficult.

There are cases in which the symptomology of at least some mental illnesses may frustrate arriving at a shared understanding in linguistic exchange, both in terms of recognising where the cooperative principle needs to be applied and in abiding by its maxims. I have outlined some cases in which it may be difficult for an individual with a mental illness to establish the speaker-meaning of her conversational partner, or for her to convey her own meaning in a way that can be understood and appreciated by said partner. Being able to understand what others say to you, and being able to accurately convey your knowledge to them is a huge part of being a successful epistemic agent: a point which is stressed by many scholars, including Fricker.

Thus, where one's condition leads to deficits in understanding, one is, almost by definition, deficient in some epistemic capacities. Limitations in understanding, combined with other deficits in rational capacities, may lead to severely diminished epistemic performance. Indeed, we must recognise that the epistemic deficiencies associated with mental illness are, on occasion, accurate (although of course, some will not be). Some judgements of testimonial or epistemic deficiency will be undue, and so constitute an epistemic injustice. Others will not be, and so will not. Establishing this will often be a complicated matter, and one fraught with difficulty. Yet, what it clear is that combatting epistemic injustice will be a more complicated matter than it is for gender and race: we must combat injustice whilst being cognisant of the epistemic limitations which can accompany at least some mental illnesses.

VII. The Role of Advocacy

Thus, combatting epistemic injustice is not a matter of assuming epistemic peerhood (as it is for gender and race). Rather, preventing epistemic injustice for mental illness will involve a balancing act between ensuring that we do not make *a priori* or undue attributions of epistemic deficiency (either in terms of an individual's rational capacities, understanding or testimony) on the basis of stereotype, whilst also acknowledging that some people with mental illnesses have genuine limitations in these capacities. Any strategy for combatting epistemic injustice must be able to support (whilst not needlessly interfering with) those individuals who have mental health issues but few epistemic limitations, yet also those whose limitations are more severe. It must also be able to cope with diverse kinds of limitations.

I suggest that advocacy is likely to be an appropriate strategy for this task. The charity Mind (2015, no pagination) describes advocacy as "getting support from another person to help you express your views and wishes, and to help make sure your voice is heard". An advocate can assist mental health service users in a variety of tasks. For instance, advocates can help service users to work out what they want to achieve in a meeting (for instance, with the psychiatrist, a GP, a social worker, or in a benefits hearing) and assist them in finding a way in which they can successfully achieve that. Service users can express concerns that they are not being listened to or that they are being dismissed. Advocates are then able to support the service user, and can either speak for them, query things on their behalf in meetings or interject where they feel that the service user is not being granted due respect or adequate time in which to express themselves. Depending upon the confidence and competency of the service user, advocates can be more or less involved in the degree of support they offer. For this reason, I suggest it may be a useful strategy by which epistemic injustice may be fought whilst attending to the epistemic limitations and differences of some service users.

How can advocacy help to combat epistemic injustice? Having an advocate present in the room can give service users the confidence to express themselves, and an advocate can back the service user up where they feel they are not being listened to. Having someone on their side empowers service users, and either they or the advocate can speak out where they feel that epistemic injustice is being committed. Indeed, service users may feel that they can speak out in the presence of an advocate where they did not feel able, or confident enough, to do this alone. Given that advocates can play a vital role in the identification and protest of epistemic injustice, it would perhaps be prudent to integrate a brief education programme about epistemic injustice and its various forms and manifestations into advocacy training. For instance, advocates could be provided with examples of real-life cases of epistemic injustice which they

may come across in their practice. This should be possible, given that advocacy training is often quite lengthy, and there is time in which this could be done.

Yet, advocacy is also capable of recognising epistemic deficiency and making sure that it does not disadvantage the service user. It acknowledges that some service users do struggle to understand what is being said to them, and equally struggle to communicate their own experiences in a clear, ordered and concise way. As such, in many cases the role of advocacy is to facilitate understanding. Advocates help service users to access information, support them to query things that they do not understand, and can explain information or procedures to them. Often, pre-meetings are held in which service users can discuss with their advocate what they want to get out of the next meeting, what they want to say, what they are concerned about, and how they want the next meeting to proceed.

By working through concerns and getting their thoughts together with an advocate, service users have an opportunity to decide on the aims of the cooperative exchange and can work through any distress, cognitive interference or anxiety with their advocate before attending a GP appointment, for example. This gives them a chance to better structure their contribution to the exchange with the GP by thinking about potential problems beforehand, and can often make service users calmer when they attend the appointment. There is a great deal of anecdotal evidence from service users that this improves their experience of the GP meeting. Advocacy (particularly when pre-meetings are held) helps to make it easier for service users to understand the purpose of the cooperative exchange and what is conveyed in the exchange. It also allows service users to plan their contributions so that they are relevant, carefully thought out and well-structured. Thus, where deficits of understanding obtain, well-conducted advocacy can

ensure that the service user is not disadvantaged and does not have a poor experience, whilst still combatting epistemic injustice. Hence, whilst advocacy services are available, I might tentatively suggest that more funding should be allocated to this area. More people should be made aware of the possibility for advocacy, and advocates themselves should be given robust training.

VIII. Conclusion

In this chapter I have explored a phenomenon which I believe is a large part of the stigma of mental illness: epistemic injustice. The mentally ill are often experience status loss. They are disempowered in social relationships, and one way in which this commonly occurs is in their capacity as epistemic agents. Some of the material in this chapter has links to chapter 2, in which I noted that the disempowerment of mentally ill people may be problematic as it removes them from the decision-making process (either regarding their own lives or about those of others). This is concerning because it exacerbates the risk of implicit stigma occurring. Further to this, a culture of testimonial injustice makes it impossible for mentally ill people to practice meaningful criticism of practices they do not endorse: for instance, where mistreatment or abuse occur. In this way, epistemic status loss may play a profound role in implicit stigma.

To explore this further, in this chapter I outlined Fricker's (2007) account of testimonial injustice. I argued that the mentally ill constitute a group which suffer identity-prejudicial credibility deficits, and provided examples of this. I then outlined Goguen's account of stereotype threat, before discussing how the mentally ill not only likely experience stereotype threat, but that it may be particularly profound given the common stereotypes of rational

deficiency. I also noted that the mentally ill were liable to experience severe epistemic spillover due to the myriad cases in which epistemic doubt (both justified and not) could be engendered within them. I suggested that stereotype threat may explicate a range of phenomena and behaviour encountered in mentally ill people and their interaction with others and services.

Finally, I noted that our strategies for combatting epistemic injustice in mental illness are likely to differ from those used in gender and race. For one, the stereotypes of rational deficiency can occasionally be accurate, and so cannot be dispensed with altogether. There are genuine epistemic limitations associated with at least some mental illnesses, and so combatting epistemic injustice is not simply a matter of establishing epistemic peerhood. Rather, a more nuanced strategy must be adopted. I have suggested a few strategies we might use to combat epistemic injustice in mental illness, including refraining from making assessments of epistemic credibility where possible, where necessary making these on a case-by-case basis, and being responsible with our application of stereotype. I concluded by suggesting that advocacy may be a suitable means of both avoiding overgeneralisation and fighting epistemic injustice, whilst at the same time appreciating and factoring in many of the epistemic limitations associated with suffering from at least some severe mental illnesses.

CHAPTER FOUR

Combatting Stigma: Labels, Stereotype and Language

I. Introduction

I noted in chapter 1 that this thesis was to be concerned with applying prominent debates in philosophy to the problem of stigma and mental illness. This chapter will focus not on aspects of stigma itself, but rather, upon ways in which we might seek to end stigma. In doing so, it brings in a range of debates within the philosophy of science and the philosophy of language. I will begin by outlining possible interventions we might hope to make on the stigma process.

As a first candidate, I will consider whether we might plausibly hope to end the stigma of mental illness by attempting to prevent labelling (the identification and naming of socially relevant forms of human difference) from occurring. However, I raise a few concerns about this project: some of which apply to other arenas in which stigma is found, and others which are distinctive of mental illness. As regards the latter, I suggest that many of the labels we use in common discourse to demarcate mental illness also furnish our systems of classification in psychiatry. Getting rid of labels for mental illness would involve not only getting rid of terms like ‘crazy’ and ‘mad’, but also potentially things like ‘schizophrenia’: for both are ways of identifying forms of social difference, marking them as relevant and affixing a language marker

to them. Yet, I suggest that psychiatry, as a scientific and a therapeutic discipline, has need of these labels. The very same labels are also beneficial to patients. As such, I suggest that it would be unwise- perhaps even impossible- to prevent the stigma of mental illness by intervening to prevent labelling.

I then move on to consider whether we would be better intervening on the stigma mechanism at the level of stereotyping. I consider the dilemma raised by scholars such as Gendler (2008, 2011) and Egan (2011): that stereotypes bring about a kind of ethical-epistemic dilemma. That is, many stereotypes generate ethical harms, but if we were to dismiss potentially relevant information, this constitutes an epistemic harm. I consider Madva's (2016) proposed solution to this dilemma, and suggest that the strategy he offers seems plausible, but will prove much harder to implement for mental illness than it will for race and gender.

II. The Stigma Mechanism: Possible Interventions

If we are to talk about ways in which we might hope to prevent stigma, it would be useful to remind ourselves once again what the process looks like. 'Stigma' is commonly used to describe the culmination of a series of different processes which can often be confused with one another. For Corrigan, stigma occurs where social differences are labelled, stereotypes are attached to those labels, the stereotypes are endorsed via prejudice and discrimination occurs as a result. For Link & Phelan (2001, p.367), stigma occurs "when elements of labelling, stereotyping, separation, status loss and discrimination co-occur in a power situation that allows the components of stigma to unfold".

As stigma is understood to be constituted by the co-existence of all these stages, it seems plausible that if one were to prevent any one of these processes, stigma would not occur. Hence, one rather broad strategy for combatting the stigma of mental illness may be to intervene upon one of its component stages. The stigma mechanism should not be simplistically understood as a chronological process. That is, whilst some stages of the process are undeniably antecedent to others, the processes mutually support one another and can occur simultaneously. Yet, there is somewhat of an order to it.

For instance, it is difficult to see how one could develop prejudice (endorsement of stereotype) if there were no stereotypes, or if one were not aware of them. Furthermore, if there are no labels to mark relevant forms of difference, then stereotypes cannot be attached to them. The hope here would be that if the cognitive-affective aspects of the stigma process do not occur, then the behavioural response which is presumably based upon them (discrimination) will not result. The idea behind challenging labelling and stereotyping seems promising.

Yet, there is another reason that labelling seems to be a good candidate for intervention: namely, that it simply appears to be an easier place to intervene than other parts of the process would be. *Prima Facie*, it is a more daunting task to challenge prejudice and discrimination. Prejudice occurs where the individual endorses the content of a stereotype, and discrimination is the behavioural result of many complicated cognitive-affective processes. Attempting to intervene on the things individuals endorse (or attempting to alter the ways in which endorsement works for each individual) is likely to be a complex and challenging process. Likewise, intervening to change behaviour will be difficult as it is guided by a plethora of factors. Intervening on both prejudice and discrimination is likely to be made more problematic

by the fact that these stages may occur implicitly (as described in chapter 2[36]). Labelling and stereotype (whilst still undeniably complex processes) appear to be a little less unwieldy. Thus, I will examine the utility of intervening on these stages in this work, starting with labelling.

How can we flesh out the suggestion that stigma in mental illness may be combatted by preventing labelling from occurring? On both Corrigan and Link & Phelan's models, labelling (construed as the identification of socially meaningful forms of difference) constitutes the first stage of the stigma mechanism. Thus, frustrating this process might simply mean that we refuse to participate in the practice of distinguishing forms of difference, or refusing to accept that any forms of difference identified are socially meaningful. Suggestions of this kind are occasionally espoused in the case of racial stigma, and can be encapsulated in the claim 'the best way to deal with racism is to do away with race altogether'. This is not to say that one cannot identify difference at all: rather, it is to say that it is only when we cease to consider race to be a *socially relevant* or *legitimate* form of difference will racial discrimination cease. The thought here is that abandoning traditional modes of distinguishing categories of people from one another will prevent stereotype, prejudice and discrimination. If the categories are not drawn, or are not considered legitimate, then people will not be sorted into them. They then cannot be labelled, and if the category is not delineated, it cannot have knowledge structures attached to it and so on. Simply put, if there is no race, there can be no racism and no racial stigma as there is no group to be subjected to these processes.

[36] Specifically, interventions will prove difficult if individuals are not introspectively aware of their implicit endorsement of stereotype and their acting upon it.

This strategy is inextricably linked to language. When we form categories by identifying socially relevant forms of difference, we generally designate a label in our language to encapsulate this process. In many senses, the label becomes the group, and constitutes the act of categorisation itself. Thus, when attempting to cease labelling, one is likely to be concerned not only with preventing the creation of new labels, but also with removing the labels traditionally used to designate differences. For instance, in academic circles, Paul Gilroy (2002) has argued for the renunciation of race categories and the language of race: his reasoning being that whilst race certainly has substantive effects on the material lives of great many people, it is essentially little more than a social construct designed for subjugation. Rather than focussing on 'race', we must, he suggests, move to a new political language of anti-racism. Thus, the project of eradicating racism has the modification of language at its heart.

Yet will this kind of strategy work for mental illness? I want to begin by noting that I am unsure that it is even desirable for categories such as race and gender[37]. In many cases, we may want to claim that it is true that racial categories have been granted undue significance, and have been affixed with many damaging and erroneous stereotypes. Race categories cannot be used to predict someone's level of criminality or their intelligence. Yet, whilst acknowledging this, do we really want to claim that race is not a relevant form of social difference, and that the category (and the language associated with it) should be done away with? For one, maintaining the category of race gives us appropriate language with which to discuss past injustices. If we do away with racial terms and the associated stereotypes altogether, then this may frustrate our ability to identify and talk about significant aspects of racism, and perhaps even identity itself. Were this language to be removed, we may encounter a situation not unlike Fricker's case of

[37] Much has been written in the Philosophy of Race (see Mallon 2006, 2007, Zack 2002, James 2017). I will not address it here, as my task here is merely to compare the situations for race, gender and mental illness.

hermeneutical injustice: where individuals lack the vocabulary with which to describe their experiences.

Further, whilst categorisation can be imposed upon a group, the group can come to (either consciously or unconsciously) accept this identity, and the attitudes and behaviours of those within the category come to strengthen it, or to re-define the boundaries of the category on their own terms. I have in mind here something akin to Ian Hacking's (1991, 1998, 1999) work on 'looping kinds' in which he notes that the category 'Homosexual' has been continually re-shaped by the culture, behaviour and beliefs of those sorted into it. Certain kinds, Hacking argues, are 'human' rather than 'natural' in that they admit of distinctive 'looping' effects in which the category is created by the work of social scientists, and this very act of categorisation changes those classified, whilst at the same time those classified can come to change the category. I do not mean to espouse a commitment to Hacking's theory here, but rather want to make the more general point that even labels and categories which are imposed on groups and are associated with problematic stereotypes can be adopted by that same community, and may even become critical to their sense of identity itself.

Crudely put, there is a great deal of anecdotal evidence to suggest that for many black people, their being a part of the category 'Black' is socially relevant for them- either because it gives them a heuristic and lexicon with which to understand their experience, or it may be inextricably linked to aspects of culture, identity, socio-political status and history which is valuable to them. Hence, even in the case of race, I am dubious about the project of intervening on labelling to halt stigma- the reason being that it may lead to a poverty of suitable language with which to challenge and talk about racial injustice, whilst simultaneously cutting off access

to aspects of identity which are inextricably (or at least profoundly) connected with the existence and lexicon of the category.

This strategy is even more problematic for mental illness. For instance, which labels would we intervene on? Presumably, all those which pick mental illness out as a relevant form of social difference, and so function as points to which problematic stereotypes can be affixed. Some labels, such as 'crazy', 'mad' or 'psycho' appear to do this, but so do our diagnostic labels- those numbered in DSM-V and ICD-10. I take it to be relatively straightforward that it would be beneficial to get rid of labels like 'crazy' and 'mad'. Indeed, this has been widely explored. Yet, even if we do this, we still have other ways of marking out mental illness as a relevant form of social difference: our diagnostic labels. But should we also get rid of these if we hope to tackle stigma?

This seems far more complicated. The labels used to designate mental illnesses- 'schizophrenia', 'anxiety' and even 'mental illness' itself- appear to track genuine and appropriate forms of difference[38], yet the very same labels are also associated with problematic stereotypes and so are part of the stigma process. The issue here is that mental illness labels are used in great many domains, by many different parties, for many different purposes. We demarcate mental illness for many reasons (some appropriate, others not), yet the labels we use to do so are generally the *same*. The member of the public who stigmatizes people with mental illnesses may use the same diagnostic label that a psychiatrist might, or someone who has that

[38] It is worth noting that there are some who would deny this. Indeed, the anti-psychiatrists (Szasz 1961, Laing 1960) deny that mental illnesses exist at all, often claiming that the act of categorising mental illnesses represents little more than an attempt to pathologize divergence from social norms, to medicalize the notion of normality and to effect social control. In this sense, some would deny that there is any difference worthy of a medical label present at all. I will not discuss this here: indeed, it will take it for granted that mental health labels do sometimes track genuine medical conditions rather than representing efforts to control a population.

mental illness. Of course, each party may understand different things by this term. For instance, a medical professional may understand the label in clinical terms, whilst a member of the public may have an ordinary language meaning of the term. Nonetheless, the psychiatrist attempting to label a patient for therapeutic and pragmatic purposes uses the same label- 'schizophrenia'- as the naïve member of the public who thinks that mentally ill people are dangerous. The former seems to be an acceptable use of the label, whilst the latter is not.

It is because the clinical and the everyday language used to label mental illness are, to a large degree, the same that the strategy of getting rid of labels (that is, labels like 'depression' and 'schizophrenia') to challenge stigma becomes problematic. Simply put, getting rid of the labels which attract stigma would also be to get rid of the labels which are used by clinicians, researchers, patients and members of the public who do not stigmatize. As I will now explore, labels provide a plethora of benefits to many different parties, and so I think it would be unwise if we were to intervene on the stigma mechanism by attempting to do away with mental illness labels.

III. Benefits of Labelling

i. Induction, Prediction and Diagnosis

Labelling (and the act of categorisation it represents) provides many benefits. Hence, there is some reason to retain labels, even if they can also be affixed with problematic knowledge structures. As I will demonstrate in the following sections, many of the tasks clinicians and medical professionals need to undertake are either made possible or expedited by the existence

of a system of classification furnished by labels. One of the primary benefits of labels is that they can be used to facilitate induction and prediction Indeed, there are many people working in mental health who need to make inferences and predictions: for instance, those working in psychiatric research and those who treat patients. Yet, whilst it is mostly medical professionals who benefit from the inductive and predictive potential of labels, it is worth noting that members of the public and people with mental illnesses are also able to access these benefits.

But how does classification facilitate induction, prediction and explanation? Categorisation is concerned with classifying entities and establishing groups based upon perceived or actual similarity. Classificatory systems track the underlying laws or relations governing or merely connecting the entities within a field of study. This allows a great deal of information to be rendered coherent or accessible to us. However, different classificatory schemas may facilitate these processes in a more robust, or less fallible manner. In order to better illustrate this, it would be prudent to briefly examine a prominent classificatory system in a prototypical science: the periodic table. The periodic table is a tabular structure in which the known elements are arranged according to increasing atomic number in rows, such that each vertical column forms a group.

For instance, the category 'Argon' is characterised as possessing atomic number 18: all putative category members must possess this property if they are to belong to the category. Significantly, this 'category-standard' characteristic is associated with several other concomitant features: namely, being remarkably unreactive and exhibiting little tendency to participate in reactions or form compounds, unless under extreme conditions. Here, the relationship between category membership and concomitant feature is akin to a law of nature

or a law-like regularity. This is because elements which possess atomic number 18 admit of a valence electron configuration which is responsible for its chemical properties. Thus, the category-standard characteristic (and thus membership in the category) *explains* the presence of the concomitant feature.

Further, we can also make *predictions* or *inductive inferences* about the behaviour of novel entities based on their category membership. If I know that substance *x* belongs to the category 'Argon', then I can soundly infer that it will have atomic number 18, and thus instantiate the relevant concomitant features. Without having to test it, I know that this substance will be unreactive, and thus suitable for usage as a 'shield gas' in welding, or other such things. By simply knowing that *x* is argon, I immediately know which chemical properties it will possess. We need not begin each investigation afresh: rather, we can utilise the shortcuts for understanding permitted by classification. Similarly, I can make some reasonable, yet not infallible, inferences about the category membership of *x* by comparing it to other substances, whose category I already know. For instance, I might observe that *x* has of many of the same observable features as *y* (which I know to belong to the category 'Argon'), and so infer that *x* may also be a noble gas, albeit perhaps not necessarily argon.

Thus, it is easy to see why classification and labelling can be so useful. However, we should not expect classification in psychiatry to be quite so robust: indeed, we must allow that our inferences and predictions will be more fallible. Why is this? The periodic table is a *monothetic* classificatory system. It specifies necessary and sufficient conditions for kind membership, or for the application of a concept. It outlines qualities which are possessed by all members of a class, and only by those members. Any inferences or predictions we make using these

categories are particularly robust because we can assume that all members of the category will behave similarly (for they all possess the same necessary features which are strongly linked to certain concomitant features), and can be properly distinguished from relevant alternatives. Indeed, because the relationship between the necessary property and concomitant features are often grounded by laws of nature, we can be confident that the possession of the former will entail the presentation of the latter in all cases.

Yet few argue that psychiatry can be fitted with a monothetic classificatory system. Many commentators have argued that mental illnesses are not natural kinds: it is not possible to specify necessary and sufficient conditions for the application of diagnostic concepts, and category membership does not entail the presence of the concomitant features as a matter of law-like regularity (see Margolis 1994, Zachar 2000, 2008, 2014). Many claim that mental illness labels are historically contingent (Kutchins & Kirk 1997, Hacking 1998, Greenberg 2013) and appear to come in and out of existence, and so they (and the disorders they purport to describe) are social constructions, and do not reflect any underlying divisions in reality: they do not 'carve nature at the joints', as Plato (1925) once put it.

Indeed, DSM-V (APA 2013 p.160), specifies that a patient qualifies for membership in the diagnostic category Major Depressive Disorder if "five (or more) of the following symptoms have been present within the same two-week period". There are no necessary conditions for category membership, nor is the presence of any one symptom individually sufficient for the application of the disease concept. Rather, the combination of the specified number of symptoms is held to be jointly sufficient. The DSM is not monothetic, but rather, *polythetic*: categories are formed based on commonly shared traits, yet no trait is essential for group

membership[39]. Two patients could both fulfil the diagnostic requirements for 'Major Depressive Disorder', yet they may not share any one symptom. Furthermore, symptoms are rarely exclusive to one diagnostic category. Whilst it is most commonly associated with schizophrenia, catatonia "can occur in several disorders, including neurodevelopmental, psychotic, bipolar, depressive, and other mental disorders (APA 2013 p.89).

As such, category membership in a mental illness category is not as informative as it is in the case of the chemical elements. The fact that patient *a* and patient *b* belong to the same diagnostic category does *not* entail that they will demonstrate the same characteristics. There is no one property that *a* and *b* must share, and they may demonstrate a different set of jointly sufficient characteristics. Thus, members of the same diagnostic category may be quite different. Further, the association between category membership and concomitant features is merely statistical, rather than law-like. If one is a member of the diagnostic category 'Major Depressive Disorder', one is statistically likely to experience a lack of motivation, but one need not necessarily.

Whilst the DSM has been subjected to a huge array of criticism (see Kutchins & Kirk 1997, Greenberg 2013), few deny that psychiatry tends to resist being furnished with a monothetic classificatory system. Yet, whilst any explanations, inductions and predictions we form using psychiatric labels will be less robust than their analogues in the prototypical sciences, this is not to say that they are no use at all. Whilst anything we do with a polythetic classification system will be more tentative or fallible, we are still able to form explanations, predictions and

[39] The concept of polythetic categorisation is connected to Wittgenstein's discussion of games and family resemblance.

inferences which are hugely useful. That is, even where a classificatory schema does not specify necessary and sufficient conditions, it remains hugely useful for that field of study.

In many cases, the main benefit of classification and the labels which furnish it is establishing and permitting the use of certain epistemic shortcuts. For instance, consider the clinical context in psychiatry, in which the practitioner is faced with a wealth of information. Practitioners encounter many patients, disorders and treatments, which would be hugely time-consuming, if not impossible, to consider on a case by case basis without relying on some prior knowledge (although of course, there is a sense in which every patient is, and should be, considered separately). Classification allows the practitioner to make use of shortcuts for understanding, or to gain insights about patients quickly on the basis of their category membership. It can also be invaluable in facilitating the identification of salient information, or in focussing the enquiry. Classification directs attention to relevant similarities pertaining between category members, and the properties they are *likely* to instantiate. If a patient has been diagnosed as suffering from Major Depressive Disorder, I cannot infer with certainty which symptoms they might present with, nor which form of treatment will be most appropriate, but I will be able to have some rough idea of where I might focus my enquiry.

In a similar vein, classification also facilitates prediction. By categorizing patients into diagnostic classes on the basis of some perceived similarity, the clinician can assume that members of this group are "generally homogeneous in the underlying nature of the illness, regardless of whether there is some variability in the presentation of symptoms or circumstances surrounding illness onset" (Garand et.al 2009, p.2). Thus, when the clinician meets someone with a particular diagnosis, she need not 'start afresh' in her enquiry. Rather,

she can make predictions about how the patient's disorder might progress, which treatment might be optimal and what risks the patient is subject to by referring to the kind-typical behaviour of the model category: that is, by looking at what has happened to similar patients. This allows the clinician to comprehend a vast array of information with relative simplicity. Hence, the process of taking a clinical history, understanding the nature of the patient's disorder and contemplating options for therapy is streamlined: hopefully minimizing the period for which the patient is in distress and allowing for quicker, more effective treatment. Classification therefore admits of a great deal of what we might term *clinical utility*, even where the classificatory system is polythetic. Hence, even where the categories cannot be characterised in terms of necessary and sufficient conditions (and thus do not meet the strict requirements set out by many metaphysicians for natural kindhood), they remain epistemically illuminating.

Indeed, one might plausibly argue that based on the above, classification is not merely useful to psychiatry, but rather, it is virtually indispensable in that it helps it to fulfil its practical aims. This is most clearly evidenced when one considers the institution of diagnosis, which could not proceed in the absence of labels and categorisation. What is diagnosis? In modern medicine, psychiatry included, the process of diagnosis is essentially the act of diagnostic labelling. That is, a clinician examines a patient's symptoms and 'matches' them with known paradigmatic disorders. For instance, the clinical interview may establish that the patient displays behaviours *x, y* and *z.* The clinician then refers to disease categories (nosologic categories), as established by specialists in the field. A disorder Φ may be characterised by the presentation of behaviours *w, x, y* and *z*. The medical professional then makes a judgement as to whether the behaviours demonstrated by the patient are sufficiently similar to those associated with Φ such that the patient might be plausibly characterised as being a member of Φ, and thus as suffering from

that particular kind of disorder. In this sense, the nosologic categories become diagnostic categories when patients are sorted into them.

Diagnosis is often held to be a vital part of therapy. From a clinical point of view, establishing category membership is a vital part of the clinicians' primary epistemic task: namely, to gain an understanding of who their patient is, and what is wrong with them. As explored above, if the category membership of patient *p* is established (imagine, for instance that it is judged that *p* does in fact suffer from Φ), then the clinician can then make predictive and inductive inferences on the basis of *p*'s category membership, and also gain access to the established body of knowledge about disorder Φ: its characteristic features, comorbidities, and common treatments, for instance. In identifying *p* as a member of Φ, one can with some reliability (yet certainly not infallibly) predict that *p* may demonstrate many of the features concomitant with Φ.

In this sense, the understanding of the individual is augmented by the common body of knowledge associated with category, to which the patient is similar enough to be considered a member. Being able to make inferences in this way reduces the cognitive load facing clinicians, expedites treatment and allows clinicians to make use of collective knowledge. Diagnosis is a vital part of psychiatric treatment, and yet it could not be conducted without categorisation and labelling. This, when combined with the explanative, predictive and inductive benefits it brings to psychiatry should give us strong reason to doubt whether the project of intervening on the stigma mechanism by preventing labelling is even feasible, let alone desirable.

ii. Psychiatry as a Scientific Discipline

The existence of a system of categorisation and the labels which demarcate its classes is arguably necessary for psychiatry in another way: namely, that it legitimises it as a form of scientific enquiry, and provides linguistic (and perhaps conceptual) cohesion where pluralism is rife. In this way, labels might plausibly benefit both medical professionals, patients, and indeed, all those touched by the discipline. The need for a classificatory system in psychiatry can be elegantly demonstrated by a brief examination of the historical context leading up to the publication of the first DSM, and so I propose to take a brief departure to examine this. The DSM was created to provide a salve to difficulties which had long plagued psychiatry: namely, the existence of psychiatric pluralism, the failure to attain diagnostic reliability and allegations of superstition.

Psychiatry is virtually unique as a modern science in that to this day there is widespread disagreement about the very nature of the discipline. Unlike medicine and the majority of the life sciences, there is often very little consensus as to the appropriate methodology psychiatry should adopt, and even what the subject of matter of psychiatry is. Indeed, Cooper (2007) argues that this diversity is so substantive that, contra Kuhn (2000), she believes that multiple paradigms exist within the discipline. I do not mean to dispute this here, and so to avoid the conceptual difficulties associated with the simultaneous existence of multiple paradigms, a weaker claim can instead be made: that psychiatry is undeniably pluralist, in a way which in non-psychiatric medicine simply is not.

In her work on scientific anthropology, Luhrmann (2000) notes that one of the most prevalent schisms is between the psychoanalytic and biologically focussed traditions of psychiatry. Such differences are not merely cosmetic disagreements about terminology. Rather, different schools use divergent investigative practices and are concerned with different elements of data collection. Generalizing somewhat, Luhrmann (2000, pp.20-23) found that psychoanalysis is interested in the case studies and histories of particular patients: "the stories behind their patients' lives" as Cooper (2007, p.89) puts it. By contrast, those with a biological orientation were concerned with genetics, hormone imbalances and neurotransmitter levels. Furthermore, individual practitioners often align themselves with a certain orientation (be this Freudian, Sullivanian, or Jungian, behavioural, biological, or humanistic). In their words, psychiatry seems to be imbued with a certain distinctive 'relativity'. It is a series of co-existing, yet often conflicting perspectives: "channels of knowledge which reveal certain aspects of patients while obscuring others" (McHugh & Slavney 1983, p.3). Indeed, this pluralism is not merely methodological. Indeed, as Parnas & Sass (2008, p.245) argue, psychiatry also tends to diverge across national and linguistic boundaries.

Unsurprisingly, this lack of a unified theoretical approach had profound implications for the task of *diagnostic reliability*: the accuracy with which different practitioners attribute the same diagnosis to the very same patient, or a patient with identical symptoms. Before the DSM was created, the process of demarcating diagnostic labels and attributing them to patients was conducted entirely according to the clinical intuition of the individual physician. However, there was widespread disagreement as to how this should be done. This failure to attain consensus became particularly injurious in the period following the Second World War, and began to severely undermine the reputability of the practice as a whole (Kutchins & Kirk 1997). Indeed, in the late 1960s a series of studies conducted by the WHO demonstrated an alarming

lack of consistency and reliability between diagnoses made in the US and in the UK (Cooper et. al 1972). Many similar studies were conducted, each of which produced equally damning results. This divergence was taken as evidence that psychiatry could not reliably diagnose mental illness, and that the praxis was disreputable.

Furthermore, as Johnstone (2000) notes, it was thought that a universal classificatory system was vital if psychiatry was ever to be considered a 'science', and to absolve itself of allegations of mysticism, superstition and ambiguity. Indeed, classification is commonly thought to be necessary for an area of enquiry to be a science. Just as our understanding of zoology is structured by the current taxonomy of species, so too should psychiatry be conducted and understood according to a classificatory system. As Kendler & Zachar (2008, p. 371) note:

> Whether we are doing epidemiologic studies, searching for genes, tracing neural circuits, or describing the pain and suffering of psychiatric patients, our results are typically organized and communicated through diagnostic categories.

Categorization and the act of labelling are often held to signal an end to the pre-scientific stage of investigation, in which different, and often competing paradigms are adhered to, and in which there is little stability or consensus within the community. They do this by providing a system in which the meaning or referents of various terms can be fixed or decided upon. In this way, classification provides a common language of fixed meaning, which allows for clearer and more precise communication between those involved (either closely or indirectly) in the discipline. Prior to the advent of the DSM, psychiatry lacked a universal classificatory system, and so was considered to be 'unscientific'.

Thus, what becomes obvious is that there was a real need within psychiatry to establish a classificatory system which could meet these challenges- pluralism, poor diagnostic reliability and lack of scientific credibility. It would be the lack of theoretical unity, combined with the relative failure to establish the kind of clinical/pathological relations found within medicine which would come to shape the resulting classification system (the DSM) most profoundly (Wallace 1994, p.31). Indeed, whilst medicine initially classified diseases by looking at their observable features (symptoms), the praxis had since identified etiological pathways for many of its diagnostic categories. This project had not been successful in psychiatry.

Given all this, one means of attaining the cohesion much needed by the discipline of psychiatry arose: establishing an *atheoretical* and *purely descriptive* classificatory schema (APA 1980, p.7). The DSM made no reference to a theoretical, methodological or etiological outlook of any kind. Rather, it was constructed according to, and was a compendium of, existing 'clinical wisdom': the *shared* wealth of knowledge accumulated by practicing psychiatrists (often gained through patient observation- a method available to all). The DSM was intended to cement these shared clinical intuitions into a manual, which could be utilised to enhance agreement within psychiatry. This is known as psychiatric operationalism, which as Parnas & Sass (2008, p.247) articulate it, amounts to the claim that "descriptions of mental or subjective phenomena should be cast at the '*lowest possible level of inference'*- that is, ideally in *external behavioural* description, or else in *simple lay language*".

Psychiatric operationalism is profoundly influenced by the work of Hempel (1965, p.123) on operational definitions:

> An operational definition of a term is conceived as a rule to the effect that the term is to apply to a particular case if the performance of a specified operation in that case yields a certain characteristic result.

Operational definitions specify ways in which we might perform an empirical check on that concept: certain objective, and crucially public, criteria are outlined for the application of a concept, which any investigator could use to ascertain whether the concept applies. One might question why it is that the criteria provided for operational definitions need be *public* in this sense. The answer is that it is neither possible nor desirable that *all* of the concepts we utilise in any definitional context need be operationally specified. Indeed, Hempel (1994, p.321) warned that such a state of affairs would result in an infinite regress. As such, it is necessary that some of our concepts are antecedently specified. Thus, certain aspects are chosen as criteria for an operational definition in virtue of their being public and easily observed, and hence simple to define beforehand[40].

Operational definitions provide a common usage for a term, and hope to ensure that everyone who utilises the concept does so in the same way. Hempel held that when public and definite criteria of application for a concept was established, then it was possible to subject that concept, and the use to which it is put in a discipline, to scientific or objective scrutiny. Crucially, this allows for the establishment of scientific objectivity, in the sense of an *inter-subjective accord* or *consensus* being reached. To see how operationalism works in psychiatry, it is useful to consult the operationalised diagnostic criteria of Panic Disorder (APA 2013, p.246):

[40] An operational definition is often a "partial criterion of application": it will not, and is not intended to, provide a full definition for the application of a concept (Hempel 1994, p.320).

A. Recurrent unexpected panic attacks. A panic attack is an abrupt surge of intense fear or intense discomfort that reaches a peak within minutes, and during which time four (or more) of the following symptoms occur;

Note: The abrupt surge can occur from a calm state or an anxious state.

1. Palpitations, pounding heart, or accelerated heart rate.
2. Sweating.
3. Trembling or shaking.
4. Sensations of shortness of breath or smothering.
5. Feelings of choking.
6. Chest pain or discomfort.
7. Nausea or abdominal distress.
8. Feeling dizzy, unsteady, light-headed, or faint.
9. Chills or heat sensations.
10. Paresthesias (numbness or tingling sensations).
11. Derealization (feelings of unreality) or depersonalization (being detached from oneself).
12. Fear of losing control or "going crazy."
13. Fear of dying.

B. At least one of the attacks has been followed by 1 month (or more) of one or both of the following:

1. Persistent concern or worry about additional panic attacks or their consequences (e.g., losing control, having a heart attack, "going crazy").
2. A significant maladaptive change in behaviour related to the attacks (e.g., behaviours designed to avoid having panic attacks, such as avoidance of exercise or unfamiliar situations).

C. The disturbance is not attributable to the physiological effects of a substance (e.g., a drug of abuse, a medication) or another medical condition (e.g., hyperthyroidism, cardiopulmonary disorders).

> D. The disturbance is not better explained by another mental disorder (e.g., the panic attacks do not occur only in response to feared social situations, as in social anxiety disorder: in response to circumscribed phobic objects or situations, as in specific phobia: in response to obsessions, as in obsessive-compulsive disorder: in response to reminders of traumatic events, as in posttraumatic stress disorder: or in response to separation from attachment figures, as in separation anxiety disorder).

Conditions A, B, C and D must be met in order for the diagnosis of Panic Disorder to be made. In general, in the DSM a diagnosis is made when the patient meets a sufficient number of the operational criteria (for further details see Klerman 1980).

The DSM sought to resolve pluralism by providing a 'common tongue' with which practitioners with often vastly divergent theoretical views could converse (and indeed, which could be used by those with no professional medical training). Cooper (2007, p.94) has remarked that the DSM functions as what Galison (1997) terms a 'contact language': a constructed dialect, which allows for unified action between parties in disagreement over fundamental issues. A contact language is established to facilitate practical action (trade, for instance) and varies in complexity in accordance with the intricacy of the linguistic need.

Galison suggests that trade languages develop even where the parties involved have radically different conceptions of the value, or even the nature, of the item traded: significant divergence does not always constitute an obstacle to co-ordination. In much the same way, scientists can coordinate on a single project, even where they have radically different conceptions of the

nature, function, or even overriding principles appropriate to the entities under study and to the discipline itself. Collaboration can occur even where there is no complete consensus in thought. Indeed, all that is needed is the desire and ability to participate in the activity: "what is crucial is the local context of the trading zone, *despite* the differences" (Galison 1997, p.803). According to Cooper, the DSM is one such contact language: possible in virtue of the fact different schools of psychiatry need to communicate with one another. It facilitates joint action by unifying psychiatry: it calls upon the vastly different parties involved (psychiatrists, healthcare workers, members of different professions) to *lay aside* most of their own theoretical commitments or worldview so that productive discussion can proceed. This enforced cohesion was thought to improve diagnostic reliability[41] and thus lend credence to the discipline.

Psychiatry continues to be beset by theoretical pluralism, and yet strives to reliably diagnose patients and maintain legitimacy as a discipline. At present, it utilises its system of classification to provide the cohesion necessary to accomplish both tasks. Whilst I acknowledge that the DSM has been widely criticised (see Kutchins & Kirk 1997, Greenberg 2013, Frances 2012, 2013), the analysis above nonetheless demonstrates how crucial a system of classification (whatever it may be) furnished with labels is to psychiatry at present. Given the undeniable and continuing reality of psychiatric pluralism, classification functions as a contact language, which itself is necessary if the discipline is to gain respect and accomplish key tasks such as reliable diagnosis. Thus, given the integral role classification plays for psychiatry, it seems inadvisable at best to do away with the labels that furnish it- even if these labels can contribute to stigma by acting as points to which problematic stereotypes can be attached.

[41] It is worth noting that the reliability of the DSM has been hotly debated. For a good overview of the reliability of DSM-V, see Cooper (2014).

iii. Benefits for Patients

Whilst labels undeniably play a role in the stigma process, it is significant that they can also provide huge benefits to those given them: patients. For this reason, we may be wary of getting rid of them. In chapter 1, I outlined many of the harms which can accompany being labelled as mentally ill. These harms were mostly concerned with either public or self-stigma. However, before making the case for the benefits of labelling, it is worth noting that labelling can generate undesirable effects or harms which are not aspects of stigma. For instance, there is some concern that in being diagnosed, patients may come to understand that their condition is an inevitable, or *essential* fact about themselves: a self-fulfilling prophecy.

Some patients may feel that their diagnosis dooms them. In being diagnosed by a professional, the patient is made aware that there is something wrong with her. She, understandably, comes to believe that she is ill (where she may not have previously). This belief may then guide her behaviour through a positive feedback loop. Because she believes herself to be ill, she will alter her behaviour accordingly: for instance, she may become pessimistic, form a belief that she cannot escape her symptoms, or think that there is something *fundamentally* wrong with her, to the extent that it cannot be remedied. She may feel forced to adopt the role of the 'sick person' (Huibers & Wessley 2006, p.4). Stress and anxiety may result, yet some aspects of the behaviour described above will also hinder recovery (particularly where one sees illness as an essential fact about oneself).

Yet, despite all this, diagnosis (the receipt of a diagnostic label) is clearly not always negative. Indeed, as Huibers & Wessely (2006, p.1) recognise:

> the act of diagnosis therefore seems to be a trade-off between empowerment, illness validation and group support, contrasted with the risk of diagnosis as self-fulfilling prophecy of non-recovery.

In many cases, labelling provides huge benefits to those affixed with them, and so there is reason to be cautious about the project of getting rid of them, even to combat stigma. For instance, being diagnosed can provide comfort and relief to patients. It reassures the patient that cases like theirs have been encountered before, and suggests that there is likely to be an established therapeutic pathway. Diagnosis reassures the patient that the clinician knows what is wrong with them, and has some idea how to fix it (in most cases). In this sense, it brings comfort and the hope of recovery. It can also serve to improve self-esteem and to 'legitimize' their suffering to others. Some patients will find that they are not believed when they try to tell those around them that they are ill. However, if a formal diagnosis is given, then this doubt may cease and those around them may be more accepting.

In some cases, patients may feel that the diagnosis absolves them from blame: there is a legitimate reason why they have not been coping with the challenges of life, it is not merely a personal failure. Huibers & Wessely (2006, p.4) put forward a particularly good analysis of this phenomenon in their study of patients suffering with Chronic Fatigue Syndrome, in which the creation of the diagnostic category was a vital step towards legitimacy. Indeed, they suggest that it was only through this recognition that the fight against the disease could commence. Thus, diagnosis can provide a socially accepted rationale for the patient's failure to cope with their situation, and thus can bring about a sense of relief or freedom, and a reduction in stress.

Further to this, diagnosis can serve to *empower* the patient (Rüsch et.al 2006a, 2006b). In being diagnosed, the individual is elevated from a state of relative ignorance- of experiencing

distressing symptoms and behaviours which she does not understand- to a state of knowledge and understanding. Diagnosis can give meaning and structure to the patient's pain or discomfort. It reassures her that there is a reason why she feels as she does. In this sense, it may be easier for patients to endure a known pain rather than an unidentified pain. Diagnosis can also allow the patient to access further knowledge. Just as clinicians can make use of the body of knowledge accumulated about a disorder, the patient can learn about her condition by accessing the shortcuts for understanding described above. She can use labels to move from being a novice about psychiatric disorder to someone well informed. For instance, she can read up about her specific condition, rather than having to contemplate a vast array of complex information.

Thus, I hope to have demonstrated that there are huge benefits to using labels to pick out mental illnesses. Labels and classification allow us to explain phenomena and make inferences and predictions. The system of classification in psychiatry plays a vital role in ensuring the cohesion (and thus stability) of the discipline. Finally, labels can have many profound benefits for patients. Thus, labelling ultimately provides benefits for numerous parties involved in psychiatry, other professions, and members of the public. Given the enormous utility of labels within psychiatry, I suggest that although the labels outlined in the DSM-V and ICD-10 can be affixed with problematic stereotypes, we should not attempt to tackle stigma by getting rid of them. They are invaluable to psychiatry in too many ways.

IV. Disregarding Stereotype?

Given that it would be unwise to attempt to do away with labelling mental illnesses, where else might we turn to combat stigma? One plausible next move would simply be to move down the

stigma mechanism, so to speak, and intervene on stereotyping. That is, if we cannot get rid of the labels, perhaps we can do away with the knowledge structures affixed to them? However, as discussed in chapter 3, not all knowledge structures about mental illness will be problematic. Some, like 'people with schizophrenia experience delusions', will describe the clinical reality quite faithfully. These stereotypes should not be dispensed with: to do so would be epistemically irresponsible as we would lose valuable information. For one, we would not be able to make use of these knowledge structures when carrying out explanation, prediction and induction- hindering both the scientific and therapeutic aims of psychiatry. So, the strategy outlined above should be qualified: perhaps we should combat stigma by getting rid of *some* stereotypes?

The natural response here would be to claim that we should attempt to suppress, forget or disregard the negative stereotypes (those which are false or prejudiced), whilst keeping the good ones (those which are accurate). Yet, recent scholarship has questioned whether it is even permissible to forget or fail to acknowledge 'problematic' stereotypes. The challenge is this: although ridding ourselves of inaccurate or prejudiced stereotypes is ethically desirable in that we can challenge stigma and prevent discrimination, disregarding or forgetting any knowledge structure imposes a significant epistemic cost upon us. Indeed, an argument of this kind has been made by Gendler (2008, 2011) and Egan (2011) about implicit bias, who both go as far as claiming that implicit biases place us in a normative dilemma, in which it is impossible to satisfy both our epistemic and moral requirements. As Madva (2016, p.194) puts is, there will be cases in which social categorisation is undeniably relevant, yet ethically problematic. Thus, we must choose between a range of epistemic goods (e.g. knowledge of stereotype, demographic facts etc.) and ethical goods (e.g. treating people fairly, adhering to egalitarian values, treating peoples as individuals etc.). These claims have been influenced by new

empirical investigations, which suggested that *mere knowledge* of stereotype seems to make it more likely that individuals will act in biased ways (for an overview see Madva 2016, p.193), even where they do not endorse the content of said stereotype.

Thus, the aims of ethical action and social knowledge appear to compete with and frustrate one another: knowledge of stereotypes makes us more likely to act in biased ways, and yet simply forgetting everything we know about stereotype and discrimination amounts to the deliberate disregarding of social knowledge. Gendler and Egan do not utilise the broader terminology of stigma, and yet the argument for implicit bias can be extrapolated outwards. They refer to cases in which merely knowing a stereotype leads to biased action, which in turn produces an ethical impetus to disregard or forget the stereotype. I suggest that this ethical impetus exists whether our target is preventing discrimination or preventing stigma in a broader sense. In both cases, what is significant is that there seems to be ethical reason to get rid of the stereotype (either to stop discrimination or to stop stigma), yet there is an epistemic cost incurred in doing so. Thus, the project suggested above (combatting stigma by getting rid of some problematic stereotypes) faces much the same challenge: an alleged normative dilemma between ethical and epistemic aims.

If this is true, what are we to do? When faced with a dilemma of this kind, Egan (2011, p.72) observes that there are three broad strategies for response:

> We can use all the categories unreflectively, and wind up with a bunch of bad stereotype-concordant inferences, judgements, attitudes etc. Alternatively, we can use the categories, but spend a bunch of cognitive resources supressing or immediately excising the bad stereotype-

concordant inferences, judgements, attitudes etc. Finally, we can avoid using the categories, and fail to code up the base rate information.

Madva (2016, p.194) notes that the first response- to use to categories unreflectively- constitutes adherence to the status quo. This is certainly not a desirable response to stigma. The second response, which he terms 'suppression', is believed by the dilemmists to be too cognitively taxing to be a viable strategy: indeed, it would require constant monitoring, and runs the risk of backfiring (see Huebner 2009, Madva 2012). Madva dubs the final kind of strategy to be 'ignorance'. Here he imagines a scenario in which we could somehow bring ourselves to forget all knowledge of stereotype, thus eliminating biased behaviour and aligning our automatic responses with those beliefs we would reflectively endorse. Notably, this is the shape that the strategy suggested above takes. Many dilemmists recognise that this method incurs epistemic costs (we lose information which could be valuable to us), yet some think this cost worth bearing, and tentatively espouse ignorance to be the most palatable strategy.

How does this apply to the issues of stigma and mental illness? It is worth noting that the dilemma outlined above only seems to bite when we consider prejudiced or inaccurate stereotypes about mental illness. For accurate stereotypes- for instance, 'people with schizophrenia experience delusions'- there appears to be little ethical cost incurred in knowing or adhering to them, and considerable epistemic gain in retaining them. Thus, there is no need to consult any of the strategies outlined above: we can merely retain them. Yet, the dilemma arises with prejudiced stereotypes. For instance, we might postulate that one way to combat stigma will be to suppress, disregard or forget stereotypes we know to be inaccurate or prejudiced (for instance, 'schizophrenics are dangerous').

Disregarding the associated stereotype 'schizophrenics are dangerous' has a great deal of ethical merit, yet it imposes epistemic costs in that we lose information which could be valuable to us. Indeed, knowledge of the prejudiced stereotype can arguably constitute an epistemic good. In knowing it we are aware of stereotypes about mental illness, we gain some idea of the kinds of beliefs other people have about mental illness, and so on. Perhaps to fight the stigma of mental illness, it is valuable to know that problematic stereotypes exist, and what their content is. In this way, we are better able to understand the problem we are working against. In this way, the prejudiced stereotype may contribute to social knowledge, even though they undoubtedly contribute to stigma. The dilemma bites, and so we must decide what to do in response.

Once again, the kind of response we mobilise will be more complicated for mental illness than it will be for analogous strategies to combat racism and sexism. As I have explored in chapter 3, few, if any stereotypes about gender and race will be 'accurate' at all, as there is rarely any intrinsic link between category membership and the associated properties. There is nothing about the nature of womanhood that makes one irrational, bad at maths, or sensitive. As such, we can treat all stereotypes in much the same way. Yet, for mental illness, some stereotypes are accurate and others are not. When we contemplate stereotypes about mental illness, our strategy must be sufficiently nuanced that we are able to treat accurate and inaccurate stereotypes differently.

Thus, one response to combat the stigma of mental illness would be to encourage individuals to forget/ disregard problematic stereotypes (accepting the epistemic costs of doing so as some dilemmists do), whilst retaining the accurate knowledge structures. This is a dual ignorance-

status quo strategy, in which individuals are called upon to practice ignorance towards prejudicial stereotypes, yet whilst simultaneously enacting a kind of status quo strategy towards viable stereotypes (although of course, even the use of accurate stereotypes should not be unreflective). Implementing this strategy would certainly have some benefits. We would incur an epistemic loss for prejudicial stereotypes, yet would retain the epistemic benefits of using viable stereotypes. Further to this, we would satisfy our moral aims in that we would get rid of the problematic stereotypes which lead to bias[42], yet retain those which are neutral and do not contribute to discrimination. Of course, the epistemic penalty of practising ignorance towards prejudicial stereotypes remains. Yet, when considering the severity of the harms stigma can generate, we might think that the epistemic loss to be worth bearing, given the considerable ethical gain.

This seems promising, yet may already be too good to be true. It may prove challenging to action, and will likely bring with it further complications. In the first place, ignorance strategies may generally prove difficult to implement. Indeed, how are we to practically go about disregarding or forgetting knowledge structures? This may be a complicated matter, although one potential solution would be to carry out protest strategies, thus reducing the visibility of problematic stereotypes in the public sphere. Disregarding or forgetting stereotype may also be complicated by the existence of implicit bias, particularly given that our awareness of stereotype and our ability to identify whether stereotype guides action may not always be introspectively available to us. It will be difficult to forget something you are not aware you are drawing on or acting upon.

[42] Madva (2016) qualifies this moral gain. That is, in some cases having knowledge of a prejudicial stereotype can facilitate social justice. In knowing the stereotype, we are alerted to common ways in which category members may experience injustice, and have access to a lexicon with which to describe this.

Yet the primary challenge of this dual theory is that it requires individuals to make a judgment about *which* prong of the strategy to implement in cases where they encounter stereotype. The strategy calls upon individuals to practise ignorance with the prejudicial stereotypes, yet simultaneously maintain the status quo where the stereotypes are accurate. One must make a judgement about the nature of the stereotype *before* one can act: individuals must first establish which stereotypes are accurate, and so can be maintained as part of the status quo, and which are not, and so must be ignored. This process is likely to consume a huge amount of cognitive resources, and may be encountered great many times even over the course of a single day- by medical professionals, but also patients and members of the public.

In this way, the advantage of the dual strategy- that it is sufficiently nuanced so as to retain useful stereotypes whilst dismissing bad ones- is also its disadvantage in that attaining this level of nuance will likely create a high cognitive load. Further to this, making the decision at all may only be possible where the mental health literacy of the person making the choice is relatively high: indeed, this dual strategy will be impossible where public health education is poor, or where correct information is not disseminated adequately. Thus, if this strategy is to be palatable at all, it must be accompanied by an extensive programme of public education. The dual ignorance- status quo strategy would certainly have significant benefit, but would likely constitute a heavy cognitive burden, and would only work in an appropriate climate.

Another plausible strategy comes from Madva (2016) who argues that the dilemma described by Gendler and Egan can be resolved through a fourth option- the *regulation* of stereotype. This would involve training ourselves to cognitively access and recall stereotypes where they are relevant, yet ignoring them when they are not. Madva argues that our epistemic goal is

unlikely to be to know *everything*, and as such, it is not immediately clear that the loss of some information will always constitute an epistemic cost. Indeed, he notes that we demonstrate a profound disposition towards categorising people, and tend to shy away from contemplating too much information. The loss or acquisition of information is not, in itself, epistemically good or bad. Rather, it is good or bad relative to one's particular aims and values. In this sense, stereotypes are not worth knowing for their own sake, but only when they meet some further end (Madva 2016, p.199). Indeed, Madva suggests that we are at fault in an epistemic sense only if our knowledge of stereotypes is accessed or activated in the wrong contexts. Mere social knowledge does not lead to biased behaviour: rather, biased behaviour is generated when social knowledge becomes *too* accessible, and we access it even where it is inappropriate (Madva 2016, p.200). As such, he suggests that regulation is the appropriate strategy, and that we must monitor the degree to which we access social knowledge in the form of stereotype.

It seems that this strategy could be applied to the issue of the stigma of mental illness. Perhaps one can access the stereotype 'schizophrenics are dangerous' when thinking about the problem of stigma and how to combat it, but one should take pains that it does not become accessible in other domains of life in which the categorisation is not relevant. In this way, the stereotype can be accessed, but it must be regulated. Yet, it must be acknowledged that once again this strategy would require the investment of great many cognitive resources. What constitutes the wrong context? Clearly, just as with the other strategy, establishing what is, and what is not, an appropriate context in which to access a stereotype may depend upon an individual possessing a sufficient degree of knowledge about mental illness and its associated stereotypes. Further, we might question how we are to regulate the activation of stereotypes, given that we may not always be aware that we are accessing them (see chapter 2). Both establishing when regulation is required and intervening where one is accessing stereotypes inappropriately may be

challenging, and will likely require a great deal of cognitive resources as one must be constantly vigilant. Indeed, Madva (2016, p.204) acknowledges that regulation will be challenging, although he suggests that stereotype accessibility can be reduced by carrying out implementation intentions of the form 'if X happens, I will do Y'.

V. Conclusion

To briefly summarize, in this chapter I have considered two ways of combatting stigma: intervening on labelling, and intervening on stereotype. I noted that many of the labels which demarcate mental illness as socially relevant are utilised by many parties and in many domains. In this way, getting rid of the labels which can be affixed with stereotypes in the stigma process is also to get rid of the labels which are used by medical professionals, researchers, patients and members of the public. I have outlined the virtues of labelling- both in terms of induction, prediction and explanation, to psychiatry as a science and to patients- and in virtue of this, I suggested that it would be inadvisable to intervene on the stigma mechanism at the level of labelling.

I then considered whether it would be advisable to halt stigma by getting rid of stereotypes: specifically, those which are prejudiced or inaccurate. I noted that there was some challenge in doing so, given that it arguably gives rise to a normative dilemma between our ethical and epistemic aims. This dilemma applies to at least some stereotypes about mental illness, although not to those which are accurate. I outlined a potential dual ignorance-status quo strategy, and one response offered by Madva (2016), but acknowledged that due to the variety of stereotypes about mental illness (specifically, whether they were problematic or not),

implementing both strategies would likely incur significant cognitive load, and would likely work only where the public were sufficiently well-informed about mental illness. Both have significant potential to combat stigma, but will prove challenging to implement.

CHAPTER FIVE

Generics and the Stigma of Mental Illness

I. Introduction

This chapter will also focus upon stereotypes, and how they feed into the stigma process. However, it will take a different perspective on the matter, and will focus upon issues in the philosophy of language and the philosophy of cognitive science. Stereotypes, as I have defined them, are knowledge structures, which can be problematic or unproblematic. Knowledge structures are most commonly passed to us and disseminated by us via language and communication. Yet, recent literature has suggested that certain linguistic forms may prove particularly problematic when disseminating knowledge structures in that they may harbour and propagate worrying ideologies in a manner which is often not appreciated by speakers. The linguistic form so often identified, and the one I will discuss here, is generics.

Sections II and III will outline what generics are, and why they are so problematic (namely, that they encourage the psychological essentialism of kinds). To do this, I will briefly summarize the literature on psychological essentialism and the evidence thought to link it to the use of generics. From this point, I will argue that when it comes to mental illness, generics are rarely, if ever, a useful means of disseminating information (section IV). Indeed, I argue

that generics conveying inaccurate knowledge structures can lead to insidious forms of social stereotyping, much as Leslie (2013, 2014) has suggested occurs for race and gender.

Yet, even where the knowledge structures in question have a degree of accuracy, generics remain an unsuitable form in which to express them. The reason for this is that it is deeply unhelpful to think of mental illness categories in essentialist terms. Doing so is likely to generate misinformation, and to encourage incorrect or problematic beliefs about mental illness. I argue that the problematic essentialist beliefs in question concern the severity and stability of mental illnesses, the degree to which the category is thought to permit particularly robust inductions and predictions, the tendency to view the category as unified and distinct from other categories, and the belief that mental illness categories may be sources of contagion.

It is because of the risk of problematic beliefs such as these developing that I suggest mental illnesses should seldom, if ever, be discussed in generic terms. Generics contribute to stigma either by perpetuating social prejudice or by propagating problematic or false beliefs. To combat the stigma of mental illness, we should avoid using generics wherever possible. We should take care to quantify utterances and challenge generics wherever we encounter them (section V).

II. What are Generics?

At the outset, it would be helpful to outline what generics are. In communication, we often express generalisations, yet broadly speaking, there are two ways of doing so. Firstly, one can make a quantified utterance: something akin to ‘most people with schizophrenia suffer from delusions’. Secondly, one can utilise a *generic* such as ‘schizophrenics suffer from delusions’.

Generics are statements which express generalizations about any given kind, however they contain no explicit quantifiers. They are generally concerned with the expression of maxims, prototypes, stereotypes, laws or conventions. Generics are commonplace within the English language, and may take a variety of syntactic forms. They can be *bare plurals*, such as 'snakes are reptiles', *indefinite singulars* such as 'a shark's skin is composed of placoid scales' or *definite singulars* such as 'the mosquito carries the West Nile virus'. However, just because a bare plural, indefinite singular or definite singular appears within a sentence does not guarantee that the utterance need be interpreted as a generic. Leslie (2014 p.3) offers two examples for consideration:

(1) Tigers are striped

A tiger is striped

The tiger is striped.

(2) Tigers are on the front lawn

A tiger is on the front lawn

The tiger is on the front lawn.

The subject matter of set (1) and (2) seem to be identical: tigers. However, there seems to be some significant disparity between the ways in which we might interpret the statements in (1) and those in (2). Indeed, group (2) appears to refer to a set of *particular* tigers- namely, those on the lawn. As such, the correct interpretation of this set seems to be *existential/specific*, rather than generic.

Using Lawler's (1973) tests, one can check whether one's existential/specific interpretation is correct by checking whether the statement is upwards-entailing. For instance, in the statement 'tigers are on the front lawn', the subject term 'tigers' can be replaced by the more inclusive 'animals', yet the statement remains true: if it is true that tigers are on the lawn, then it must also be true to claim that there are animals on the lawn. Indeed, it is because the statement refers to particular tigers that this upwards entailment is possible. However, the same is not true when the statement in question admits of a generic interpretation. Consider the group of utterances in (1). These statements do not refer to particular animals, but rather to the kind or category *'tigers'*. Unlike (2), (1) makes claims about tigers *in general*. As such, one finds that the statements in (1) are not upwards-entailing: one cannot claim that because 'tigers are striped' that 'animals are striped'. Hence, the statements in (1) should be properly understood as generics, rather than existential/specific statements.

Generics do not make claims about individuals, nor do they function by describing how many members of a kind possess the predicate attributed to them. As Carlson (1997) righty notes, when asked how many ravens were black, one could not usefully reply with the generic 'ravens are black'. In this instance, a quantified answer such as 'all', 'most', or 'some' is required. However, generic statements lack such explicit quantifiers. As Geurts (1985 pp.248-249) observes, it is tempting to think that generic statements are quantified, albeit implicitly, and are thus reducible to quantified utterances. For instance, the phrase 'all sparrows are birds' appears to express exactly the same sentiment as the bare plural generic 'sparrows are birds'. However, it would be a mistake to think that generics can simply be re-read as quantified statements.

This is because generics admit of a rather striking characteristic, one which is described as their having 'fluctuating truth conditions' by Carlson and exhibiting "truth conditional laxity" by

Leslie (2014 p.1). The truth conditions of generic utterances are very difficult to establish. For the utterance 'sparrows are birds' to be true, it requires that *all* sparrows are birds. Indeed, to encounter a sparrow which was not a bird would render the generic false. Generics like this seem like they can be converted to quantified statements with the addition of a universal quantifier.

However, not all generic statements require such strict truth conditions. Consider the indefinite singular generic 'a tiger has stripes'. For this statement to hold, it is *not required* that all possible tigers have stripes. Indeed, the truth of the generic is compatible with the existence of a few non-conformists or anomalies: albino or melanistic tigers which do not possess stripes. There are examples of generics which have even laxer truth conditions. For instance, the definite singular generic 'the zebra shark has spots' remains true even though juvenile zebra sharks have stripes (which gradually morph into spots as they reach maturity), and that sightings of partially, or totally, albino variants of adult zebra sharks have been documented. Finally, the generic 'leopards have spots' is accepted into common parlance even though the melanistic variant, the black leopard, is very common.

Here we see why 'translating' generics to quantified statements is so challenging. In most of the cases outlined above, for the generic to be considered true it is not required that *all* members of the kind instantiate the trait. Thus, we cannot simplistically translate generics by adding in a universal quantifier. Yet, it is also true that some of the generics above do not require that even *most* members of a kind demonstrate the characteristic. A tempting move here would be to claim that we can still translate the generic, but we must say that 'some' members must have the property, rather than most. But even this is not an option. Once again, not all generics

require that even *some* of its subjects demonstrate the named predicate. 'Some' seems to imply relative frequency, yet generics such as 'mosquitos carry the West Nile virus' are acceptable to us, even though less than one percent of mosquitos carry the virus.

As I have already explored briefly in chapter 2, Leslie claims that generics where very few category members instantiate the property remain acceptable to us because the properties in question are often striking ones: properties which we would do well to avoid, and so take particular care to do so. However, it is worth noting that there are limits to this truth-conditional laxity. Generics are not entirely unconstrained by truth conditions. Indeed, there are generic statements which we deem false, or at least reject. For instance, Leslie remarks that we are perfectly happy to accept the generic 'mosquitos carry the West Nile virus', even given that less than one percent do, yet we would reject the statement 'books are paperbacks', although over eighty percent of them are.

Generics also raise other interesting issues regarding truth in that they tend to render our normal processes of verification problematic. This phenomenon is illustrated particularly well by Geurts (1985, p.249). Consider the generic statement 'giraffes have four legs'. It seems plausible that when attempting to conduct a truth-conditional analysis of this statement, we would allow that not all giraffes must have four legs in order for the generic to be acceptable: the truth of the statement is compatible with a few giraffes having three, two, or perhaps even no legs as a result of birth defects or accidents. However, imagine a scenario in which some ecological disaster occurs, which rapidly makes it the case that there are only 3 giraffes left in existence, two of which possess only three legs. Now what are we to make of the claim that 'giraffes have four legs', given that not even half of the category demonstrate the property?

One common response is to argue that the generic still stands, as the predicate need not describe the actual state of affairs in the category. Rather, 'giraffes have four legs' refers to a useful generalization about giraffes- that it is *normal* for them to have four legs, or that this property is something which giraffes *will* possess*, given acceptable circumstances* (no accidents, birth defects). The generic therefore articulates what we think is normal for the kind *'*Giraffe'.

At first glance, this seems like an adequate response to the kind of problems exemplified by the giraffe case. However, Geurts adds a further complication: what if the ecological disaster which led to the decline in giraffe population numbers and the loss of their limbs also had some effect on their genetic structure, such that the surviving giraffes produced offspring with only three legs when bred together? Perhaps here, we would be far less willing to continue to claim that the generic utterance was true, presumably because having four legs would no longer be part of the normal course of development for the given kind. On the other hand, perhaps we might also claim that in virtue of the significant change in the genetic structure, these animals are now no longer giraffes, but rather, animals of a different kind or species. In this case, the generic about giraffes still stands.

Indeed, these examples aptly illustrate yet another interesting feature of generic utterances: that, relatively speaking, they seem to almost *resist* verification. As Geurts observes, establishing the truth value of generic sentences is a complex matter. It is not merely a function of the actual state of affairs at any one time. One might plausibly think that the generic 'giraffes have four legs' remains acceptable in the case where two of the three remaining giraffes have three leg. However, we would be far less likely to continue to claim that the quantified utterance 'most giraffes have four legs' is true. Why is this? Generics seem to have a truth, or perhaps

more appropriately, a *resonance* above that of traditional notions or truth or falsity. Generics appear appropriate in cases like the one above, where quantified utterances do not, because they are not factual statements: they do not report how the world *is*, but rather how it *normally is*, or how it would be most useful for us to think of it as being. Generics are not concerned with mere representation of fact and demonstrate a sort of 'inertia' to variation in the truth conditions operating within the real world (Geurts 1985). Hence, whilst generics are certainly based upon reality, they do not literally denote it.

Following this brief explanation of generics, I will now outline why they are thought to be a problematic mode of phrasing and disseminating knowledge structures, before commenting on what should be done about it. Many of the features of generics outlined above will become relevant once again later in this chapter: indeed, as I will explore in section IV, because generics demonstrate truth-conditional laxity and both complicate and frustrate the establishment of truth conditions, certain kinds of responses to the problem generics pose for mental illness will not be feasible.

III. Why are Generics Problematic?

One reason to think that generics are problematic is because they encourage hearers to treat the kind referred to in the generic in an essentialized way: that is, to perceive members as sharing some deeply hidden fundamental nature or essence, which then warrants the projection of properties observed in one member of the kind onto other perceived members. A similar, yet

related, worry is that generics tend to be *about* kinds which are already highly essentialized. In this way, Leslie (2013, 2014) and Haslanger (2011) argue that generics foster essentialist beliefs about the kind, encouraging us to subject them to certain common, yet often implicit, heuristics for understanding. There are several things to be explored here. Firstly, what kind of essentialist beliefs are Leslie and Haslanger talking about? Secondly, how do generics encourage us to look at kinds in this way? Thirdly, how might this apply to mental illness, and why is it a problem? In the interests of clarity, I will consider each of these questions in turn.

What kinds of essentialist beliefs are fostered by generics? There are two parts to this question. First, are these beliefs philosophically essentialist, or essentialist in a folk or psychological sense? Second, what is the content of these beliefs: precisely what beliefs might we develop about a kind when we hear a generic about them? In order to answer the first question, it is worth outlining the difference between philosophical essentialism and psychological essentialism (which, echoing Leslie I will refer to as *quintessentialism* for clarity of exposition).

Philosophical essentialism is a kind of metaphysical essentialism, which might be broadly, although not uncontroversially, characterised as "the doctrine that (at least some) objects have (at least some) essential properties" (Robertson & Atkin 2016). Essential properties are those which contribute to the essence of any given thing, play a key role in determining its identity and make it the kind of thing that it is. Metaphysical essentialism is concerned with the *actual* ontological structure of reality, and it makes certain claims about how the world is really constituted: namely, that (at least some) objects have (at least some) essential properties. The 'essences' in which it recommends belief really are located in the world, and so essentialism makes a claim about how reality really is, in itself. In this sense, it is *metaphysical*: concerned with the fundamental nature of what exists.

In contrast, quintessentialism refers to the belief that individuals and kinds possess a fundamental essence or nature. It is a kind of *representational essentialism* in that it is not concerned with delineating the true nature of reality, nor does it make any of the ontological claims characteristic of its metaphysical variant. Rather, quintessentialism "addresses how people construe reality (in their belief systems, language, and cultural practices)" (Gelman 2004, p.405). As such, it is concerned with the way in which we *view* the world and those entities in it, rather than the way the world *in fact is*. Quintessentialism is a heuristic: a method of facilitating discovery or learning which is not guaranteed to bring about truth, nor the optimal epistemic outcome, but which nonetheless is sufficient for the purposes at hand. Leslie, amongst other scholars, has suggested that quintessentialism is a heuristic we commonly use to process the vast amounts of information with which we are presented on a daily basis: a cognitive short-cut which allows for progress in enquiry and reasoning, which formulates broad generalizations and tries to find similarities between things. Indeed, it appears to be a near universal form of cognition, for which there is a huge amount of empirical evidence. I will not summarize this here, given that the literature is so substantive. Instead, I will simply direct the interested reader to the sources specified in the reference[43].

We can now specify the claim above: generics encourage *quintessentialist* beliefs about the kinds they name, rather than metaphysically essentialist beliefs. However, before we identify what quintessentialist beliefs look like, it would be useful to make a few further clarifications about quintessentialism. For one, it is not a single or cohesive system, which is always applied

[43] See Leslie (2007, 2008, 2013, 2014), Gelman (2003, 2005), Haslanger (2011), Gelman & Wellman (1991), Taylor (1996), Taylor, Rhodes & Gelman (2009), Hirschfield & Gelman (1994), Hirschfeld (1996), Gelman & Markman, (1986) Jaswal & Markman (2002), amongst others. For an overview of all the communities in which quintessentialism has been observed, see Mahalingam (2003) Meyer et. al (2013), Gelman & Diesendruck (2001) Sousa et al. (2002), Waxman et al. (2007), Atran et al. (2001), Walker (1999) and Gil-White (2001).

in its totality to quintessentialized kinds. Rather, Leslie calls it a 'syndrome' comprised of many closely related, yet distinct, beliefs and attitudes. Leslie (2013, p. 115) stresses that:

> It is crucial to understand the hypothesis of psychological quintessentialism correctly. It does not involve the claim that to be a Quintessentialist is to subscribe to *all* the views listed above; rather quintessentialism is best thought of as a syndrome, which can be manifested in a variety of default implicit beliefs and ways of interpreting one's world.

It is also notable that some of the quintessentialist outlooks described in the material below contradict each other. Indeed, because it is not a cohesive outlook and the beliefs themselves are often irrational, contradictory quintessentialist beliefs can be held simultaneously. Finally, quintessentialism is a *folk* mode of thought, which describes the ways in which ordinary members of the public generally process information, or understand and structure the world around them.

Quintessentialism is a reasoning process which does not appear to be the product of education. Indeed, children as young as four have demonstrated quintessentialist beliefs (Gelman 2003, Leslie 2013). Leslie notes that these children lack the evidence and experience required to form such beliefs, yet they demonstrate them all the same. She takes this as evidence that quintessentialist modes of thought are *innate*, and that we are born with access to them. In other words, they are schemas which come 'pre-installed' in our conceptual machinery, and can be used prior to experience. Indeed, as Leslie (2014, p.210) puts it, quintessentialism represents "our most basic form of generalizing, one which we exploit from our earliest days".

Finally, quintessentialism is most commonly an *implicit* belief system. Often, we are not aware that we possess these beliefs, or that we think in this way (much as with some implicit biases

in chapter 2). Because such beliefs are implicit, and often introspectively inaccessible, we should not expect others to profess them aloud in most cases. Further, we cannot simply dismiss the existence of implicit quintessentialist beliefs by claiming that, following some thought, we do not think we have beliefs like that. Indeed, as Leslie (2013, p.115) notes, introspection can only reveal to us what our *explicit* beliefs are: it says nothing about the existence of beliefs of which we are not cognisant. Indeed, I hope that stressing that quintessentialist beliefs are often implicit will make them appear a little less strange to the reader.

i. Quintessentialism

What are the quintessentialist beliefs that Leslie and Haslanger suggest that generics expose kinds to? By means of introduction, I invite the reader to consider this list of (seemingly unrelated) phenomena, as outlined by Gelman (2005):

1. The president of Harvard recently suggested that the relative scarcity of women in "high-end" science and engineering professions is attributable in large part to male-female differences in intrinsic aptitude (Summers 2005).
2. In a nationally representative survey of Black and White Americans, most adults agreed with the statement, "Two people from the same race will always be more genetically similar to each other than two people from different races" (Jayaratne 2001).
3. Nearly half the U.S. population reject evolutionary theory, finding it implausible that one species can transform into another (Evans 2001).
4. A recent study of heart transplant recipients found that over one third believed that they might take on qualities or personality characteristics of the person who had donated the heart (Inspector, Kutz, & David, 2004). One woman reported that she sensed her donor's "male energy" and "purer essence" (Sylvia & Novak 1997, pp. 107, 108).

5. It is estimated that roughly half of all adopted people search for a birth parent at some point in their lives (Müller & Perry 2001).
6. People place higher value on authentic objects than exact copies (ranging from an original Picasso painting to Britney Spears's chewed-up gum; Frazier & Gelman 2005).

These examples seem diverse indeed. However, Gelman argues that each claim might be best understood or rationalised in quintessentialist terms. That is, what unites all these phenomena is the existence of some shared belief, the content of which is something like "certain categories (e.g., women, racial groups, dinosaurs, original Picasso artworks) have an underlying reality or true nature that one cannot observe directly" (Gelman 2005).

That is, quintessentialists believe that certain categories are in possession of an underlying, yet non-observable, *quintessence*. Leslie (2014, p.211) argues that quintessentialism entails the following:

> That is, we believe implicitly or explicitly, that each animate individual has an underlying nature or essence- an almost substance-like entity that pervades its insides, and which causally grounds its more stable and enduring properties. Further, we believe that certain kinds "carve essence at its joints"- that is, certain kinds (e.g. animal kinds) pick up on genuine differences and similarities in the essences of individuals.

Further to this, *The MIT Encyclopaedia of the Cognitive Sciences* describes quintessentialism as:

> …any folk theory of concepts positing that members of a category have a property or attribute (*essence*) that determines their identity (Gelman 1999 p.282).

Leslie's definition is quite dense, and so it would be useful to unpack it somewhat to better analyse some of the beliefs identified. On her model, animate individuals possess their own

quintessence. To be a quintessentialist (that is, to be one of the people that think in this way), one need not know precisely what the content of any given quintessence *is*. This is significant, as there are many instances in which the key quintessential features of individuals or categories are either difficult, or perhaps impossible, for the lay public to access. Indeed, we commonly rely upon what Prentice & Miller (2007) term 'essence placeholders': we simply assume that there is such a quintessence, even if we are ignorant of its true nature. Hence, quintessentialism requires only that we believe that each animate entity is in possession of *a* quintessence, even if we cannot say precisely what it is.

Quintessence is held to be the main determiner of identity: it dictates what the entity it is located within is going to be *like*. As Leslie (2013, p.112) puts it, quintessences have "causal powers", in that they make certain characteristics obtain in the individual. Quintessence is thought to be causally potent, and responsible for the instantiation of long-lasting, stable and non-accidental properties: for instance, gender, race, traits and personality. It is what causes the individual in which it obtains to demonstrate the properties or traits characteristic of it. However, the relationship between quintessence and trait is not one of necessity. As Leslie (2013, p.112) puts is, quintessences cause properties in a "defeasible way". Should circumstances not be favourable, then it may not be possible for the property to become manifest at all, or perhaps it may not be possible for it to continue to be manifested.

Quintessences are 'substance-like', and permeate the entity in which they are located entirety. Leslie (2013, p.114) expands upon this by noting that:

> …quintessences are substance-like entities in that they occupy- or better, pervade- space-time regions. They can also mix with each other, albeit only under unusual circumstances.

> Quintessences can permeate objects, including inanimate objects, and thus sometimes be transmitted. Notably, such transmissions almost always require a physical bearer.

Quintessence is normally understood to be located inside things, rather than on the surface. Indeed, quintessentialists believe that internal structure is more important than surface perceptual similarity when attempting to establish identity: a belief which has been demonstrated even in preschool children (see Keil 1989, Gelman & Markman 1986 and Jaswal & Markman 2002). Interestingly, the fact that quintessence is physically located (and thus 'substance-like') entails that it can be removed, as any part of our inner machinery can be. We can remove some inner part of an animate thing (for instance blood, tissue or an entire organ), and in doing so possess a portion of that individual's physical matter: imbued with, or containing, its quintessence.

As quintessence is held to be causally potent, quintessentialists believe that when the quintessence-bearing matter comes into contact with other animate individuals, it is capable of exercising this causal prowess upon them (Leslie 2013). For instance, Leslie (2013) and Gelman (2003) both observe that recipients of organ donations have reported believing that they have somehow taken on some of the properties of the donor. However, Leslie notes that quintessences retain some 'causal potency', even where the contact between recipient and the individual is not quite so substantive. Quintessentialists also believe that the transferral of quintessences is possible in another way: namely, that they can be passed on through physical contact. Leslie (2013, p.114) notes that whilst internal parts are held to the primary bearers of quintessence, quintessentialists often believe that one's quintessence can 'rub off' onto things one touches. For instance, items of clothing can become imbued with the quintessence of the wearer.

Finally, Leslie notes that quintessentialism can have a profound impact upon the perception of kinds, and how we go about categorising. In quintessentialism, there are multiple ways of conducting classification (for details, see Gelman 2003). As Leslie (2013, p.116) puts it:

> the Quintessentialists also believe that there are a number of levels or degrees of similarity in quintessence, all of which are real and objective, in effect constituting a taxonomic hierarchy of kinds. At the lowest levels of this taxonomy, there are considerable similarities between the quintessences of members of some distinct kinds, while at the higher levels, there is considerable variation between the quintessences of members of the same kind.

That is, we can form kinds on the basis of similarity between individual quintessences. However, there are a number of ways of characterising the similarity, and thus numerous ways of delineating categories or kinds: all of which are 'real and objective' insofar as they are determined by objective facts about quintessence. Thus, in creating categories on the basis of quintessential similarity, we construct a 'taxonomic hierarchy of kinds'. As Leslie notes, kinds can be formed where the membership all exhibit high degrees of quintessential similarity, whereas others will demonstrate significant variation in their quintessences. However, it is generally thought that there exists such thing as a "privileged level of this subjective taxonomy", which Leslie (2013, p.111) argues:

> occupies a "sweet spot" in this trade-off between within-kind variation in quintessence, and cross-kind quintessential distinctness. At this level, individual members of the same kind have only minimal differences in their quintessences, and these quintessences are quite dramatically different from the quintessences had by members of other kinds.

The project of discovering the objectively determined 'privileged' or 'basic-level' kind is concerned with the maximization of both intra-group quintessential similarity and inter-group quintessential dissimilarity. Indeed, it is because the entities in basic-level kinds have very similar quintessences to one another (and very different quintessences to members of other basic-level kinds), that quintessentialists often feel that kinds founded upon quintessence can truly 'carve nature at the joints'. Membership in a category is thought to be an intrinsic matter: established solely by the possession of the relevant quintessence. Just as it makes little sense to talk of an entity only partially demonstrating a particular quintessence, so too is it the case that membership cannot be uncertain. Rather, membership in the kind is an all or nothing matter, with kinds themselves being *discrete* in nature and admitting of *sharp boundaries* on membership.

Quintessence thus determines both category membership and serves as the causal basis for all those properties of an entity which are stable or enduring. As such, it is unsurprising that quintessence plays a vital role in quintessentialists' inductive practices. Basic-level kinds, for instance, are those in which the members of the kind admit of very similar quintessences, and thus exhibit the same kinds of properties. In this sense, kind-membership is hugely informative: if we know that one member of a basic-level kind demonstrates a property which is caused by quintessence (as opposed to being accidental), then we can infer that another member of the same basic-level kind will also demonstrate the property in question. Hence, basic-level kinds allow us to make inductive inferences about members of a category on the basis of their category membership. Given that basic-level kinds attempt to maximize both inter-group homogeneity and inter-group heterogeneity (with both characterised in terms of quintessence), these inferences, whilst certainly not infallible, are generally reliable.

To summarize, quintessentialism is the phenomenon by which we perceive kinds as having a fundamental underlying essence or nature that is shared by its members. As Leslie & Lerner (2016) put it:

> We believe that, while some kinds may group individuals together on the basis of superficial properties (as an extreme example, consider the kind *trinkets*), other kinds group their members together on the basis of deep, intrinsic similarities (animal kinds are paradigmatic examples here). Kinds that fall in the latter category are said to be *essentialized* in the psychological sense.

Kinds that are quintessentialized (that is, essentialized in the psychological sense) are liable to having quintessentialist heuristics applied to them. That is, its members will be thought to share a similar quintessence, which is not observable, yet somehow constitutes and dictates their fundamental nature. Their quintessence determines what kind of thing it will be, and which features it will possess. This quintessence is thought to be substance-like, causally potent, and capable of being transferred onto objects or even to other individuals. Quintessentialists think that at least some kinds 'carve nature at the joints', and so believe that the category permits a particularly rich kind of induction and has discrete boundaries (where membership is an all-or-nothing, intrinsic matter).

ii. Quintessentialism and Generics

The link between quintessentialism and generics is that generics tend to be *about* kinds which are already highly essentialized, or they *encourage* quintessentialist beliefs about a kind. Leslie

argues that if we convey a knowledge structure or stereotype using a generic (of any syntactic form) then we encourage the quintessentialization of that kind. Generics which are already about quintessentialized kinds encourage this tendency even further. But how do generics do this? As this is not the main focus of my work, I will offer only a brief summary of the mechanism and the evidence here. Simplistically, human beings utilise linguistic clues (amongst other forms of experience) to establish whether a kind should be thought of in quintessential terms: generics such as '*Xs* are Φ' (and generics of other syntactic forms) suggest that the kind is an essentialized one (Anthony 2016, p.187).

Evidence shows that from our infancy, information disseminated to us via the mechanism of generic language tends to be interpreted as meaning that the property attributed to the kind is a stable and enduring one, whether the kind in question is social or natural (Cimpian & Erickson 2012 and Cimpian & Markman 2009, 2011). Research by Gelman, Ware & Kleinberg (2010) and Rhodes, Leslie & Tworek (2012) demonstrates that both children and adults who hear generic language used to describe a novel animal or social kind come to quintessentialize that kind. Rhodes et al. (2012) found that pre-schoolers who heard generic language used to describe properties belonging to a novel social group (dubbed the 'Zarpies') quickly came to quintessentialize that group, even though the novel group contained a diversity of genders, races, ethnicities and ages. Rhodes (2012, no pagination) describes a simplified version of the methodology used in this experiment:

> Children were shown Zarpies one at a time. As they saw each new Zarpie, some of the children heard generic sentences (*e.g.*, "Look at this Zarpie! Zarpies are scared of ladybugs"), while others heard non-generic sentences (*e.g.*, "Look at this Zarpie! This Zarpie is scared of ladybugs"). A few days later, we examined what children in the two groups thought about Zarpies. Hearing generic sentences had striking effects on children's beliefs. First, several days

later, children who had previously heard generics more often assumed that Zarpies would share new traits. For example, we told them a new trait of a particular new Zarpie (*e.g.*, "This Zarpie makes a buzzing sound"). Children who had previously heard generics were more likely to assume that other Zarpies, even those that differed in sex, race, and age, would share the new traits (*e.g.*, would also make buzzing sounds).

I will not say more here. However, what can be demonstrated is that there is a robust link between the use of generic language and the quintessentialization of kinds. Now that the link between generics and quintessentialism has been broadly outlined, we are now able to see why stereotypes expressed as generics may be problematic for attitudes and beliefs towards mental illness and the mentally ill.

IV. Generics and Mental Illness

I have already noted that generics are one way in which we might convey a knowledge structure. Knowledge structures for mental illness can generally be inaccurate, or they might be accurate. Both accurate and inaccurate stereotypes can be conveyed in generic forms. For instance:

(1) Schizophrenics are dangerous

(2) Schizophrenics experience delusions

(1) conveys a knowledge structure which is inaccurate, whereas (2) conveys a knowledge structure which is generally accurate (many, although not all, people with schizophrenia experience delusions). When conveyed in this generic form, neither (1) nor (2) strike us as immediately false in the same way that 'books are paperback' might. Indeed, at least to my

mind, both appear to be the kind of utterance which might plausibly be made in a conversation, or which someone might read somewhere. Yet, I suggest that there is something objectionable about both, albeit for different reasons.

Let us commence with (1). The truth of this generic does not seem to require that all schizophrenics are dangerous, nor indeed, most. The generic seems acceptable, or at least, not obviously false. Crucially, whilst the acceptability of the generic may be attributed to ignorance (perhaps the hearer doesn't know that people with mental illnesses are actually more likely to be the victims of violent crime than the perpetrators), there is another explanation for its continuing resonance. There are cases in which the use of the generic form makes an utterance seem acceptable to a hearer, even where that hearer is aware that the knowledge structure underlying it (here, the claim that it is common, or even relatively common, for people with schizophrenia to be dangerous) is not accurate. It is the use of the generic form which makes the utterance acceptable to us, even though if the same knowledge structure were expressed in a quantified form, it would likely be rejected.

How can we explain this? I have already described one answer from Leslie (2013, 2014): namely, that the generic is accepted at low prevalence levels because the property predicated of the kind (here, 'dangerousness') is a striking one- one which it would be in our interests to avoid. This, Leslie argues, allows us to explain a particularly insidious form of social stereotyping. As Brownstein (2016, no pagination) puts it:

> even if just a few members of what is perceived to be an essential kind (e.g., pit bulls, Muslims) exhibit a harmful or dangerous property, then a generic that attributes the property to the kind likely will be judged to be true.

Thus, merely one purported or actual incidence of someone with schizophrenia being dangerous will be sufficient for (1) to be judged true, or at least, for it not to be rejected. In this way, the generic can convey the prejudicial stereotype described in (1) even though its truth conditions are very lax. Thus, generics can convey problematic knowledge structures about mental illness which are not factually accurate, thus contributing to stigma by disseminating and propagating stereotyping. They are problematic in that they encourage hearers to generalise properties of perceived quintessential kinds to other members of that kind, unfairly depicting all people with schizophrenia as being dangerous, even though it is very unlikely to be true.

Yet, beyond even this, generics give particular cause for concern in that they are commonly accepted where a quantified utterance expressing the same knowledge structure would not be. Why is this? For instance, why is it that if I were to say something like 'all schizophrenics are dangerous' or 'most schizophrenics are dangerous', these utterances would be more likely to be rejected than 'schizophrenics are dangerous'? The quantified statements would likely be rejected because they do not reflect 'the facts', or the actual empirical state of the world. Yet, as I noted in the material above, generics have often been described as having a 'resonance' which transcends traditional notions of truth or falsity. Establishing the truth conditions for generic utterances is a complicated matter, and one that is not simply established as a function of the actual state of affairs at any one time. They demonstrate an 'inertia' towards variation in truth conditions in the real world (as aptly demonstrated by Geurts' giraffe example).

Generics appear acceptable to us where a similar quantified utterance would not, and the reason for this is that generics are *not factual statements*: they do not report how the world *is*, but rather how it *normally is*, or how it would be most useful for us to think of it as being. Whilst

generics are certainly based upon reality, they do not literally denote it. Whilst the quantified utterance can be rejected because it is an inaccurate depiction of the real world, the generic will likely not be, as its resonance does not depend upon truth in as strict a manner as the quantified utterance. In this way, the generic appears to bypass our normal processes of acceptance and verification, and smuggles the problematic stereotype past us, where a quantified utterance would likely not succeed (Leslie 2013). Generics convey problematic knowledge structures very effectively, as they are particularly well-placed to bypass our normal methods of detecting falsity. Later in the paper I will outline how this feature impacts upon the methods available to us to challenge generics about mental illness.

Further to this, we should be worried that generics, despite being so difficult to submit to truth conditional analysis, seem to be particularly easy to extract quintessentialist information from. Recall Rhodes et. al's (2012) work on the Zarpies. In this case, very young children were able to understand the generic and glean quintessentialist information from it after very few exposures. Empirical evidence suggests that children are able to make these cognitive leaps, even though it would be virtually impossible to describe to a child what a 'quintessential' kind was, or to ask them why they were making the predictions and inferences they were on the basis of kind membership. Thus, generics can convey a great deal of information, yet are simple enough that they can be understood by children, and information can be extracted from them after very few exposures.

This is certainly cause for concern. Indeed, this suggests that young children who hear (1) even just a few times will be able to understand that 'schizophrenics' are a quintessential kind, and that the characteristics of one member will likely be shared by other members of that kind. In

this way, prejudicial stereotypes about mental illness can be conveyed to children in a manner which is both simple and rapid. Indeed, similar results have been replicated in adults. Thus, generics propagate problematic stereotypes or knowledge structures in a way which is particularly pernicious, underhand and both difficult to monitor and combat.

We can now summarize why generics like (1) are objectionable. Simply put, the use of the generic form smuggles a problematic knowledge structure past our normal methods of evaluating truth and falsity. We find them acceptable even where we do not find a quantified utterance conveying much the same knowledge structure so. As Leslie argues, even in cases where very few members of the kind demonstrate the property, if the property is striking then it can be easily generalised to other members of the kind. Thus, generics like (1) suggest that the category 'schizophrenics' is a quintessential kind, and that all kind members will have much the same characteristics. Striking property generics in particular contribute to insidious and wide-reaching forms of social stereotyping. Compounding this yet further is the concern that generics are easily transmissible, and can be understood by even very young children: they convey problematic knowledge structures very effectively, and to a wide audience.

Generics like (1) certainly raise problems for the stigma of mental illness, and so must be addressed. I will say more about how this can be done in a later section. Yet, the way in which these kinds of generics are problematic for mental illness isn't particularly *distinctive* to mental illness. Indeed, the account above very closely matches that offered by Leslie and other scholars regarding race and gender. Generics like (1) are problematic in much the same way that other kinds of generics with prejudicial content are: for instance, 'blacks are criminals' or

'women are too emotional'. Generics like (1) contribute to social stereotyping (and thus, stigma) in a more general way.

Yet, there is a way in which generics about mental illness (and some medical diseases) cause distinctive issues. (2) conveys a knowledge structure which is generally accurate: many people with schizophrenia experience delusions. Indeed, experiencing delusions is one of the five diagnostic criteria for schizophrenia in DSM-V (APA 2013). For the diagnosis to be made, 2 of the 5 symptoms must be demonstrated for a certain time. As explored in chapter 4, it is therefore not necessary that one experience delusions to have schizophrenia (as no symptom is individually necessary for the application of the disease concept), but it is true that many people with schizophrenia do experience delusions. Insofar as generics like (2) express a knowledge structure which corresponds to clinical reality, we might think that they could be useful for us, or at least, not problematic in the same way that generics like (1) are.

But, there is reason to think that generics like (2) will also be problematic, and will contribute to stigma or misinformation. Why is this? It cannot be that the knowledge structure conveyed by the generic is the source of the difficulty. Indeed, the knowledge structure (that it is relatively common for people with schizophrenia to experience delusions) is accurate. If we were to express it in a quantified manner (e.g. '*many* people with schizophrenia experience delusions', or even '*some* people experience delusions') this would not be problematic. Yet, the utterance 'schizophrenics experience delusions' *is* problematic. Indeed, it is the use of the generic form to make utterances about mental illnesses which is so deeply unhelpful, rather than the accuracy of the knowledge structure being conveyed.

Why is the use of the generic so problematic? The answer lies in how it encourages us to perceive the category it names. As noted above, generics are often used to describe kinds which are already highly quintessentialized, and when the generic form is used, the kind it names is set up as a candidate for a quintessential kind. (2) encourages the hearer to think of the category 'schizophrenics' as a quintessential kind, much as (1) does, as both are generics. In doing so, it sets the category up as being generally homogenous, and thus very solid ground for making inferences and predictions.

We may well think that this is not particularly troubling. Indeed, the category 'schizophrenics' does pick out *relatively* similar individuals, and in chapter 4 I have already stated that diagnostic categories are commonly used in explanation, induction and prediction to great effect. Indeed, we might even think that because generics convey complicated information in a simple and accessible way, they may actually be beneficial for the purposes of learning. Why is this? A generic like 'tigers have stripes' (which expresses an accurate knowledge structure) may run into problems when there are extreme changes in the real world (just as Geurts' example about giraffes), yet we might think that this is worth the pay off. The generic allows us to make useful generalisations about tigers. Upon seeing a creature with stripes, we know that there is a possibility it will be a tiger. Indeed, without ever having seen a tiger, if we were called upon to describe one then we could: if something is a tiger, then it will likely be striped and have 4 legs. Admittedly, this has its limits. For instance, a tiger may not have stripes (it may be albino or melanistic). However, we might think that although fallible, these inferences are valuable. Hence, there is an argument that generics convey knowledge structures efficiently, quickly and simply. This could be hugely beneficial when attempting to educate the public about mental illness.

Indeed, we might claim that a similar argument can be made for (2). The kind 'schizophrenics' is based on certain similarities like experiencing delusions or having hallucinations (and so is generally unified), and we already treat the categories as if they are informative (see chapter 4). It remains the case that if a drastic change occurred such that the generic no longer accurately reflected the real state of affairs in the world, we would encounter the same problem of inertia as in the giraffe example. However, perhaps as with the tiger case, the benefits to understanding afforded by the generic would be worth this pay off. Indeed, if generics are so good at conveying a wealth of information to audiences who would struggle to understand it if explained in other terms, then perhaps generics could be harnessed to quickly yet effectively improve mental health literacy?

The above seems somewhat plausible, but I maintain that the use of generics to describe mental illness remains problematic. It must be appreciated that utilising the generic does not merely encourage the perception of a kind as unified and founded upon some loose similarity. Rather, it encourages that it be viewed as a *quintessential* kind. Where generics are used to describe mental illness, it encourages us to view the kind 'mental illness' and the individuals within it in quintessential terms, and to apply the quintessentialist beliefs outlined in section III.i to them. These beliefs include the perception of the individual as possessing a fundamental nature or essence which is causally potent, substance-like and potentially transferrable. Further, that some kinds do carve nature at the joints, where membership in the kind is an intrinsic matter, and there are sharp boundaries on membership.

These beliefs, when applied to mental illness, are deeply unhelpful. Thinking about mental illness and the mentally ill in quintessential terms will likely generate a great deal of

misinformation, some of which will be profoundly damaging to people with mental illnesses and public perception of them. Thus, insofar as generics encourage the development of quintessentialist beliefs about mental illness and the mentally ill, they should be avoided. Generics like (2) are deeply problematic- not because the knowledge structure they convey is inaccurate, but rather, because the use of the generic form sets the category 'mental illness' up as a candidate for a quintessential kind, which then generates misinformation.

In what follows I will say a little more about the potentially damaging consequences of subjecting mental illness categories and mentally ill people to quintessentialist outlooks, and the kinds of problematic interpretations which might be encouraged by generics. The first misconception that may be generated is that all mental illnesses are stable, serious and enduring. Beliefs such as these may affect public and private attitudes towards illness and recovery. The second is that using generics may bring us to think that mental illness categories are profoundly unified and rich sources for induction and prediction: an interpretation which does not adhere with the accepted clinical understanding of mental illness. The third concern is that quintessentializing mental illness categories may also lead to entitative thinking, which itself is liable to exacerbate the social distance between the healthy and the sick. Finally, quintessentialist interpretations of mental illness may lead to 'contagion' thinking, which may contribute to stigma and frustrate some existing anti-stigma initiatives for mental illness.

i. Mental Illness as Stable, Serious and Enduring

One prominent kind of erroneous or problematic belief that might be generated from perceiving schizophrenia, for instance, as a quintessential kind is that schizophrenia may be seen as

inherent or incurable. In quintessentialist thinking, membership in a kind is thought to be an intrinsic matter, determined by quintessence. Quintessential nature is generally thought to be unchanging, and the kinds themselves are discrete, with sharp boundaries on membership. If schizophrenia were to be seen in this way, then having the illness may plausibly be seen as being part of one's nature. It may come to be thought that an individual's quintessence causes her to have schizophrenia and the associated symptoms (and it is in virtue of her quintessence that she belongs to the illness category), and that having schizophrenia is simply part of her fundamental or underlying nature.

Notably, this is similar to the claim that schizophrenia has a biological *cause*. Indeed, there are many similarities between the gene concept and quintessence. Leslie (2013) observes that in modern thought, genes are often considered to be part of one's quintessence. Genetic structure cannot change, much as quintessence cannot. Further, both are understood to cause certain properties associated with them, often in a defeasible way. As Griffin (2016) reports, scientists have long attempted to establish a biological cause of schizophrenia, and there is some evidence to suggest that progress is being made. Should a cause for schizophrenia be identified, this would surely have great medical benefit.

However, there is a sense in which the language of genes or biology, and of quintessence, can be unhelpful for people with mental illnesses. Corrigan (2016) and Phelan (2006) have observed that programmes which encourage us to describe mental illness in biological or genetic terms have proved successful in reducing blame. Where mental illness is explained in this way, individuals appear to blame people with mental illnesses less for their behaviour or transgressions. Yet, with this another phenomenon was observed: that using genetic and

biological language to describe mental illnesses also increased the degree to which individuals believed the mental illness in question to be serious and permanent. I suggest that using the language of quintessence will also do this. Quintessentialists believe that the stable, significant and enduring properties of individuals are caused by quintessence, and further, that our quintessence is unchanging, and characterises our fundamental nature or identity. Insofar as the symptoms or characteristics of mental illness may be seen as significant properties, then they are liable to being understood as quintessential features of the individual: simply part of her nature, and unchanging.

If it is true that the use of generics may encourage quintessentialist beliefs about the severity and stability of mental illnesses, this may prove problematic. The most obvious issue this may cause concerns attitudes towards recovery. It has long been a problem for people with mental illnesses that once someone is found to be mentally ill at one time, this then becomes hard to escape. Others, or even the individual herself, may come to think that she will always have that illness. This may be frustrating for people who once attracted diagnoses. Further, it may lead to individuals who are now in recovery being treated as if they still had the illness. In this way, even those who no longer fit into an illness category may be subjected to the perceptions outlined in chapter 1. If illness is seen as a permeant or quintessential property, this will likely cause many parties (both those who have a mental illness and those who do not) to be sceptical of recovery, or perhaps even to doubt that recovery is possible. This may drastically impact upon the self-esteem, self-efficacy and hope for the future of many people with mental illnesses.

Furthermore, if members of the public had beliefs like these about illness categories, then there is a danger that they may 'write off' individuals they believe to have a mental illness, and will not revise their opinion or treatment of them. People who are now in recovery may find that this is not appreciated, and that they continue to be patronised, or are treated as if they are incompetent or lack capacity. Individuals who once had episodes of psychosis, for example, may find that people are unwilling to approach them afterwards: the reason being that their being ill is perceived as a stable, enduring and permeant characteristic. This kind of attitude may also lead to a failure to revise judgements of epistemic or testimonial credibility as described in chapter 3, thus propagating epistemic injustice.

If illness is seen as a quintessential and enduring property, then little effort may be made to adjust to people with mental illnesses. For instance, imagine that an employee has a mental illness. In this case, an employer may take note of the employee's condition (and those factors which exacerbate it), and where possible, make adjustments so that the employee can work effectively whilst avoiding situations which may be distressing. For instance, one may afford an employee with anxiety more time to complete projects, or one might not assign them to deal with a particularly aggressive customer. However, if one were to think that mental illness is always serious and enduring, one may wonder why one should bother making adjustments of this kind- after all, the employee will never get any better. Attitudes like this may prove very harmful to people with mental illnesses, and may make it impossible for them to continue working in certain climates. Thus, one of the problems involved with using generics about mental illness is that it encourages quintessentialism, which itself discourages belief in meaningful recovery (construed as a return to functional life) and change.

Yet, there is another sense in which using a generic to quintessentialize a kind can be problematic. Haslanger (2011) notes that some generics, such as 'women are submissive' seem objectionable to us, yet may be 'true' in the sense that a significant number of category members instantiate the associated property (for instance, many women may actually depend upon men for their livelihood). She argues that 'true' generics of this kind remain problematic because presenting the kind as a quintessential one serves to contribute to various forms of social oppression. They do this by making problematic quintessentialist claims about the nature of the subject-predicate relationship.

For instance, consider the claim 'roughly half of people with schizophrenia experience delusions'. In this quantified example, the subject-predicate relationship is presented as matter of close statistical association. It suggests that it is *empirically* the case that roughly half of people with that illness have that symptom. According to Haslanger, claims such as these report the way the world actually is. In contrast, generics express a far more problematic kind of predicate-subject relationship: quintessentialism[44]. Haslanger argues that to utter a generic is to make a claim about the grounding of a particular subject-predicate relationship in nature: namely that it is part of the quintessence of F, for instance, to be G. This is a much stronger claim that the quantified variant.

Haslanger (2011, p.2) suggests that generics smuggle quintessentialist ideologies ("representations of social life which in some way serve to undergird social practices") into the linguistic common ground in which we all share. Where generics are used, a falsehood about

[44] Haslanger uses the term 'essentialism', but she is referring to the psychological variant. Here, I have paraphrased her as 'quintessentialism' for ease of exposition.

the nature of the subject-predicate relationship is pragmatically implicated. For instance, the generic 'women are submissive' suggests that it is not merely circumstantial or contingent that many women rely on men for their livelihood. Rather, it suggests that it is part of the quintessence of women that they are submissive. This is why generics which appear to be 'true' (in that a significant number of category members seem to demonstrate the property) can remain problematic- they perpetuate social oppression by making quintessentialist claims about the relationship between category members and characteristics.

Haslanger's work has interesting implications for mental illness. To say 'schizophrenics experience delusions' is to claim that it is part of the quintessence of being schizophrenic that one experiences delusions. This is a much stronger claim than the quantified variant 'many people with schizophrenia experience delusions', yet both appear to be expressing much the same knowledge structure. Yet, in the former the subject-predicate relationship is a matter of quintessence, rather than statistical regularity as in the latter. Further, consider the utterance 'worriers make for poor employees', where 'worriers' is used to refer to people with anxiety. To many, this statement seems problematic. However, what is it that grounds our desire to reject the statement? Strangely, the generic may actually be *true,* given a set of standard metrics detailing what it is to be a good employee. For instance, it may turn out, empirically speaking, that those employees with anxiety tend to be absent from work for longer periods than their non-anxious counterparts, or perhaps that they tend to exhibit lower productivity, or simply do not contribute as well to the maintenance of a happy working climate.

Yet, using Haslanger's model, we can express why this generic is problematic. It pragmatically implicates the falsehood that it is just something about the quintessence of people with anxiety

that they are poor employees. Yet, this concerns us because this is simply not true. Indeed, many people with anxiety excel in the workplace (particularly where accommodating measures such as flexible working hours were implemented). We want to allow that in many cases, it is a contingent matter that people with anxiety will perform poorly at work, andnot a quintessential characteristic of theirs. In this way, by misconstruing the nature of the subject-predicate relationship, the generic 'worriers make for poor employees' contributes to forms of social oppression by depicting the relationship between category and property as one of quintessence rather than contingency.

ii. Quintessential Similarity, Induction and Prediction

In chapter 4, I acknowledged that mental illness categories are often fruitfully used to carry out explanation, induction and prediction. However, viewing mental illness categories in quintessential terms can prove damaging. As noted above, quintessentialists tend to think of categories as *richly* informative. Kinds are formed around similarities in quintessence, where members share deep underlying similarities in virtue of their both being members of the same category. For instance, a quintessential understanding of the category 'Panthera Tigris' states that category membership is a matter of having the right quintessence, and that members of that category share this quintessence. From here, we can infer that tigers will be very similar to one another, and knowing that something is a member of the category 'Panthera Tigris' allows us to make certain predictions about it (that it will very likely have four legs, be striped, be a carnivore etc.). Here, category members are thought to be deeply similar to one another, and membership in the category can be utilised as a basis for a wide array of inductive and predictive inferences.

Yet, this kind of thinking does not reflect the clinical reality of many mental illnesses. As explored in chapter 4, it is not the case that all members of the category 'Panic Disorder' will be deeply alike. In DSM-V (APA 2013, p.246), the diagnosis is afforded if a minimum of 4 of 13 different criteria (symptoms) apply. Two individuals (*a* and *b*) could be afforded a diagnosis of Panic Disorder, yet both might demonstrate entirely different symptoms. Just because *a* reports that she experiences the sensation of choking (one of the diagnostic criteria), we cannot assume that *b* also does in virtue of the fact that she belongs to the same diagnostic category as *a*.

Indeed, members of mental illness categories are similar to one another in that they broadly share in some of the same characteristics or symptoms, yet the kind is not as unified as the quintessentialist interpretation would suggest. As Garand et.al (2009, p.2) rightly observe, establishing categories allows us to assume that the category is:

> generally homogeneous in the underlying nature of the illness, regardless of whether there is some variability in the presentation of symptoms or circumstances surrounding illness onset.

Yet, although categorization permits us to assume a degree of homogeneity, it does not allow us to think that members of the category are likely to be deeply similar: and this, problematically, is what quintessentialist thinking encourages. Insofar as generics encourage quintessentialist interpretations of mental illness categories, they spread misinformation by presenting the kind as more unified than accepted clinical understanding says it is. As a result, it misrepresents how robust the explanations, predictions and inferences we might make using the category are.

Membership in a mental illness category *is* informative (and thus grounds for some further explanatory, inductive or predictive manoeuvres), yet it is not *as richly* informative as quintessentialist thinking suggests. Indeed, mental illness categories are far more heterogeneous than categories such as 'Panthera Tigris'. Some are not particularly unified at all. For instance, some mental illness categories seem to be semi-arbitrarily drawn (i.e. diagnoses which fall into the 'not otherwise specified' category). We cannot predict which symptoms someone with Panic Disorder will demonstrate just from knowing their diagnosis, and we also cannot assume that someone with Panic Disorder will behave exactly like someone else who we know to have the same diagnosis. Thus, generics about mental illness are liable to spread misinformation by depicting mental illness categories are more unified than they in fact are, and misrepresenting the robustness of any explanations, predictions and inferences we may form on the basis of category membership.

iii. Entiativity

Work in cognitive science has suggested that quintessentialism about categories also causes individuals to engage in *entiative* patterns of thought. The precise relationship between the two phenomena is the subject of much debate. However, a convincing case has been put forward by Yzerbyt et. al (2004), in which the authors claim that quintessentialism and entiative thought mutually sustain one another. Whereas quintessentialism is broadly defined as the belief in underlying quintessences, entiativity is understood as the general tendency to search for, or perceive, similarities and connections between members of a category (Yzerbyt et. al 2004). As Haslam at. al (2000, p.116) put it:

> entiativity represents the extent to which a social aggregate is perceived to be a coherent, unified, and meaningful entity, and is a function of Gestalt principles of similarity, proximity and common fate.

Entiativity thus represents the extent to which the members of a category are similar, or will behave in comparable ways. Entiative tendencies, such as the disposition to seek out similarities between group members, reinforce quintessentialism in that the existence of an underlying quintessence is often posited to explicate said similarities. Conversely, the tendency to quintessentialize categories and uncover deep underlying characteristics is likely to encourage the search for similarities. However, it is often the case that the similarities we settle upon are trivial, or merely surface level. Even so, entiative thinking has been shown to be incredibly significant for our perception of social categories.

For instance, Yzerbyt et.al (2004, p.141) have conducted experiments in which participants were randomly divided into two groups. The first group were informed that they had been put into that category because all members of the group had something in common, although they were not informed precisely what it was (i.e. they were assigned a trivial category label). By implying that there was some shared similarity that served as the rationale for drawing the category in that way, the investigators implicitly encouraged the first group to think in an entiative manner. The second group were informed that the grouping was random.

Encouraging the first group to think in this way had several interesting effects. Firstly, their perception of intragroup entiativity led them to believe that there were sharp boundaries on group membership, which in turn led to increased group identification (Yzerbyt et.al 2004,

p.145). Yzerbyt et. al note that these findings seem to suggest that entiative thinking gives rise to quintessentialist conceptions of group membership. Conversely, the members of the second group (in which there was perceived to be a very low degree of group entiativity) displayed little tendency to identify with the grouping, as they saw no underlying meaning in its delineation.

Several interesting results were observed. Of greatest import to my work is that it appeared to be the case that quintessentializing a category encouraged participants to form entiative beliefs about said category. As Yzerbyt et.al note, the imposition of the trivial category label provided a minimal set of quintessentialist beliefs: it conveyed the notion that the category must be in possession of some underlying properties which justified its being named as a distinct grouping. As such, the first group exhibited increased entiative thought in that they exaggerated both the intra-group similarities and the inter-group differences: engaging in entiative patterns of thought led the participants to polarize the groups, perceiving them as particularly distinct. Hence, those groups with high entiativity were thought to admit of a greater degree of homogeneity, whereas lower entiativity groups were thought to be more internally diverse.

If Yzerbyt et. al are right that quintessentialism and entiativity mutually support one another, this may have consequences for mental illness. Generic utterances may engender not only quintessentialist thinking, but also entiative thinking. How might this impact upon mental illness? Diagnostic labelling is way of forming categories and drawing boundaries between groups: crudely, between people who are ill and those who are not. Crucially, the rationale behind this category is not arbitrary: it is not a *trivial* category label, but rather, one that is thought to describe some real differences between people. If even trivial category labels

encourage entiative thought, empirically founded category labels are likely to exacerbate this tendency yet further. Hence, as mental illness labels are thought to be meaningful, it is likely that there will be some danger that entiative thinking will lead to the two categories being polarized.

This polarization is unlikely to help the fight against the stigma of mental illness. Indeed, it is likely to frustrate it to a profound degree. As noted by Link & Phelan (2001), the perception that 'they' are very different or even totally distinct from 'us' can lead to the dehumanisation of out-groups, which can lead to appalling treatment. In any case, perceiving people with mental illnesses to be very dissimilar to us is likely to encourage distrust, avoidance and perhaps even violence. In many ways, entiative thinking might widen the social distance and lack of contact between those with mental illnesses and those without.

This will likely frustrate contact strategies for combatting the stigma of mental illness. Contact strategies work by encouraging the public to see people with mental illnesses as individuals, and to confront the public with counter-stereotypical instances of people with mental illnesses, thereby hopefully discrediting the stereotype. Yet, if the mentally ill and the healthy are polarized and seen as radically different, then this may not be possible. For instance, individuals are unlikely to attend to the humanity, emotions and particular circumstances of those they perceive to be very different to themselves. In this way, one of the dangers of quintessentialism about mental illness (and thus of encouraging quintessentialism by using generics) is that research suggests that it in turn encourages entiativity, which itself leads to the polarization of groups. This may actively encourage stigma, and will likely frustrate contact anti-stigma initiatives.

iv. Contagion

Leslie (2013, 2014) observes that one prominent quintessentialist belief is that individuals possess a quintessence which is substance-like and causally potent. Quintessence is thought to be contained in the inner parts of an entity, and can be removed, rubbed off onto other things or transplanted. Where this occurs, the quintessence can cause the thing into which it is introduced to take on some of the properties it causes. Thus, Frank's quintessence pervades his inner machinery (in this case, his heart). When his heart is transplanted into Dee, she may come to believe that his quintessence is now inside her. As Frank's quintessence is causally potent, she believes that she comes to take on some of his traits. Perhaps she reports that she feels his 'male energy', as in Gelman's (2005) example.

This quintessentialist belief seems incredibly similar to a phenomenon Frazer (1976) called *contagion*: a quasi-magical belief in which the subject believes either implicitly or explicitly that the fundamental nature of an animate entity can be 'rubbed off' or transferred to another animate entity or object. Indeed, contagion-thinking appears to be a sub-set of broader quintessentialist thought, and so one kind of belief which quintessentialized categories and individuals may be subjected to. Insofar as generics about mental illness encourage us to perceive the category as a quintessential one, they also encourage us to think about mental illness and the mentally ill in terms of contagion: namely, that the quintessence which causally grounds significant properties like illness could be transferred. This is likely to be a profoundly unhelpful way of thinking about mental illness, and may cause avoidance and revulsion, whilst frustrating contact strategies.

Yet, to explore this further it is necessary to outline precisely what contagion is. There is a well-documented, yet puzzling, set of phenomena, all of which appear to be united by a common theme: namely, the belief that the fundamental nature of an entity or object can be in some sense transferred from it onto something else- be it another animate individual or a different object. This is known as the *contagion principle*: one of two laws of 'sympathetic magic' put forward most clearly by the anthropologists Mauss and Frazer in their respective works '*A General Theory of Magic*' and '*The Golden Bough*'[45]. Mauss and Frazer believed that these laws represented universal patterns of thought, which could be found in a plethora of cultures and served to explicate a large variety of magical, ritualistic and spiritual beliefs or practices. The law of contagion states that (Rozin & Nemeroff 1994, p.159):

> people, objects, and so forth, that come into contact with each other may influence each other through the transfer of some or all of their properties. The influence continues after the physical contact has ended and may be permanent (Frazer 1976). According to Mauss, "once in contact, always in contact".

This transfer of properties is thought to be caused where an entity or object, *x*, comes into contact with a contagious entity, *c*, often characterised as the 'vital essence' or 'soul stuff' of another entity, *y*. The quintessence of *y* is a source of contagion: where *y's* quintessence comes into contact with *x* (where *x* can be another individual, an animal or an object), it can potentially infect and influence *x* through the transfer of some of its properties. This contact is most often characterised as physical, where contagion occurs through touch, eating, drinking etc.

In many cases, this transferral of properties through contagion is characterised as a *negative* phenomenon, bringing about undesirable changes in the 'infected' individual or object. Indeed,

[45] The other law- similarity- will not be discussed here, but see Nemeroff & Rozin (1994).

it is this kind of contagion with which I will primarily be interested. However, it is worth noting that contagion is not always a bad thing. Indeed, there appear to be cases of *positive contagion*: that is, instances in which the value of an object can be increased through contact with a contagious entity. Here, the influence of the contagious entity on the object (and the resulting transferral of properties) is well-received. For instance, it is almost uniformly the case that an object which has come into contact with a famous person fetches a higher sale price than the exact same object absent of this relationship might (Hood & Bloom 2008, Frazier & Gelman 2009, Newman et. al 2011). Yet, the same phenomenon which gifts added value to objects formerly owned by notable persons can also serve to make the objects in question *profoundly undesirable*. Here contact with a contagious entity (microbes, dirt, faeces) serves to devalue the individual or object in question. However, social or moral properties also seem to be a significant source of negative contagion (for instance, where the person coming into contact with the object was evil, or hated by the participant), and rendered the object less valuable or desirable (see Nemeroff & Rozin 1994).

There is reason to think that contagion thinking may apply to mental illness. Namely, that quintessentialists may (in virtue of the belief that quintessence is causally potent and transferrable) think that mental illnesses (or aspects of them) may be 'caught' from mentally ill people by coming into contact with their quintessence. How might this be? I will start with the notion that 'aspects' of the illness may be transferred. Mental illnesses may be thought of as undesirable in two ways. For one, they are socially or morally problematic. Mental illnesses have long been associated with undesirable traits such as criminality, immorality, untrustworthiness, being incompetent etc. Perhaps individuals may fear that these morally and socially charged 'aspects' of mental illness may be transferred to them. In this way, a person with a mental illness may be seen as a source of negative contagion insofar as their quintessence

is thought to be associated with these properties. Individuals may avoid close contact with people with mental illnesses for fear of this kind of moral or social contagion, wherein they worry that contact will cause them to take on these negative characteristics. Here, quintessentialist thinking may encourage avoidance, fear and revulsion because individuals may worry that they will take on certain properties of mentally ill people.

Yet, there is another way in which mental illness is undesirable: it is an illness, and so almost inherently a bad thing to have. Indeed, the contagion principle is often applied where individuals fear catching an illness (which itself is negatively evaluated) from another person. Here, contagion may again apply to mental illness. Namely, that quintessentialists may (in virtue of their beliefs that quintessence is causally potent and transferrable) think that mental illnesses may be 'caught' from mentally ill people through contact with their quintessence: but 'caught' in the literal sense that individuals fear developing the same illness. That is, just as I might fear being infected with Tuberculosis, perhaps I also fear that I will 'catch' schizophrenia from someone who has it. In this way, we may avoid, malign and generally stigmatize people with mental illnesses in order to prevent catching them ourselves.

There has been quite a lot written on infectious disease, contagion and stigma. One prominent example is Kurzban & Leary's (2001) convincing analysis of the evolutionary origins of stigma, wherein they claim that what has traditionally been referred to as stigma is in fact a series of complex psychological mechanisms which have been cultivated by the processes of natural selection to deal with some of the problems inherent to being a social animal: simply put, stigma is an adaptation which allows human beings to limit contact with socially undesirable parties. Yet, surely there is no reason to think the same would apply for mental

illness? Indeed, it seems implausible that quintessentialism could engender a serious belief in a literal kind of contagion (i.e. infection). Whilst public mental health literacy is perhaps not as high as it should be, none amongst us truly believe that schizophrenia can be transmitted as *Staphylococcus* is.

Yet, whilst a literal medical contagion of mental illness seems initially implausible, there is reason to suggest that some of us do think in this way. It is easy to see how the misinformation spread by generics and quintessentialism (namely, that quintessence grounds significant properties and can be transferred) may not immediately be identified as false. For one, public knowledge of the biomedical model of germ transmission is relatively poor. Further, even among experts there is little known about the causes of mental illness. Perhaps this lack of knowledge, combined with the fact that the language used to describe mental illness is becoming increasingly biological (which itself may be misinterpreted), may lead to some ill-founded fears that one can catch a mental illness through close contact with someone who has one, or something they have touched.

Indeed, although it may strike many of us as preposterous, there is reason to believe that at least some members of the population believe that there is a physical risk of contagion from the mentally ill, akin to that posed by bacteria, viruses and the like. Marsh & Shanks (2014, pp.1020-10220 conducted an experiment utilising 122 undergraduate students, in which the students were presented with a series of mental and medical illnesses. They were then asked to consider to what extent they believed each disease to be communicable: "how likely you think it would be for someone to catch [disorder name] through close contact with someone with that disorder" on a scale of 0% to 100% chance. Participants were also asked to make some remarks

about how long they thought they would need to be exposed to the contagious entity to catch the disease. Interestingly, a small yet notable number of the participants (4%)[46] reported that mental illnesses could be transmitted via 'physical contact': for instance, by touching the same object as someone else, or being sneezed upon. Whilst this view was certainly in the minority, should the experiment prove to be representative of society as a whole, it may be the case that a substantial number of people hold beliefs such as this. Indeed, this is particularly concerning, given that we would hope that undergraduate students were relatively well-educated when compared to the general public.

More research needs to be done here. However, I suggest that we should not discount the possibility that some contagion beliefs about mental illness will take the form of a literal transmission (i.e. catching the disease). Indeed, what can be taken from this section is that generics, by encouraging quintessentialism, may also encourage contagion-style thinking about mental illness and the mentally ill. Namely, because quintessence is causally potent yet transferrable, in close contact there is a risk that one will come to take on the same properties as the mentally ill person, where the properties in question could be moral/social characteristics or the illness itself.

This is worrying, as contagion thinking about mental illness will likely contribute to (and may even constitute) stigma and the perception of the mentally ill as dangerous. Contagion thinking provides justification for avoidance or dislike, and will likely act as a barrier to equal treatment. Contagion-thinking may also frustrate contact strategies, which often require close contact between the mentally ill person and a member of the public. Many commentators (e.g. Corrigan

[46] For precise coding details, see Marsh and Shanks (2014, p.1021-1022).

2016, Rüsch & Xu 2017) have noted that *in vivo* or face to face contact strategies have been found to be better at breaking down stereotype than cases where an individual's story is merely read about or seen on video. Yet, if contagion thinking leads us to (perhaps explicitly, but likely implicitly) resist or avoid coming into contact with people with mental illnesses, it is likely that face to face contact strategies will be frustrated. People who quintessentialize mental illness may be unlikely to attend contact sessions for fear of taking on some of the negative (quintessence-based) properties of the ill person.

Indeed, contact strategies will likely encounter the same problem experienced by education strategies: namely, that those who attend are already 'on board' with the message, and so do not need to be targeted. Indeed, perhaps only those who do not already think of mentally ill people as potential sources of negative contagion will want to meet them. If this is the case, then attending to the particular kind of contagion feared may help in encouraging contact. That is, if individuals fear that they might 'catch' the illness, then perhaps education strategies to disseminate correct information about microbial transmission and the causes of mental illness may be required. Of course, this may be challenging given the relative complexity of the material and the general lack of knowledge available even to experts.

v. Drawing Together

Of course, as I will now explore, one way in which we might prevent contagion thinking (as well as the other undesirable effects of applying quintessentialism to mental illness listed above) is to challenge the linguistic form which encourages mental illness to be viewed in quintessential terms: generics. To recap briefly- in the sections above I suggested that generics

can be used to convey two broad kinds of knowledge structure about mental illness: those which are broadly accurate, and those which are not. Generics conveying inaccurate stereotypes, for instance (1), are problematic in that they generalize properties to members of quintessential kinds and can be complicit in pernicious forms of social stereotyping. However, generics like (2) are also problematic, for reasons apart from the accuracy of the knowledge structure they convey. That is, there is no issue with the content they express, but in using the generic, we encourage the mental illness category to be viewed as a quintessential kind. This is worrying as it is deeply unhelpful to think about mental illness categories in this way, for the reasons explored above.

Indeed, I suggest that using generics to talk about mental illness is seldom, if ever, appropriate. Mental illnesses are not quintessential kinds, and so should not be depicted this way. As such, whether the stereotype is accurate or not, I suggest that we should avoid generic utterances altogether when discussing mental illness. In doing so, we will hopefully reduce our chances of running into the difficulties outlined above (although, if Leslie is right that quintessentialism is innate and universal, we are unlikely to avoid thinking in this way altogether). Where we want to express an accurate knowledge structure about mental illness, we can of course still do so, but we should use a non-generic form. Although generics play a valuable role in our cognition (for Leslie, they represent a primitive form of generalising), we are able to convey knowledge structures in other ways.

As noted in section II, we cannot straightforwardly translate generics into quantified utterances. Thus, we should not think about trying to translate the generic into a quantified utterance, but rather, we should phrase the knowledge structure expressed in the generic in a different way.

Thus, for accurate stereotypes, we should attempt to express them as quantified utterances where possible: for instance, to say something like 'most people with schizophrenia experience delusions' or 'roughly half of people with schizophrenia experience delusions' rather than 'schizophrenics experience delusions'. Likewise, problematic stereotypes may be less concerning where the generic form (e.g. 'schizophrenics are dangerous') is avoided and the knowledge structure is expressed in different ways (e.g. 'very few people with schizophrenia are dangerous' or 'of the violent crimes committed in the last year, *x* were committed by people with schizophrenia').

Using quantified utterances is useful in several ways. Firstly, the form of expression does not propose to quintessentialize the kind, avoiding the difficulties outlined above. Secondly, quantifying strongly encourages us to attend to the truth conditions of an utterance. In explicitly quantifying an utterance, our attention is inevitably drawn to the relative prevalence with which category members demonstrate a property. This means that we might be better able to appreciate how common it is for someone with schizophrenia to be dangerous or experience delusions, for instance: a fact which might be obscured if we were to utilise the generic formulation. Crucially, as I will explore in a little more detail later, quantifying an utterance also ensures that it is the kind of thing that can be refuted, where generics tend to evade refutation due to truth conditional laxity and inertia. Finally, using quantified utterances to express knowledge structures allows us to retain the epistemic benefit of using it, whilst minimizing the potential ethical and epistemic harms which might result from our expressing it. In this way, it may contribute somewhat to alleviating the dilemma outlined in chapter 4.

In a forthcoming paper, Leslie suggests one further way in which we might shape our utterances to combat stigma: namely, avoiding generic noun phrases when describing characteristics of social identity. That is, we should use adjectives or descriptive phrases when ascribing social identity properties, rather than generic constructions or labelling nouns Anthony (2016, p.188). As Leslie (forthcoming, pp. 37-38) puts it:

> Instead of *labelling* a person as *a Muslim*, we might instead *describe* the person- if needed- as, say, *a person who follows Islam*, thus emphasizing that *person* is the relevant kind sortal, and that *following Islam* is a particular property that the individual happens to possess.

Leslie holds that even a minor change in the language we use may halt the formulation of striking property generics, thus reducing social stereotyping. This seems to be a plausible suggestion for mental illness. For instance, rather than using the labelling noun 'schizophrenics', it would be better to say 'people with schizophrenia'. In this way, the illness appears to be characterised as a property the person simply happens to possess at the time, rather than being construed as a necessary or unchanging feature of them. When talking about mental illness (either in public information campaigns, on TV, film, the media, or as clinicians, members of the public, or as category members ourselves), we would do well to ensure that we say 'people with *X*…' rather than '*Xs*….'. Doing so will hopefully go some way to reducing pernicious forms of social stereotyping. However, given the prevalence of generics, this may well be a considerable task, and will require constant monitoring.

V. What should we do when we Hear Generics?

It is likely that our obligations do not conclude there. Indeed, what are we to do if we hear generics used around us, or if we see a generic form used in the media etc.? Several suggestions

spring to mind, some of which will be appropriate, and others not. One initially plausible strategy might be to dispute or reject the generic. Indeed, when we hear something we disagree with, or which misrepresents a category member, our normal inclination is to challenge it. We normally do this by refuting it, or providing evidence to the contrary. For instance, if I overhear someone saying 'schizophrenics attack people', my initial intuition might be to say 'no they don't', and perhaps to offer up the relevant statistics to back up my refutation. Perhaps, as Leslie notes, I may try to provide a counterexample to the problematic claim: I may bring up an example of someone with schizophrenia who is not dangerous. In these cases, I attempt to discredit the utterance by pointing out its falsity, or by demonstrating that it does not reflect the actual state of affairs in the world faithfully.

However, this kind of strategy is unlikely to be as tractable for the rejection of generics as it would be for the rejection of quantified utterances. For instance, consider the generic 'mosquitos carry the West Nile virus'. Striking property generics like this one can occur even where just one member of the kind displays the named property. The generic is not obviously false in the same way its quantified analogue might be. For instance, the generic 'schizophrenics attack people' may also appear acceptable to us, given that at least some category member has demonstrated the property. In this case, it may resist factual refutation, just as the mosquito generic does.

Indeed, as Leslie and Anthony observe, striking property generics are particularly resistant to counterexample. The generic is likely to resonate with us even where we are presented with examples of people who do not fit the stereotype. One way of analysing this is to suggest that, (much as the literature on the limitations of contact theories suggests) people often tend to

make exceptions for counter-stereotypical people, but maintain an adherence to the rule or stereotype nonetheless. In this way, they might say 'well yes, this person with schizophrenia didn't attack me, but schizophrenics attack people all the same' (where 'schizophrenics' refers to the kind rather than any individual). Here, the contact may succeed in causing the individual to develop an individualised counter-stereotypical belief about a person with schizophrenia, but maintain the generic belief regardless on the grounds that it will be true of some category members, and thus retains a degree of truth or resonance. In this way, generics, and the prejudice they generate, may be relatively impervious to refutation or discrediting by counterexample.

Another way of looking at this is to claim that because generics have truth conditions which are both hard to analyse and not dependent upon the actual state of things, attempting to reject them on the grounds of their being inaccurate representations of the world may prove fruitless. This is because generic utterances can have a resonance beyond mere truth and falsity. As many generics seem acceptable to us even if they do not track the actual truth conditions in the world particularly well, attempting to challenge them by demonstrating that the world is not as they report may not prove useful.

i. Metalinguistic Blocking

If this is true, then what are we to do when we hear problematic generics used? One plausible response comes from Haslanger (2011). Haslanger argues that certain generics propagate problematic quintessentialist ideologies, which contribute to social stigma. I have suggested that generics like (1), but also interestingly, those like (2), do just this. In response to

problematic generics, Haslanger argues that we must strive to engage in what she calls *metalinguistic blocking*. She begins by noting that language and communication play a vital role in the dissemination of ideologies (problematic or not). Ideologies are not imposed upon us, but rather are created and constituted by our own choices and actions. Indeed, she suggests that all people are equally involved in the construction and representation of the world-particularly social life and its continuation. In this sense, we are all capable of shaping social reality in some way. Crucially, ideology and representation of the social world has a profound impact upon action: what we do, and do not do, is shaped by the ideologies and forms of consciousness which operate upon us, at both the conscious and unconscious level.

Ideology is thus "a background cognitive and affective framework that gives actions and reactions meaning within a social system, and contributes to its survival" (Haslanger 2011, p.3). According to Haslanger ideologies can be belief-like (akin to propositional statements), or they can take the form of primitive dispositions or habits (hegemony). An ideology might be said to be hegemonic where it is *particularly embedded* within the shared conception of social reality. The hegemony is those beliefs which are taken for granted and commonly accepted (to the point at which they seem like natural facts about the world), with their status as conceptions of social reality forgotten. As Haslanger (2011, p.4) puts it, the hegemony comes to be seen as natural, which it turn makes it natural because people treat it as such: "once we constitute our social world, descriptions of it not only appear true, but are true". The perception that a particular form of social reality is natural renders it immune from criticism as we cease to perceive it as social in origin.

For Haslanger, ideologies enter into the hegemony when they become common items within the linguistic common ground, and thus part of what is widely understood to be *background knowledge*. For communication to be at all possible, we must assume that there is such a thing as the common ground, or a set of common assumptions. Indeed, communication requires that both parties assume that there are a common set of presuppositions, and that their conversational partner is also aware of them. How are these presuppositions selected? It is worth noting that the common ground can vary between social locales, and even between different conversational partners.

However, generally speaking, Haslanger argues that we are reliant upon conversational maxims like those put forward by Grice. The common ground is founded on something resembling the cooperative principle, although the particular content of the presupposed background knowledge will obviously vary in accordance with what you might reasonably expect each partner to know. The common ground can be added to over the course of an interaction. I can add new presuppositions to the common ground by simply asserting them. For instance, if I say 'my dog is called Jasper' in a conversation, then this adds several presuppositions to the common ground: that I have a dog, that it is likely a male etc. The common ground may also be added to via implicature or omission (i.e. 'at least *one* thing has gone right today' adding the presupposition that lots of things have gone wrong today).

However, as Haslanger notes, we are not passive in the formation of ideologies, nor powerless regarding the content of the linguistic common ground. Should my conversational partner attempt to add an objectionable presupposition to the common ground (for instance one I know to be false), then I am entirely at liberty to reject it. This may be conducted via a simple counter-

assertion or negation, either in regard to the literal content of the statement, or the implicature, or both. For instance, should you say to me 'Joe is coming home at 5pm tonight' but I know that he had to stay late at work, then I can simply reply by saying 'no, he has to stay late so he'll be home at 7pm'. My linguistic act effectively blocks the initial presupposition from being accepted into the common ground, and adds another. This is an instance of *metalinguistic blocking*: actively preventing some undesirable presupposition from entering the common ground[47].

Haslanger believes that this simple case provides us with the template to challenge generics, even where they seem true. That is, for one to engage in metalinguistic blocking, it need not be the case that the objectionable presupposition is literally false. For instance, an act of metalinguistic blocking denies the problematic implicature even whilst the statement remains *literally true*. If I hear the generic 'women are submissive' (which may be literally true in that it may statistically be the case that many women depend on men for their livelihoods) I can prevent it from entering the linguistic common ground by saying 'no, that's not right', or even 'it isn't helpful to say that'. These responses are designed to block the entry of the problematic presupposition into the linguistic common ground (which it itself pragmatically implied by the generic). However crucially, they do so without making reference to truth conditions or truth/falsity. Indeed, in metalinguistic blocking we do not reject the generic on the grounds that it is empirically false: rather, we reject it because it pragmatically implicates harmful falsehoods. The generic is not being rejected because it misrepresents the frequency with which the named subject demonstrates the predicate.

[47] This blocking move need not necessarily add anything new to the common ground.

Rather, it is rejected because it misrepresents the *nature* of the subject-predicate relationship (in this case, it suggests that women are submissive *by nature*). Whereas refutation works by claiming that an utterance is false or not factually accurate, metalinguistic blocking works by pointing out that the generic smuggles in problematic ideologies, and so recommends that they do not enter into the linguistic common ground. In this sense, the rationale for rejecting the generic differs between refutation and metalinguistic blocking: the former claims that the generic does not accurately reflect the state of things in the world, and the latter claims that the generic smuggles in problematic ideologies.

I have suggested that refutation is not an adequate means of rejecting generics, for generics have a resonance beyond truth conditions. Instead, if we are to reject generics, we should do so via the mechanism of metalinguistic blocking, on the basis that generics smuggle in problematic ideologies. However, how are we to capture this act linguistically? Some utterances will not be suitable. For instance, if I were to hear the generic 'schizophrenics experience delusions' and attempt to block it from entering the common ground, I cannot simply say 'no, that's not true'. For one, it may appear that I am trying to *refute* the generic (and so claim that it should be rejected on the grounds of inaccuracy/ falsity). In this case, my rejection may not succeed as someone can simply say 'well, lots of people with schizophrenia do experience delusions', and thus the generic stands. Even in cases where the generic is not particularly accurate (for example, 'schizophrenics attack people'), attempting to refute it by saying 'no, that's not true' mischaracterises the rationale behind the rejection, and makes it appear that the generic should be rejected on the grounds of inaccuracy/ falsity. Thus, it does not capture the rationale behind metalinguistic blocking. Saying something like 'no, that's not true' or 'that's wrong' may also backfire in other ways. For instance, when said in response to the generic 'schizophrenics experience delusions', it may inadvertently suggest that the

opposite is true (i.e. that people with schizophrenia don't experience delusions), which is also not the desired effect.

Thus, we must find another way of responding to generics about mental illness which better captures our reason for rejecting them: namely, that they smuggle in problematic and overly general quintessentialist ideas. There are some utterances which are suitable for this. For instance, if I were to hear the generic 'schizophrenics experience delusions', I can attempt metalinguistic blocking by saying something like 'people with schizophrenia have all kinds of individual differences' or 'some people with schizophrenia experience delusions, but we can't generalise'. Likewise, I might respond to the generic 'schizophrenics attack people' by saying 'it's not helpful to generalise like that'. I may also explicitly point out why generalising like this is unhelpful, and how it may contribute to stigma or poor treatment of people with mental illnesses. Further, after carrying out metalinguistic blocking, I could express the same knowledge structure in a quantified form. Thus, the utterances we use to reject generics should avoid reference to falsity or incorrectness, but rather, they should highlight that the use of the generic could lead to misleading presuppositions.

It is because of this that refuting generics is distinctively difficult: we must find a linguistic expression which encapsulates the act of metalinguistic blocking, and which does not appear to be an act of refutation. Often, the kinds of utterances which will be appropriate for this task will be like those outlined above ('it's not helpful to generalise' or 'people with schizophrenia have all kinds of individual differences'). These capture what is objectionable about the generic, but may well sound quite weak or nit-picky, thus adding to the challenge posed by metalinguistic blocking. Indeed, metalinguistic blocking may not always be well-received for

this reason, but nonetheless, it is a valuable means of preventing the spread of problematic ideologies.

ii. Going Forward

Thus, even though generics resonate with us in a way which goes beyond mere truth conditions, we can still reject them- although not on the grounds of empirical falsity. We are, as Lewis (1979) observes, 'scorekeepers in a language game', and capable of exercising some power over the content of the linguistic common ground. In this way, we can deny generics entry into the common ground whether they are 'true' or not, effectively bypassing the difficulties regarding the factual refutation of generics. Sustained metalinguistic blocking can, Haslanger argues, prevent problematic ideologies from entering the linguistic common ground, thus in turn preventing them from conveying or constructing problematic forms of social reality. Hence, one way of challenging generics about mental illness is to carry out metalinguistic blocking: to refuse them entry into the common ground. Indeed, this might broadly be characterised as a form of protest strategy, in which problematic depictions of mental illness are challenged. As I have argued, this applies to all generics about mental illness, not merely those which convey inaccurate knowledge structures.

If we want to eradicate stigma, we must appreciate the role of language in constructing forms of social reality, understand which linguistic forms are problematic, endeavour to not use them ourselves, and challenge others through metalinguistic blocking where appropriate. For the first two points, perhaps a simplified public education strategy would be plausible. Indeed, following the Zarpie experiments, Rhodes and collaborators have written some public

engagement pieces which are very accessible, and could potentially explain generics and their links to stigma to non-philosophers (Rhodes 2012). Perhaps material such as this could be incorporated into an education campaign and disseminated accordingly.

Yet, it is likely that the final step- the moderation of our own behaviour and challenging others- is likely to be challenging or contentious for several reasons. For one, it will be labour-intensive, particularly given that generics are often used. Further, challenging others on what they say may not always be met favourably. Calling others out on making contributions to the linguistic common ground which are not acceptable may cause some tension or embarrassment, and individuals may be unwilling to challenge generics for these reasons. Indeed, a similar problem can be evidenced where individuals fail to challenge the use of damaging labels like 'crazy', 'psycho' or 'nuts' for fear of being accused of 'political correctness gone mad', or some analogous statement.

This is likely to be a significant issue. Perhaps the best way an individual might carry out metalinguistic blocking is primarily around those they are close with. In this sense, where individuals do not need to worry as much about receiving negative reactions, they might gently correct family members or friends in order to gain confidence. Indeed, it is worth noting that the manner in which metalinguistic blocking is carried out is likely to have a great bearing on how it is received. Whilst it seems obvious, metalinguistic blocking will likely be most successful where it is firm and assertive, but not aggressive, patronising or smug.

This strategy may face further challenges. For one, one might wonder how successful an individual carrying out metalinguistic blocking can be- how much change will one person

quibbling about language bring about? This will obviously depend upon the social status and power of the individual, yet for most of us, we might wonder whether our actions will ever effect meaningful change. In response to this, I suggest that even at the personal level, metalinguistic blocking may well be effective in altering the common ground amongst those close to oneself. Given that even individual acts of prejudice or discrimination may be very upsetting, altering the mindset of just one potential stigmatizer is likely to be worthwhile.

Yet, the most effective mode of conducting metalinguistic blocking will likely be to join together to form an organisation or group which then conducts the challenge. In this way, bodies of people conducting metalinguistic blocking- perhaps by challenging generics which are visible in the public sphere in film, television or even policy- could exert a huge impact upon the common ground in which great many of us share. Once again, I suggest that this might provide a role for advocacy. Advocacy is conducted at both the individual and the collective level, with collective advocacy being defined by the Scottish Independent Advocacy Alliance (2017, no pagination) as enabling "a peer group of people, as well as a wider community with shared interests, to represent their views, preferences and experiences", where "the group as a whole may campaign on an issue that affects them". The rationale behind collective advocacy is that groups are harder to ignore than individuals, and that raising difficult issues in a group can alleviate many of the difficulties associated with raising said issues on one's own (for instance, stress, fear, isolation or worry). In much the same way, collective advocacy could speak out against generic language on the grounds that it creates problems for people with mental illnesses, and may contribute to the poor treatment they receive.

VI. Conclusion

The aim of this chapter has been to demonstrate that generic utterances plausibly contribute to the stigma of mental illness. I have done this by outlining what generics are, and how it is that they are thought to encourage the quintessentialization of the categories they describe. I have suggested that generics are a problematic way to describe mental illness on two counts. Firstly, where a knowledge structure is 'inaccurate'- that is, where it does not describe reality very faithfully- phrasing that knowledge structure as a generic can obscure this. That is, because generics demonstrate such a strange relationship with truth conditions, we often find that generics are acceptable to us, even where category members very rarely demonstrate the named property. Leslie notes that this is very commonly the case where the property in question is a striking one. In this way, even if just one member of a category demonstrates a striking property, the property can be simplistically attributed to the broader kind. In this way, generics can contribute to damaging forms of social stereotyping.

Yet, we might also use generics to convey 'accurate' stereotypes about mental illness: for instance, 'schizophrenics experience delusions'. I have suggested that the use of the generic form is problematic, even where the knowledge structure behind it may not be. This is because quintessentializing mental illness kinds is deeply unhelpful, and engenders misinformation and social distance. As such, I have suggested that to fight stigma, we should attend to the use of language. To combat stigma we should avoid using generics to talk about mental illness, and instead use quantified utterances. We must also seek to challenge generic utterances where we find them, and insist on quantified utterances in their stead. I have suggested that whilst generics appear to evade direct refutation, one promising mode of challenging them is to

engage in what Haslanger terms ‘metalinguistic blocking’: where the challenge in question does not claim that the generic is strictly ‘false’, but rather claims that the use of the generic pragmatically implicates a falsehood which can lead to the cultivation of damaging forms of social reality. I have acknowledged that fighting stigma by attending to language is likely to be a time-consuming process, and one which might prove quite difficult. Yet, I suggest, the project is a necessary one, and will likely be fruitful.

FINAL THOUGHTS

I. Summary

This thesis has centred around philosophical perspectives on the stigma of mental illness. Given that each chapter has contained short summary, I will recap the shape of the work only briefly here. Chapter 1 served as an introduction to the problem of the stigma of mental illness. It outlined what stigma was and the forms it can take before demonstrating the myriad harms it creates for people with mental illnesses. It also outlined the three types of strategies used to combat the stigma of mental illness: education, contact and protest.

Chapter 2 made the point that stigma extends beyond explicit expressions, and that an individual may act in a discriminatory manner even where doing so goes against her explicit and avowed beliefs. If we are to tackle the stigma of mental illness, we must tackle implicit stigma. Yet, this will likely require methods other than traditional education strategies, which are ill-suited to tackling forms of implicit social cognition. Indeed, I put forward a method of selecting 'target groups' for anti-stigma interventions who would both be easy to deliver training to (as they are commonly members of professional bodies) and whom are at greatest risk of carrying out unintentional discrimination (in virtue of their being asked to make decisions for and about people with mental illnesses, whilst existing in working climates which prime the activation of type-1 processes).

Chapter 3 focussed upon epistemic injustice as one form of status loss and discrimination to which people with mental illnesses are often exposed. It discussed Fricker's notion of epistemic injustice, before noting that establishing what constitutes an 'undue' deficit in credibility is a complicated matter for mental illness. As such, I suggested that accurate stereotypes may be consulted when making credibility judgements, but this should be done only where necessary, and in line with a 'capacity' approach: that is, credibility should be established relative to the task required, and we should be ready to revise these judgements where appropriate. This chapter also explored issues of mental illness and stereotype threat, before suggesting that this phenomenon may explain some of the behaviour displayed by people with mental illnesses.

Chapter 4 discussed potential ways of intervening on the stigma mechanism to halt mental illness stigma. The first suggestion- that we could potentially cease labelling- was dismissed on the grounds that much of the language used to mark out and stigmatize mental illness is the very same as that used by members of the psychiatric profession, people with mental illnesses and the well-intentioned public. To get rid of the labels used to stigmatize is also to get rid of labels which are hugely beneficial and practically necessary. For this reason, the strategy is untenable. Chapter 4 also discussed the possibility of intervening on stereotype, and described a potential ethical-epistemic dilemma which may occur as a result of disregarding stereotype. I tentatively suggested that this may be fruitful, but establishing which stereotypes to maintain and which to disregard would be a time-consuming matter, and would only be possible where public mental health literacy is suitably high.

Chapter 5 focussed upon how language can contribute to stigma. It explored how generics about mental illness contribute to insidious forms of social stereotyping (in the way Leslie

2013, 2014 describes), yet are also problematic in that they encourage quintessentialist thinking about the category and its members. As I have outlined, thinking in this way about mental illness is deeply unhelpful, and generates misinformation about the severity and permanence of psychiatric disorder, whilst also setting the categories up as particularly rich bases for induction and prediction. Insofar as generics may also promote entiative and contagion-style thinking, they will exacerbate social distance, encourage separation, contribute to stigma and frustrate current anti-stigma initiatives. As such, I have suggested that we should avoid the use of the generic form altogether when talking about mental illness (whether the knowledge structure conveyed by the generic is 'accurate' or not), and instead use quantified utterances where appropriate. Where we do hear generics (either in common discourse, used by authority figures, or in policy or in the media), we should reject their entry into the linguistic common ground by carrying out metalinguistic blocking (Haslanger 2011).

II. Recommendations for Action

In this work, some practical recommendations for combatting the stigma of mental illness have been suggested. At this stage, it would be useful to tease these out. One of the most pressing recommendations is that research into implicit bias and mental illness should be conducted. Indeed, chapter 2 demonstrated that stigma can (and likely does) occur through implicit mechanisms. I have suggested that if we are to combat the stigma of mental illness, we would do well to turn our attention to the possibility (and indeed, likelihood) that there are implicit biases about mental illness, and that these may cause discriminatory behaviour whilst being introspectively inaccessible to the individual.

I have suggested that there is good reason to think that such biases exist, but in order to formulate strategies to combat them, we must know how numerous they are, what their content is, and how they impact upon behaviour. That is, we must discover more precisely what these biases are and uncover their heterogeneity if we are to formulate strategies to combat them. Thus, I suggest that research into implicit biases and mental illnesses will be a crucial next step in the fight against stigma. Indeed, it will be necessary if we are to formulate precise strategies to combat biases. Research of this kind if already being conducted by Project Implicit, but could be hugely expanded, with more of a focus on biases which are likely to contribute directly to stigma (for instance, biases regarding mental illnesses and dangerousness).

I have also recommended that education strategies be pursued in order to combat mental illness stigma. It is noteworthy that in order for the strategy of intervening on stereotype to halt stigma (as described in chapter 4) to be successful, a good degree of public mental health literacy is required. Indeed, this would also be beneficial when considering implicit bias, epistemic injustice and generics about mental illness. As such, a general recommendation can be made that mental illness stigma will be easier to fight if the public are better informed about mental illness and how it presents more generally. Thus, a general public education campaign would be useful.

However, more specific education strategies can also be suggested. Although I have acknowledged that they may have their limitations, it seems advisable that education programmes about implicit social cognition be initiated (specifically to target groups). These will be basic at first, but later furnished with the data about the precise nature of implicit biases about mental illness uncovered from the research suggested above. Although we are not in a

position at present to say much about the specific biases we have and how precisely to address them, an education programme would be useful in that it would alert the public to the possibility that they may well demonstrate behaviours which they would not reflectively endorse. Hence, I would recommend that an education programme about implicit social cognition be launched.

Education on other topics may also prove fruitful in the fight against mental illness stigma. Much like implicit bias, our practices of assigning credibility are often not transparent to us, and occur below the level of conscious awareness. Yet, as Fricker (2007) has demonstrated, these practices can often result in injustice, and I suggest, in stigma more broadly. As such, alerting the public to potential injustices in our epistemic practices is an important part of combatting the stigma of mental illness. The literature on epistemic injustice has become expansive indeed, which may be an indication that it is relatively 'wieldy' as philosophical concepts go. That is, it can be explained in simple terms and applied to many aspects of daily life. For this reason, it may well be suitable for a public education campaign, or at least, for delivery to advocates (whose practice centres upon combatting epistemic injustice against service users with mental health issues). Indeed, I suggest that educating advocates and those who frequently come into contact with people with mental health issues about epistemic injustice will be very useful.

Further still, Rhodes (2012) has produced some work about generics, their acquisition, and the beliefs/ attitudes they pragmatically implicate which has been tailored for non-academic audiences. In much the same way as above, this suggests that educating the public or specialist groups about generic utterances and the beliefs they encourage may be very possible. Indeed, explaining why generic utterances are problematic may well be more effective than simply

telling people not to use them, or criticising them where they do- the reason being that if an explanation is offered then people may better appreciate the purpose in monitoring the language they use. Indeed, explaining why it is problematic to talk about mental illnesses in generic terms will hopefully create a motive or imperative for members of the public to examine and modify their own conduct. Hence, education about generics and the effects they have on our perceptions of people and categories may be very useful in the fight against mental illness stigma.

The crux of all the education strategies outlined above is to raise awareness of processes which may contribute to mental illness stigma, but of which we are not usually cognisant (i.e. acting on implicit biases, making judgements about the epistemic capabilities of individuals based on prejudicial stereotypes, using language which encourages further problematic beliefs). In this way, education allows us to make the public aware of the many things they do which might contribute to the stigma of mental illness. Yet, of course, this will be but the first step. Indeed, I acknowledged in chapter 2 that any attempts to combat implicit bias will likely be time-consuming and difficult. The same will likely be true of combatting epistemic injustice and the use of generics about mental illness.

Indeed, attempting to monitor one's epistemic practices such that one only makes said judgements where necessary, and only insomuch as is required by the circumstance, will likely prove difficult and a matter of some personal endeavour. Likewise, monitoring one's language and that of those around oneself may require constant vigilance and some significant effort. This thesis recommends that individuals carry out exactly these tasks, whilst also taking measures to assess whether they are afflicted by implicit biases. Yet, more structural responses

can also be recommended. We must examine institutions, professions, policies and procedures to make sure that they do not smuggle in unwarranted or *a priori* assumptions about the epistemic deficiencies of people with mental illnesses. Similarly, we must ensure that generics about mental illness are not used by those in positions of influence (e.g. healthcare professionals, the media, persons of power), nor utilised in legislation, policy, research and clinical practice.

Thus, if we want to combat the stigma of mental illness, we must examine and modify our own conduct on a regular basis, whilst also subjecting others, institutions, the media and policies to scrutiny. This programme of critical reflection should ideally be accompanied by protest where problematic depictions of mentally ill people, problematic knowledge structures or generic language about mental illness are used. In this way, protest may turn out to be a good way of breaking down common associations between mental illness and negative characteristics by limiting our exposures to them. This may help to prevent the activation of type-1 processes (by limiting the circumstances under which associations are primed) or reduce the number of situations in which stereotypes seem salient.

These recommendations are only the beginning, but I suggest that they follow from the application of philosophical tools of analysis to the issue of the stigma of mental illness. There is much more work to be done here. Indeed, I think it likely that there is much more to be gained from the application of philosophy to debates on stigma and mental illness generally. I hope that in future this area in the literature will continue to grow, and produce further recommendations as to how we can best combat the stigma of mental illness.

BIBLIOGRAPHY

ABIRI, S., OAKLEY, L. D., HITCHCOCK, M. E. and HALL, A. 2016. Stigma Related Avoidance in People Living with Severe Mental Illness (SMI): Findings of an Integrative Review. *Community Mental Health Journal*. **52**(3), pp. 251-261.

AMERICAN PSYCHOLOGICAL ASSOCIATION. 2013. *Diagnostic and Statistical Manual: Mental Disorders*. 5th Edition. Washington, DC: American Psychiatric Association.

AMODIO, D. M. and DEVINE, P. G. 2006. Stereotyping and evaluation in explicit race bias: Evidence for independent constructs and unique effects on behaviour. *Journal of Personality and Social Psychology.* **91**, pp. 652-661.

ANTHONY, L. M. 2016. Bias: Friend or Foe? Reflections on Saulish Scepticism. *In:* M. BROWNSTEIN and J. SAUL (eds.) *Implicit Bias and Philosophy, Volume 1.* Oxford: Oxford University Press, pp. 157-190.

ANGERMEYER, M., and MATSCHINGER, H. 1994. Lay beliefs about schizophrenic disorder: the results of a population survey in Germany. *Acta Psychiatricia Scandinavica. Supplementum.* **382**, pp. 39-45

ANGERMEYER, M., and MATSCHINGER, H. 2003. Public beliefs about schizophrenia and depression: Similarities and differences. *Social Psychiatry Psychiatric Epidemiology*. 38, pp. 526-53.

ATRAN, S., MEDIN, D. L, LYNCH, E., VAPNARSKY, V., UCAN EK' and SOUSA, P. 2001. Folkbiology Doesn'tCcome from Folkpsychology: Evidence from Yukatec Maya in Cross-Cultural Perspective. *Journal of Cognition and Culture.* **1**, pp. 4–42.

BACKER, T. E. 1985. Powers Untapped: Enhancing Mass Media Depictions of Mental Illness. *Proceedings of the First Annual Rosalynn Carter Symposium on Mental Health Policy: Stigma and the Mentally Ill.* Atlanta: Emory University School of Medicine, pp.18-29.

BANAJI, M. and GREENWALD, A. 2013. *Blindspot*. New York: Delacorte Press.

BANAJI, M. and HARDIN, C. 1996. Automatic Stereotyping. *Psychological Science.* **7**, pp. 136-141.

BELLACK, A. S., MORRISON, R. L. WIXTED, J. T. and MUESER, K. T. 1990. An Analysis of Social Competence in Schizophrenia. *British Journal of Psychiatry.* **156**, pp. 809-818.

BERTRAND, M., CHUGH, D. and MULLAINATHAN, S. 2005. Implicit discrimination. *American Economic Review*, pp. 94–98.

BLAIR, I., STEINER, J. and HAVRANKEK, E. 2011. Unconscious (implicit) bias and health disparities. *The Permanente Journal.* **15**(2), pp.71-78.

BORTOLOTTI, L. 2009. *Delusions and Other Irrational Beliefs.* International Perspectives in Philosophy and Psychiatry. Oxford, Oxford University Press.

BORTOLOTTI, L. 2017. *The Three Stigmas about Mental Health we need to Deconstruct.* TED Talk at TEDXBRUM. Presented October 2017. Available from: http://www.lisabortolotti.com/ [Accessed 04/01/2018].

BOYCE, J. et al. 2015. Mental health and contact with police in Canada, 2012. *Catalogue no. 85-002X*. Ottawa, ON: Juristat, Canadian Centre for Justice Statistics, Statistics Canada [Accessed March 17, 2016]. Available from http://www.statcan.gc.ca/pub/85-002-x/2015001/article/14176-eng.htm?WT.mc_id=twt

BRENER, L., ROSE, G., VON HIPPEL, C. and WILSON, H. 2013. Implicit attitudes, emotions, and helping intentions of mental health workers toward their clients. *Journal of Nervous and Mental Disease.* **201**(6), pp. 460-463.

BROWNSTEIN, M. 2016. Implicit Bias. *In:* E. N. ZALTA (ed.) *The Stanford Encyclopaedia of Philosophy* [online]. Available from: https://plato.stanford.edu/archives/spr2017/entries/implicit-bias/>.

BROWNSTEIN, M. 2017. What can we learn from the Implicit Association Test? A Brains Blog Roundtable. *Philosophyofbrains.com* [online]. Available from: http://philosophyofbrains.com/2017/01/17/how-can-we-measure-implicit-bias-a-brains-blog-roundtable.aspx

BROWNSTEIN, M. and SAUL, J. 2016. *Implicit Bias and Philosophy.* Oxford: Oxford University Press.

CARLSON, G. 1977. *Reference to Kinds in English*. PhD. Dissertation University of Massachusetts.

CIMPIAN, A. and ERICKSON, L. C. 2012. The effect of generic statements on children's causal attributions: Questions of mechanism. *Developmental Psychology*. **48**(1), pp. 159-170.

CIMPIAN, A., and MARKMAN, E. M. 2009. Information learned from generic language becomes central to children's biological concepts: Evidence from their open-ended explanations. *Cognition.* **113**(1), PP. 14-25.

CIMPIAN, A., and MARKMAN, E. M. 2011. The generic/ nongeneric distinction influences how children interpret new information about social others. *Child Development.* **82**(2), pp. 471-492.

COOK, J. A. 2006. Employment Barriers for Persons with Psychiatric Disabilities: Update of a Report for the President's Commission. *Psychiatric Services.* **57**(10), pp. 1391-1405.

COOPER, J. E. et. al. 1972. *Psychiatric Diagnosis in New York and London.* London: Oxford University Press.

COOPER, R. 2007. *Psychiatry and Philosophy of Science*. Stocksfield: Acumen.

COOPER, R. 2014. *Diagnosing the Diagnostic and Statistical Manual of Mental Disorders*. London: Karnac Books Ltd.

CORRELL, J., PARK, B. JUDD, C. and WITTENBRINK, B. 2002. The police officer's dilemma: Using race to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*. **83**, PP. 1314–1329.

CORRIGAN, P. W. 1998. The Impact of Stigma on Severe Mental Illness. *Cognitive and Behavioural Practice.* **5**, pp. 201-222.

CORRIGAN, P. W. 2000. Mental Health Stigma as Social Attribution: Implications for Research Methods and Attitude Change. *Clinical Psychology.* **7**(1), pp. 48-67.

CORRIGAN, P. W. 2004. How stigma interferes with mental health care. *The American Psychologist.* **59**(7), pp. 614-625.

CORRIGAN, P.W. 2007. How Clinical Diagnosis Might Exacerbate the Stigma of Mental Illness. *Social Work*. **51**(1), pp.31-39.

CORRIGAN, P. W. 2016. Lessons learned from unintended consequences about erasing the stigma of mental illness. *World Psychiatry.* **15**(1), pp. 67-73.

CORRIGAN, P. W. and CALABRESE, J. D. 2005. Strategies for assessing and diminishing self-stigma. *In:* P.W. CORRIGAN (ed.) *On the stigma of mental illness. Practical strategies for research and social change.* Washington, DC: American Psychological Association, pp.239–256.

CORRIGAN, P. W., DRUSS, B. G., and PERLICK, D. A. 2014. The impact of mental illness stigma on seeking and participating in mental health care. *Psychological Science in the Public Interest*. 15, pp. 37–70.

CORRIGAN, P. W. and KLEINLEIN, P. 2005. The impact of mental illness stigma. *In:* P. W. CORRIGAN (ed.) *On the stigma of mental illness: practical strategies for research and social change*. Washington: American Psychological Association, pp. 11–44.

CORRIGAN, P. W., MICHAELS, P. J. and MORRIS, S. 2015. Do the effects of anti-stigma programs persist over time? Findings from a meta-analysis. *Psychiatric Services*. **66**(5), pp. 543–546.

CORRIGAN, P. W., MORRIS, S. B., MICHAELS, P. J., RAFACZ, J. D. and RÜSCH, N. *2012.* Challenging the public stigma of mental illness: a meta-analysis of outcome studies. *Psychiatric Services.* **63**, pp. 963-73.

CORRIGAN, P. W. and PENN, D. L. 1999. Lessons from social psychology on discrediting psychiatric stigma. *The American Psychologist.* **54** (9), pp. 765-776.

CORRIGAN, P. W. and WATSON, A. C. 2002. Understanding the impact of stigma on people with mental illness. *World Psychiatry.* **1**(1), pp. 16-20.

CORRIGAN, P. W., WATSON, A. C., and BARR, L. 2006. The self-stigma of mental illness: Implications for self-esteem and self-efficacy. *Journal of Social Clinical Psychology*. **25**, pp. 875–884.

CORTINA, L. M. 2008. Unseen Injustice: Incivility as Modern Discrimination in Organizations. *Academy of Management Review.* **33**(1), pp. 55-75.

CORTINA, L. M. et al. 2011. Selective Incivility as Modern Discrimination in Organizations: Evidence and Impact. *Journal of Management.* **20**(10), pp. 1-27.

CROCKETT, M. 2013. Models of Morality. *Trends in Cognitive Science.* **17**(8), pp. 363-366.

CUTCLIFFE, J. and HAPPELL, B. 2009. Psychiatry, mental health nurses, and invisible power: Exploring a perturbed relationship within contemporary mental health care. *International Journal of Mental Health Nursing.* **18**, pp. 116-125.

DASGUPTA, N. 2013. Implicit attitudes and beliefs adapt to Situations: A decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept. *Advances in Experimental Social Psychology.* **47**, pp. 233-279.

DE HOUWER, J. 2006. What are implicit measures and why are we using them. *In:* WEIRS, R. W. and STACY, A. W. (eds.) *The Handbook of Implicit Cognition and Addiction.* Thousand Oaks, CA: Sage, pp.1-28.

DE HOUWER, J., TEIGE-MOCIGEMBA, S., SPRUYT, A. and MOORS, A. 2009. *Psychological Bulletin.* **135**(3), pp. 347-368.

DEVINE, P. G. 1989. Prejudice and out-group perception. *In*: A. TESSER (ed.) *Advanced social psychology*. NewYork: McGraw-Hill, pp. 467–524.

DIESENDRUCK, G. 2001. Essentialism in Brazilian Children's Extensions of Animal Names. *Developmental Psychology*. **37**, pp. 49–60.

DOVIDIO, J. F., KAWAKAMI, K., and GAERTNER, S. L. 2002. Implicit and Explicit Prejudice and Interracial Interaction. *Journal of Personality and Social Psychology.* **82**(1), pp. 62-68.

DRICKEY, J. 1990. Reducing Stigma: A Discussion About Accurately Portraying People with Serious Mental Illness. *Television and Families*, pp.30-35.

DU BOIS, W. E. B. 1903. *The Souls of Black Folks.* Chicago: A. C. McClurg & Co.

EGAN, A. 2011. Comments on Gendler's 'The epistemic costs of implicit bias'. *Philosophical Studies.* **156**, pp. 65-79.

ESTROFF, S. E. 1989. Self, Identity, and Subjective Experiences of Schizophrenia: In Search of the Subject. *Schizophrenia Bulletin.* **15**(2), pp.189-196.

EVANS, E. M. 2001. Cognitive and Contextual Factors in the Emergence of Diverse Belief Systems: Creation versus Evolution. *Cognitive Psychology*. **42**, pp. 217-266.

FARINA, S. 1998. Stigma. *In:* K. T. MUESER and N. TARRIER (eds.) *Handbook of social functioning in schizophrenia*. Needham Heights: Allyn & Bacon, pp. 247–279.

FELDMAN, D. B., and CRANDALL, C. S. 2007. Dimensions of mental illness stigma: What about mental illness causes social rejection*? Journal of Social and Clinical Psychology*. 26, pp. 137–154.

FISKE, S. T. 2000. Stereotyping, prejudice, and discrimination at the seam between the centuries: evolution, culture, mind, and brain. *European Journal of Social Psychology.* **30**, pp. 399-422.

FORSCHER, P. S. et. al. 2013. A Meta-Analysis of Change in Implicit Bias. *Open Science Framework* [online]. Available from: https://osf.io/awz2p/

FOY, S. L. 2013. Branded: How Mental Disorder Labels Alter Task Performance in Perception and Reality. *Dissertation submitted to Duke University.* Available from: https://pdfs.semanticscholar.org/8cdc/262fde1b4fbf8ca82f0eb6a7876681fbb06c.pdf

FRANCES, A. 2012. DSM-F Field Trials Discredit the American Psychiatric Association. Huffington Post Science. The Blog. Posted 31 October 2012. Available at: www.huffingtonpost.com/allen-frances/dsm-5-field-trials-discre_b_2047621.html. [Accessed 5/2/2015].

FRANCES, A. 2013. *Saving Normal*. New York: Harper Collins.

FRANKISH, K. 2016. Playing Double: Implicit Bias, Dual Levels, and Self-Control. *In:* M. BROWNSTEIN and J. SAUL (eds.) *Implicit Bias and Philosophy, Volume 1.* Oxford: Oxford University Press, pp. 23-46.

FRANKISH, K. 2017. What can we learn from the Implicit Association Test? A Brains Blog Roundtable. *Philosophyofbrains.com* [online]. Available from: http://philosophyofbrains.com/2017/01/17/how-can-we-measure-implicit-bias-a-brains-blog-roundtable.aspx

FRAZER, J. G. 1976. *The Golden Bough: A Study in Magic and Religion.* London: Macmillan & Co.

FRAZIER, B. N. and GELMAN, S. A. 2005. Adults' ratings of different types of authentic objects. Poster to be presented at the 2005 American Psychological Society Annual Convention, Los Angeles, CA.

FRAZIER, B.N. and GELMAN, S.A. 2009. Developmental Changes in Judgments of Authentic Objects. *Cognitive Development.* **24**, pp.284-292.

FRICKER, M. 2007. Epistemic Injustice: Power and the Ethics of Knowing. Oxford Scholarship online. Available from: DOI: 10.1093/acprof:oso/9780198237907.001.0001

FRIDELL, J. and STRAUB. 2016. Implicit Bias versus the "Ferguson Effect": Psychosocial Factors Impacting Officers' Decisions to Use Deadly Force. *The Police Chief*.

GAEBEL, W., RÖSSLER, W. and SARTORIUS, N. (eds.) 2017. *The Stigma of Mental Illness- End of the Story?* Switzerland: Springer International Publishing.

GALISON, P. 1997. *Image and Logic: A Material Culture of Microphysics*. Chicago, IL: University of Chicago Press.

GARAND, L., LINGLER, J.H., CONNER, K.O. and DEW, M.A. 2009. Diagnostic Labels, Stigma, and Participation in Research Related to Dementia and Mild Cognitive Impairment. *Research in Gerontological Nursing*. **2**(2), pp.112-121.

GARLOW, S., ROSENBERG, J. and MOORE, J. et al. 2008. Depression, desperation, and suicidal ideation in college students: Results from the American foundation for suicide prevention college screening project at Emory University. *Depression Anxiety*. **25**, pp. 482–8.

GATES, H. L. and STEELE, C. 2009. A conversation with Claude Steele: Stereotype threat and black achievement. *Du Bois Review.* **6**(2), pp. 251-271.

GELMAN, S. A. 1999. Essentialism. *In*: R. A. WILSON and F. C. KEIL (eds.) *The MIT Encyclopedia of the Cognitive Sciences*. Cambridge, MA; London: The MIT Press, pp. 282-284.

GELMAN, S. A. 2003. *The Essential Child: Origins of Essentialism in Everyday Thought*. Oxford.

GELMAN, S. A. 2004. Psychological Essentialism in Children. *Trends in Cognitive Sciences*. **8**(9), pp. 404-409.

GELMAN, S. A. 2005. Essentialism in Everyday Thought. *Psychological Science Agenda* [online]. Available from: http://www.apa.org/science/about/psa/2005/05/gelman.aspx

GELMAN, S. A. and DIESENDRUCK, G. 2001. Essentialism in Brazilian Children's Extensions of Animal Names. *Developmental Psychology*. **37**, pp. 49–60.

GELMAN, S. A. and MARKMAN, E. M. 1986. Categories and Induction in Young Children. *Cognition.* **23**, pp. 183-209.

GELMAN, S. A. and WELLMAN, H. M. 1991. Insides and essences: Early understandings of the nonobvious. *Cognition. **38**, pp.* 213-244.

GELMAN, S. A., WARE, E. A. and KLIENBERG, F. 2010. Effects of Generic Language on Category Content and Structure. *Cognitive Psychology.* **61**(3), pp. 273-301.

GENDLER, T. S. 2008. Alief in action (and reaction). *Mind and Language.* **23**(5), pp. 552-585.

GENDLER, T. S. 2011. On the Epistemic Costs of Implicit Bias. *Philosophical Studies.* **156**(1), pp. 33-63.

GIBB, S., FERGUSSON, D. and HORWOOD, L. 2010. Burden of psychiatric disorder in young adulthood and life outcomes at age 30. *British Journal of Psychiatry.* 197, pp. 122–7.

GIL-WHITE, F. J. 2001. Are Ethnic Groups Biological "Species" to the Human Brain?: Essentialism in Our Cognition of Some Social Categories. *Current Anthropology.* 42, pp. 515–54.

GLASER, J. and KNOWLES, E. 2008. Implicit motivation to control prejudice. *Journal of Experimental Social Psychology*. **44**, pp. 164–172.

GOFFMAN, E. 1973. *Stigma: Notes on the management of spoiled identity*. Englewood Cliffs, NJ: Prentice Hall.

GOGUEN, S. 2016. Stereotype Threat, Epistemic Injustice and Rationality. *In:* M. BROWNSTEIN and J. SAUL (eds.) *Implicit Bias and Philosophy: Volume 1.* Oxford: Oxford University Press, pp. 216-237.

GREENBERG, G. 2013. *The Book of Woe: The DSM and the Unmaking of Psychiatry.* New York: Penguin Group.

GREENWALD, A., MCGHEE, D., and SCHWARTZ, D. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*. 74, pp.1464–1480.

GREENWALD, A., NOSEK, B. and BANAJI, M. 2003. Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of personality and social psychology*. **85**(2), pp. 197–216.

GREENWALD, A., POEHLMAN, A., UHLMANN, E. L. and BANAJI, M. R. 2009. Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity. *Journal of Personality and Social Psychology.* **97**(1), pp. 17-41.

GRICE, H.P. 2010. Logic and Conversation. *In*: A.P. MARTINICH (ed.) *The Philosophy of Language*. Fifth Edition. Oxford: Oxford University Press, pp. 171-181.

GRIFFIN, A. 2016. Scientists find biological cause for schizophrenia in study that could open way to curing disorder. *The Independent* [online]. Available from: http://www.independent.co.uk/life-style/gadgets-and-

tech/news/scientists-find-biological-cause-for-schizophrenia-in-study-that-could-open-way-to-curing-disorder-a6838716.html

GUERTS, B.1985. Generics. *Journal of Semantics.* **4**(3), pp. 247-255.

HACKING, I. 1991. A Tradition of Natural Kinds. *Philosophical Studies*. **61**(1-2), pp. 109-126.

HACKING, I. 1998. *Mad Travellers.* Cambridge, MA: Harvard University Press.

HACKING, I. 1999. *The Social Construction of What*? Cambridge, Massachusetts: Harvard University Press.

HAMILTON, D. L. and SHERMAN, J. W. 1994. Stereotypes. *In:* R. S. WYER, JR and T. K. SRULL (eds.) *Handbook of Social Cognition (2nd ed, vol 2).* Hillsdale, NJ: Lawrence Erlbaum, pp. 1-68.

HANEY-LÓPEZ, I. 2000. Institutional racism: Judicial Conduct and a new theory of racial discrimination. *The Yale Law Journal.* **109**(8), pp. 1717-188.

HASLAM, N., ROTHSCHILD. L. and ERNST, D. 2000. Essentialist Beliefs about Social Categories. *British Journal of Social Psychology.* **39**(1), pp. 113-127.

HASLANGER, S. 2011. Ideology, Generics, and Common Ground. *In:* C. WITT (ed.) *Feminist Metaphysics: Explorations in the Ontology of Sex, Gender, and the Self*. Dordrecht: Springer, pp. 179–209.

HAWLEY, K. 2011. Knowing How and Epistemic Injustice. *In:* In J. BENGSON and M. A. MOFFETT (eds.) *Knowing How: Essays on Knowledge, Mind, and Action*. Oxford: Oxford University Press, pp. 283-99.

HENRY, J. D. et al. 2010. Threat perception in schizophrenia-spectrum disorders. *Journal of International Neuropsychology and Sociology*. 16, pp. 805–812.

HEMPEL, C. 1965. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. The Free Press.

HEMPEL, C. 1994. Fundamentals of Taxonomy. *In*: J.Z.SADLER, O.P.WIGGINS and M.A.SCHWARTZ (eds.). *Philosophical Perspectives on Psychiatric Diagnostic Classification*. Baltimore and London: The John Hopkins University Press, pp.315-331.

HILTON, J. L. and VON HIPPEL, W. 1996. Stereotypes. *Annual Review of Psychology.* **47**(1), pp.237-271.

HIPES, C., LUCAS, J., PHELAN, J. C. and WHITE, R. C. 2016. The stigma of mental illness in the labour market. *Social Science Research.* **56**, pp. 16-25.

HIRSCHFELD, L. A. 1996. *Race in the Making*: *Cognition, Culture, and the Child's Construction of Human Kinds.* Cambridge, MA: MIT Press.

HIRSCHFELD, L. A. and GELMAN, S. A. (eds.) 1994. *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge: Cambridge University Press.

HO, A. 2014. Epistemic Injustice. *In:* B. JENNINGS (ed.) *Encyclopedia of Bioethics, Edition: 4th*. Macmillan.

HOLROYD, J. and SWEETMAN, J. 2016. The Heterogeneity of Implicit Bias. *In:* M. BROWNSTEIN and J. SAUL (eds.) *Implicit Bias and Philosophy, Volume 1.* Oxford: Oxford University Press, pp. 80- 103.

HOOD, B.M. and BLOOM, P. 2008. Children Prefer Certain Individuals Over Perfect Duplicates. *Cognition.* **106**, pp.455-462.

HUEBNER, B. 2009. Trouble with Stereotypes for Spinozan Minds. *Philosophy of the Social Sciences.* **39**, pp. 63-92.

HUEBNER, B. 2016. Implicit Bias, Reinforcement Learning, and Scaffolded Moral Cognition. *In:* M. BROWNSTEIN and J. SUAL (eds.) *Implicit Bias and Philosophy, Volume 1.* Oxford: Oxford University Press, pp.47-79.

HUIBERS, M.J.H. and WESSELY, S. 2006. The Act of Diagnosis: Pros and Cons of Labelling Chronic Fatigue Syndrome. *Psychological Medicine*, pp.1-8. .

HYLER, S. E., GABBARD, G. O. and SCHNEIDER, I. 1991. Homicidal maniacs and narcissistic parasites: stigmatization of mentally ill persons in the movies. *Hospital Community Psychiatry*. **42**, pp. 1044–1048.

INSPECTOR, Y., KUTZ, I. and DAVID, D. 2004. Another person's heart: magical and rational thinking in the psychological adaptation to heart transplantation. *Israel Journal of Psychiatry and Related Sciences*. **41**, pp. 161-173.

JAMES, M. 2017. Race. *In:* E. N. ZALTA (ed.) *The Stanford Encyclopaedia of Philosophy* [online]. Available from: https://plato.stanford.edu/entries/race/

JASWAL, V. K. and MARKMAN, E. M. 2002. Children's Acceptance and Use of Unexpected Category Labels to Draw Non-Obvious Inferences. *In*: W. GRAY and C. SCHUNN (eds.) *Proceedings of the twenty-fourth annual conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum, pp. 500-505.

JAYARATNE, T. 2001. National Sample of Adults' Beliefs about Generic Bases to Race and Gender. Unpublished raw data.

JOHNSTONE, L. 2000. *Users and Abusers of Psychiatry: A Critical Look at Psychiatric Practice*. London: Routledge.

JORM, A. *2012.* Mental health literacy: empowering the community to take action for better mental health. *American Psychologist. 67, pp.* 231-*43.*

JUSSIM, L., MCCAULEY, C. and LEE, Y. T. 1995. Why study stereotype accuracy and inaccuracy? *In:* Y. T. LEE, L. JUSSIM, & C. MCCAULEY (eds.) *Stereotype accuracy: Toward an appreciation of group differences*. Washington, DC: American Psychological Association, pp. 1-23.

KEIL, F. 1989. *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: MIT Press.

KENDLER, K.S. and ZACHAR, P. 2008. The Incredible Insecurity of Psychiatric Nosology. *In*: K, S. KENDLER and J. PARNAS (eds.). *Philosophical Issues in Psychiatry*. The John Hopkins University Press: Baltimore, pp.370-383.

KESSLER, R. C., BERGLUND, P. A., BRUCE, M. L., KOCH, R., LASKA, E. M., LEAF, P. J., et al. 2001. The prevalence and correlates of untreated serious mental illness. *Health Services Research.* 36, pp. 987–1007.

KIDD, I. J. and CAREL, H. 2016. Epistemic Injustice and Illness. *Journal of Applied Philosophy*, pp. 1-19.

KINDY, K. and ELLIOTT, K. 2015. 990 people were shot and killed by police this year: here's what we learned. *Washington Post* [online]. Available from: https://www.washingtonpost.com/graphics/national/police-shootings-year-end/

KINNEAR, S.H., LINK, B.G., BALLAN, M.S. et al. 2016. Understanding the Experience of Stigma for Parents of Children with Autism Spectrum Disorder and the Role Stigma Plays in Families' Lives. *Journal of Autism and Developmental Disorders*. **46**(3), pp.942-953.

KLERMAN, G.L. 1980. Affective Disorders. *In*: H.I. KAPLAN, A.FREEDMAN and B.J. SADDOCK (eds.) *Comprehensive Textbook of Psychiatry (3rd ed., Vol 2)*. Baltimore: Williams & Wilkins, pp.1305-1331.

KOPERA, M. et al. 2015. Evaluating Explicit and Implicit Stigma of Mental Illness in Mental Health Professionals and Medical Students. *Community Mental Health Journal*. **51**(5), pp. 628-34f

KRUEGER, J. 1996. Probabilistic National Stereotypes. *European Journal of Social Psychology.* **26**(6), pp. 961-980.

KRUPA, T. M., KIRSH, B., COCKBURN, L. and GEWURTZ, R. 2009. Understanding the stigma of mental illness in employment. *Work*. **33**(4), pp. 413-425.

KUHN, T. 2000. *The Road Since Structure*. Chicago: University of Chicago Press.

KUNDA, Z. and OLESON, K. C. 1995. Maintaining stereotypes in the face of disconfirmation: Constructing grounds for subtyping deviants. *Journal of Personality and Social Psychology*. **68**, pp. 565–79.

KURZBAN, R. and LEARY, M. R. 2001. Evolutionary Origins of Stigmatization: The Functions of Social Exclusion. *Psychological Bulletin.* **127**(2), pp. 187-208.

KUTCHINS, H. and KIRK, S.A. 1997. *Making Us Crazy: DSM- the Psychiatric Bible and the Creation of Mental Disorders*. London: Constable.

LAING, R.D. 1960. *The Divided Self*. London: Tavistock Press.

LAKEMAN, R. 2010. Epistemic injustice and the mental health service user. *International Journal of Mental Health Nursing.* **191**, pp. 151-153.

LANE, K., KANG, J. and BANAJI, M. 2007. Implicit Social Cognition and Law. *Annual Review of Law and Social Science*. **3**, pp. 427–451.

LAWLER, J. M.1973. *Studies in English generics*. University of Michigan Papers in Linguistics.

LESLIE, S. 2007. Generics and the Structure of the Mind. *Philosophical Perspectives*, pp. 375–405.

LESLIE, S. 2008. Generics: Cognition and Acquisition. *Philosophical Review*. **117**(1), pp.1–49.

LESLIE, S. J. 2013. Essence and Natural Kinds: When Science Meets Preschooler Intuition. *Oxford Studies in Epistemology*. **4**, pp.108-165.

LESLIE, S. J. 2014. Carving Up the Social World with Generics. *Oxford Studies in Experimental Philosophy.* **1**, pp.208-232.

LESLIE, S. J. Forthcoming. The original sin of cognition: Fear, prejudice, and generalization". *The Journal of Philosophy*.

LESLIE, S. J. and LERNER, A. 2016. Generics and Experimental Philosophy. *In:* W. BUCKWALTER and J. SYTSMA (eds.) *A Companion to Experimental Philosophy*. Oxford: Wiley-Blackwell, pp. 404-416. Print.

LEWIS, D. 1979. Scorekeeping in a Language Game. *Journal of Philosophical Logic.* **8**, pp. 339- 359.

LINK, B.G. 1987. Understanding Labelling Effects in the Area of Mental Disorders: An Assessment of the Effects of Expectations of Rejection. *American Sociological Review*. **52**, pp.96–112.

LINK, B. G. and PHELAN, J. C. 2001. Conceptualizing Stigma. *Annual Review of Sociology.* **77**, pp. 363-385.

LINK, B. G. and PHELAN, J. C. 2006. Stigma and its public health implications. *Lancet.* **367**, pp. 528-529.

LOZANO, R. et al. 2012. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet.* **380** (9859), pp. 2095-2128.

MACHERY, E. 2016. De-Freuding Implicit Attitudes. *In:* M. BROWNSTEIN and J. Saul (eds.) *Implicit Bias and Philosophy, Volume 1*. Oxford: Oxford University Press, pp. 104-129.

MADVA, A. 2012. The Hidden Mechanisms of Prejudice: Implicit Bias and Interpersonal Fluency. *PhD Dissertation Columbia University.*

MADVA, A. 2016. Virtue, Social Knowledge, and Implicit Bias. *In:* M. BROWNSTEIN and J. Saul (eds.) *Implicit Bias and Philosophy, Volume 1.* Oxford: Oxford University Press, pp. 191-215,

MAHALINGAM, R. 2003. Essentialism, Culture, and Power: Representations of Social Class. *Journal of Social Issues.* **59**, pp. 733–49.

MAKANJUOLA, V., ESAN, Y., OLADEJI, B. et al. 2016. Explanatory model of psychosis: impact on perception of self-stigma by patients in three Sub-Saharan African cities. *Social Psychiatry and Psychiatric Epidemiology* [online], pp.1-10. Available from: http://link.springer.com/article/10.1007/s00127-016-1274-8

MALLON, R. 2006. Race: Normative, Not Metaphysical or Semantic. *Ethics*. **116**(3), pp. 525–551.

MALLON, R. 2007. A Field Guide to Social Construction. *Philosophy Compass*. **2**(1), pp. 93–108.

MALLON, R. 2016. Stereotype Threat and Persons. *In:* M. BROWNSTEIN and J. Saul (eds.) *Implicit Bias and Philosophy, Volume 1*. Oxford: Oxford University Press, pp. 130-154.

MARGOLIS, J. 1994. Taxonomic Puzzles. *In*: J.Z.SADLER, O.P.WIGGINS and M.A.SCHWARTZ (eds.). *Philosophical Perspectives on Psychiatric Diagnostic Classification*. Baltimore and London: The John Hopkins University Press, pp.104-128.

MARKOWITZ, F. E. 2001. Modelling processes in recovery from mental illness: Relationships between symptoms, life satisfaction, and self-concept. Journal of Health and Social Behaviour. **42**, pp. 64-79.

MARSH, G. 2011. Trust, Testimony, and Prejudice in the Credibility Economy. *Hypatia*. **26**(2), pp. 280-293.

MARSH, J. K. and SHANKS, L. L. 2014. Thinking you can catch mental illness: how beliefs about membership attainment and category structure influence interactions with mental health category members. *Memory & Cognition.* **42**(7), pp. 1011- 1025.

MCHUGH, P.R. and SLAVNEY, P.R. *The Perspectives of Psychiatry: Second Edition.* Baltimore and London: The John Hopkins University Press.

MCCKINNON, R. 2014. Stereotype Threat and Attributional Ambiguity for Trans Women. *Hypatia.* **29**(1), pp. 857-872.

MIND. 2015. *Advocacy in Mental Health* [online]. Available from: https://www.mind.org.uk/information-support/guides-to-support-and-services/advocacy/#.WXUFR-mQzIU

MIND. 2016. *Soap characters and news readers can save lives: people with mental health problems seek help following media coverage* [online]. Available from: https://www.mind.org.uk/news-campaigns/news/soap-characters-and-news-readers-can-save-lives-people-with-mental-health-problems-seek-help-following-media-coverage/#.WXT4FemQzIX

MENTAL HEALTH FOUNDATION. 2016. *Fundamental Facts about Mental Health.* [Online]. London: The Mental Health Foundation. [Accessed 07/2017]. Available from: http://www.mentalhealth.org.uk/content/assets/PDF/publications/fundamental-facts-15.pdf?view=Standard

MEYER, M., GELMAN, S.A., LESLIE, S.J. and STILWELL, S.M. 2013. Essentialist Beliefs about Bodily Transplants in the United States and India. *Cognitive Science.*

MULAY, A. L. 2016. Crisis Intervention Training and Implicit Stigma Toward Mental Illness: Reducing Bias Among Criminal Justice Personnel. *International Journal of Forensic Mental Health.* **15**(4), pp. 369-381.

MÜLLER, U. and PERRY, B. 2001, Adopted Persons' Search for and Contact with Their Birth Parents I: Who Searches and Why? *Adoption Quarterly*.**4,** pp.5-37.

NEMEROFF, C. and ROZIN, P. 1994. The Contagion Concept in Adult Thinking in the United States: Transmission of Germs and Interpersonal Influence. *Ethos*. **22**, pp.158-186.

NEWMAN, G. E., DIESENDRUCK, G. and BLOOM, P. 2011. Celebrity Contagion and the Value of Objects. *Journal of Consumer Research.* **38**, pp. 215-228.

NOSEK, B. A. and BANAJI, M. R. 2001. The Go/ No-go Association Task. *Social Cognition.* **19**(6), pp. 625-664.

NOSEK, B. A., BANAJI, M. R. and GREENWALD, A. G. 2002. Harvesting intergroup implicit attitudes and beliefs from a demonstration Web site. *Group Dynamics.* **6**(1), pp. 101-115.

NOSEK, B. A., GREENWALD, A. G. and BANAJI, M. 2005. Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin*. **31**(2), pp.166–180.

NOSEK, B. A., GREENWALD, A. G. and BANAJI, M. R. 2007a. The Implicit Association Test at Age 7: A Methodological and Conceptual Review. *In:* J. A. BARGH (ed.) *Automatic Processes in Social Thinking and Behaviour*. Philadelphia: Psychology Press.

NOSEK, B. et al. 2007. Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*. **18**, pp. 36-88.

NOSEK, B., HAWKINS, C. and FRAZIER, R. 2011. Implicit social cognition: from measures to mechanisms. *Trends in cognitive sciences*. **15**(4), pp. 152–159.

OSWALD, F. L. et al. 2013. Predicting Ethnic and Racial Discrimination: A Meta-Analysis of IAT Criterion Studies. *Journal of Personality and Social Psychology.* **105**(2), pp. 175-192.

PACHANKIS, J. E. 2007. The psychological implications of concealing a stigma: A cognitive-affective-behavioural model. *Psychological Bulletin*.**133**, pp 328–345.

PARNAS, J. and SASS. L.A. 2008. Varieties of Phenomenology: On Description, Understanding, and Explanation in Psychiatry. *In*: K, S. KENDLER and J.PARNAS (eds.). *Philosophical Issues in Psychiatry*. The John Hopkins University Press: Baltimore, pp. 243-276.

PAYNE, K., CHENG, C. M., GOVORUN, O. and STEWART, B. D. 1999. An Inkblot for Attitudes: Affect Misattribution as Implicit Measurement. *Journal of Personality and Social Psychology*. **89**(3), pp. 277-293.

PENN, D. L., GUYNAN, K., DAILY, T. SPAULDING, W. D., GARBIN, C. P. and SULLIVAN, M. 1994. Dispelling the stigma of schizophrenia: What sort of information is best? *Schizophrenia Bulletin*. **20**(3), pp. 567-578.

PENN, D.L., CORRIGAN, P.W., BENTALL, R.P., RACENSTEIN, J.M. and NEWMAN, L. 1997. Social cognition in schizophrenia. *Psychological Bulletin*. **121**, pp. 114–132.

PERIS, T. TEACHMAN, B., and NOSEK, B. 2008. Implicit and explicit stigma of mental illness: Links to clinical care. *Journal of Nervous and Mental Disease*. **196**, pp. 752–760.

PFEIFFER, E.J. & MAITHYA, H.M.K. 2016. Bewitching sex workers, blaming wives: HIV/AIDS, stigma, and the gender politics of panic in western Kenya. *Global Public Health* [online], pp.1-16. Available from: https://www.researchgate.net/profile/Elizabeth_Pfeiffer/publication/307593609_Bewitching_Sex_Workers_Blaming_Wives_HIVAIDS_Stigma_and_the_Gender_Politics_of_Panic_in_Western_Kenya/links/57ce314708ae582e069240f2.pdf

PHELAN, J. C., LINK, B. G., STUEVE, A., and PESCOSOLIDO, B. 2000. Public conceptions of mental illness in 1950 and 1996: What is mental illness and is it to be feared? *Journal of Health and Social Behaviour*. **41**(2), pp. 188–207.

PHELAN, J. C. 2006. Geneticization of Deviant Behaviour and Consequences for Stigma: The Case of Mental Illness. *Journal of Health and Social Behaviour*. **46**(4), pp. 307-322.

PLATO. 1925. *In*: H. N. FOWLER (ed.) *Plato in Twelve Volumes, Vol. 9*. Cambridge, MA, Harvard University Press; London, William Heinemann Ltd.

PRENTICE, D. A. and MILLER, D. T. 2007. Psychological Essentialism of Human Categories. *Current Directions in Psychological Science.* **16**(4), pp. 202-206.

QUINN, D. M, KAHNG, S. K. and CROCKER, J. 2004. Discreditable: Stigma effects of revealing a mental illness history on test performance. *Personality and Social Psychology Bulletin*. **30**, pp. 803–815.

RHODES, M. and GELMAN, S.A. 2009. Five-Year-Olds' Beliefs About the Discreetness of Category Boundaries for Animals and Artifacts. *Psychonomic Bulletin and Review.* **16**, pp. 920–924.

RHODES, M. 2012. How Generic Language Leads Children to Develop Social Stereotypes. Available from: http://www.huffingtonpost.com/marjorie-rhodes-phd/generic-language-social-stereotypes_b_1753667.html

RHODES, M. LESLIE, S. L. and TWOREK, C. M. 2012. Cultural transmission of social essentialism. *Proceedings of the National Society of Sciences of the United States of America.* **109**(34), pp. 13526-13531.

ROBERTSON, T. and ATKINS, P. 2016. Essential vs. Accidental Properties. *In*: E.N. ZALTA (ed.c) *The Stanford Encyclopedia of Philosophy* [online]. [Accessed 23/07/2017]. Available from: <http://plato.stanford.edu/archives/win2013/entries/essential-accidental/>.

ROY, P. 2002. *There Ain't no Black in the Union Jack': The Cultural Politics of Race and Nation*. London: Routeledge.

RÜSCH, N., ANGERMEYER, M. and CORRIGAN, P. W. 2005. Mental illness stigma: Concepts, consequences, and initiatives to reduce stigma. *European Psychiatry.* **20**, pp. 529-539.

RÜSCH, N. et. al 2006a. Self-stigma in women with borderline personality disorder and women with social phobia. *Journal of Nervous and Mental Disease*. **194**, pp. 766–73.

RÜSCH, N. et. al. 2006b. Self-Stigma, Empowerment, and Perceived Legitimacy of Discrimination among Women with Mental Illness. *Psychiatric Services*. **57**, pp.399-402.

RÜSCH, N., ZLATI, A., BLACK, G. and THORNICROFT, G. 2014. Does the stigma of mental illness contribute to suicidality? *The British Journal of Psychiatry.* **205**(4), pp.257-259.

RÜSCH, N. and XU, Z. 2017. Strategies to Reduce Mental Illness Stigma. *In:* W. GAEBEL, W. RÖSSLER and N. SARTORIUS, N. (eds.) *The Stigma of Mental Illness- End of the Story?* Switzerland: Springer International Publishing, pp. 451-468.

SANATI, A. and KYRATSOUS, M. 2015. Epistemic injustice in assessment of delusions. *Journal of Evaluation in Clinical Practice*. **21**, pp. 479-485.

SARTORIUS, N. 2005. *Reducing the stigma of mental illness: A report from the Global Programme of the World Psychiatric Association*. Cambridge: Cambridge University Press.

SARTORIUS, N. and CALLARD, F. 2012. *Mental illness, discrimination, and the law fighting for social justice. Chichester*: Wiley-Blackwell.

SAUL, J. 2013. Implicit Bias, Stereotype Threat, and Women in Philosophy. Available from: https://www.sheffield.ac.uk/polopoly_fs/1.394073!/file/saul_implicit.pdf

SCOTTISH INDEPENDENT ADVOCACY ALLIANCE. 2017. Types of Advocacy [online]. Available from: https://www.siaa.org.uk/us/independent-advocacy/need-advocate/

SCHWABE, L. and WOLF, O. 2013. Stress and multiple memory systems: From 'thinking' to 'doing'. Trends in Cognitive Sciences. 17(2), pp. 60-69.

SCHWITZGEBEL, E. 2010. Acting Contrary to our Professed Beliefs or the Gulf Between Occurrent Judgement and Dispositional Belief. Pacific Philosophical Quarterly, pp. 531-553.

SHAPIRO, J. and AARONSON, J. 2013. Stereotype Threat. *In:* STANGOR and CRANDALL, pp. 95-117.

SHEEHAN, L., NIEWEGLOWSKI, K. and CORRIGAN, P. W. 2017. Structures and Types of Stigma. *In:* W. GAEBEL, W. RÖSSLER and N. SARTORIUS, N. (eds.) *The Stigma of Mental Illness- End of the Story?* Switzerland: Springer International Publishing, pp. 43-66.

SINGAL, J. 2017. Psychology's Favourite Tool for Measuring Racism Isn't Up to the Job. *New York Magazine.* Available from: http://nymag.com/scienceofus/2017/01/psychologys-racism-measuring-tool-isnt-up-to-the-job.html

SIREY, J. A., BRUCE, M. L., ALEXOPOULOS, G. S., PERLIC, D. A., RAUE, P., FRIEDMAN, S. J., and MEYERS, B. S. 2001. Perceived stigma as a predictor of treatment discontinuation in young and older outpatients with depression. *American Journal of Psychiatry.* 158, 479–481.

SOUSA, P., ATRAN, S. and MEDIN, D. L. 2002. Essentialism and Folkbiology: Evidence from Brazil. *Journal of Cognition and Culture*. **2**, pp. 195–223.

STAFFORD, M. C. and SCOTT, R. R. 1986. Stigma deviance and social control: some conceptual issues. *In:* S. C. AINLAY, G. BECKER and L. M. COLEMAN (eds.) *The Dilemma of Difference.* New York: Plenum, pp. 77-91.

STEELE, C. M. and ARONSON, J. 1995. Stereotype vulnerability and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*. 69, pp. 797-811

STIER, A. and HINSHAW, S, P. 2007. Explicit and implicit stigma against individuals with mental illness. *Australian Psychologist*. **42**(2), pp. 106-117.

STUART, H. 2004. Stigma and work. *Healthcare Papers.* **5** (2), pp. 100-111.

STUART, H. 2006. Media portrayal of mental illness and its treatments: what effect does it have on people with mental illness? *CNS Drugs.* **20**(2), pp. 99-106.

STULL, L. G. et al. 2013. Implicit and Explicit Stigma of Mental Illness: Attitudes in an Evidence-Based Practice. *Journal of Nervous and Mental Disease.* **201**(12), pp.1072-1079.

SUMMERS, L. H. 2005. *Remarks at NBER Conference on Diversifying the Science & Engineering Workforce*. Cambridge, MA, January 14, 2005. [Accessed 07/05/2005]. Available from:http://www.president.harvard.edu/speeches/2005/nber.html

SYLVIA, C. and NOVAK, W. 1997. *A Change of Heart*. Boston: Little, Brown.

SZASZ, T. 1961. *The Myth of Mental Illness*. London: Harper & Row Publishers.

TAYLOR, M. G. 1996. The Development of Children's Beliefs about Social and Biological Aspects of Gender Differences. *Child Development*. **67**, pp. 1555–71.

TAYLOR, M. G., RHODES, M. and GELMAN, S. A. 2009. Boys Will Be Boys; Cows will be Cows: Children's Essentialist Reasoning about Gender Categories and Animal Species. *Child Development*. **79**, pp. 1270–87.

TIME TO CHANGE. 2017. Available from: https://www.time-to-change.org.uk/mental-health-statistics-facts

THE MENTAL HEALTH FOUNDATION. 2015. *Fundamental Facts about mental health 2015* [online]. Available from: https://www.mentalhealth.org.uk/publications/fundamental-facts-about-mental-health-2015.

THE MENTAL HEALTH FOUNDATION. 2016. *Fundamental Facts about mental health 2016* [online]. Available from: https://www.mentalhealth.org.uk/publications/fundamental-facts-about-mental-health-2016.

THE NHS INFORMATION CENTRE FOR HEALTH AND SOCIAL CARE. 2009. *Adult Psychiatric Morbidity in England: Results of a Household Survey.* London: NHS.

THORNTON, J. A. and WAHL, O. F. 1996. Impact of a newspaper article on attitudes toward mental illness. *Journal of Community Psychology.* **24**(1), pp. 17-25.

TIME TO CHANGE. 2014. *Attitudes to Mental Illness 2013 Research Report.* [Online]. London: Time to Change. [Accessed 30/11/2015]. Available from: http://www.time-to-change.org.uk/sites/default/files/121168_Attitudes_to_mental_illness_2013_report.pdf

TRIGGLE, N. 2017. Winter pressure 'busts NHS budget'. *BBC News website.* Available from: http://www.bbc.co.uk/news/health-39029265

VERHAEGHE, M., BRACKE, P., & BRUNYNOOGHE, K. 2008. Stigmatization and self-esteem of persons in recovery from mental illness: The role of peer support. *International Journal of Social Psychiatry*. 54, pp. 206–218.

VOGEL, D. L., WADE, N.G. and HACKLER, J.H. 2007. Perceived public stigma and the willingness to seek counseling: The mediating roles of self-stigma and attitudes toward counselling. *Journal of Counselling Psychology.* **54**(1), pp.40-50.

VOS, T. 2015. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet.* **386** (9995), pp. 743-800.

WAHL, O. F. 1995. *Media Madness: Public Images of Mental Illness.* New Brunswick, NJ: Rutgers University Press.

WAHL, O. F. 1999. Mental health consumers' experience of stigma. *Schizophrenia Bulletin*. **25**, pp.467–78.

WALKER, S. 1999. Culture, Domain-Specificity, and Conceptual Change: Natural kind and Artifact Concepts. *British Journal of Developmental Psychology*. 17, pp. 203–19.

WALLACE, E. R. 1994. Psychiatry and its nosology: A historico-philosophical overview. *In:* J. Z. SADLER, O. P. WIGGINS and M. A. SCHWARTZ (eds.) *Philosophical Perspectives on Psychiatric Diagnostic Classification.* Baltimore: John Hopkins University Press, pp. 16-88.

WALTON, G. and COHEN, G. 2007. A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology.* **89**(1), pp. 82-96.

WASHINGTON, N. and KELLY, D. 2016. Who's Responsible for This? Moral Responsibility, Externalism, and Knowledge about Implicit Bias. *In:* M. BROWNSTEIN and J. SAUL (eds.) *Implicit Bias and Philosophy, Volume 2.* Oxford: Oxford University Press, pp.11-36.

WAXMAN, S. R., MEDIN, D. L., and ROSS, N. 2007. Folkbiological Reasoning from a Cross- Cultural Developmental Perspective: Early Essentialist Notions Are Shaped by Cultural Beliefs. *Developmental Psychology*. **43**(2), pp. 294–308.

WHITEFORD, H. A., FERRARI, A. J., DEGENHARDT, L., FEIGIN, V., and VOS, T. 2015. The Global Burden of Mental, Neurological and Substance Use Disorders: An Analysis from the Global Burden of Disease Study 2010. *PLoS ONE*. **10**(2). Available from: http://doi.org/10.1371/journal.pone.0116820.

WRIGHT, I. C., RABE-HESKETH, S., WOODRUFF, P. W., DAVID, A. S., MURRAY, R. M. and BULLMORE, R. T. 2000. Meta-analysis of regional brain volumes in schizophrenia. *The American Journal of Psychiatry.* **157**(1), pp. 16-25.

VOGEL, D. L., WADE, N. G, and HACKLER, A. H. 2007. Perceived public stigma and the willingness to seek counselling. *Journal of Counselling Psychology*. **54**, pp. 40–50.

YZERBYT, V. Y., JUDD, C. M. & CORNEILLE, O. 2004. *The Psychology of Group Perception: Perceived Variability, Entitativity, and Essentialism.* London: Psychology Press.

ZACHAR, P. 2000. Psychiatric Disorders are not Natural Kinds. *Philosophy, Psychiatry, & Psychology.* **7**, pp.167-182.

ZACHAR, P. 2008. Real Kinds but no True Taxonomy. *In*: K, S. KENDLER and J. PARNAS (eds.). *Philosophical Issues in Psychiatry*. The John Hopkins University Press: Baltimore, pp.327-354.

ZACHAR, P. 2014. Beyond Natural Kinds: Towards a “Relevant” “Scientific” Taxonomy in Psychiatry. *In:* H. KINCAID and J.A. SULLIVAN (eds). *Classifying Psychopathology: Mental Kinds and Natural Kind*s. London, Cambridge, MA: The MIT Press, pp.75-104.

ZACK, N. 2002. 2002, *Philosophy of Science and Race*. New York: Routledge.