

TANGLED INFERENCES: AN INVESTIGATION OF INFERENCE WEBS IN
METAPHYSICS, EPISTEMOLOGY, AND THE PHILOSOPHY OF SCIENCE

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Frances Heather Fairbairn

August 2019

© 2019 Frances Heather Fairbairn

TANGLED INFERENCES: AN INVESTIGATION OF INFERENCE WEBS IN
METAPHYSICS, EPISTEMOLOGY, AND THE PHILOSOPHY OF SCIENCE

Frances Heather Fairbairn, Ph. D.

Cornell University 2019

This dissertation is a collection of three papers. The second and third papers are primarily about inference webs. An inference web is a structured collection of claims and principles in which inferential relationships (including association and common use) dictate the covert meaning of the individual parts of the structure. Here, I introduce the notion of inference webs and argue that, while they are an essential part of our epistemic practices, they can be problematic and, when they are, they obscure our theorizing and wrong those who operate under them.

In the third paper I argue that the mental illness literature operates under a problematic inference web whereby legitimate principles are imbued with empiricist assumptions that cause them to go awry. For example, they assume that the social and the scientific are fundamentally distinct. Most philosophers of science will deny this principle if asked directly about it, but I will show that the same principle is operative in many of the inferences that are routinely made in the literature.

In the second paper, I argue that inference webs play a role in epistemic injustice. For example, women who have heart attacks are frequently misdiagnosed comparative to their male peers. I argue that this is, in part, due to the fact that someone's being a woman blocks the inference to 'this person is having an heart attack' even when there is both i) a good understanding of what heart attacks are, and ii) good evidence for a heart

attack. In these cases, the individual notions are well understood, but inference webs prevent us from correctly inferring one from the other (alternatively, encourage us to incorrectly infer one from the other).

The first paper represents a part of my philosophical work that is independent of the above project. I offer a solution to the problem of advanced modalizing which is supposed to show that genuine modal realism is unable to accommodate claims like 'possibly, there are many possible worlds.' I argue that the ontology of modal realism rules out those claim as category mistakes.

BIOGRAPHICAL SKETCH

Fran grew up in Manchester, England, the youngest of four children. While in high school, she displayed an early interest in Philosophy and Fine Arts. After achieving A-Levels in each of Philosophy, Fine Art and Textiles, Fran enrolled in the Philosophy BA program at the University of Leeds. During her time in Leeds, Fran became interested in metaphysics and completed a dissertation under the supervision of Dr. John Divers titled 'Does Possible World Realism Offer an Adequate Reduction of Modality?'. She graduated with First Class Honors in 2011. Fran continued her interest in metaphysics through a Masters' program at the University of Leeds. During her Masters' degree, Fran chaired the British Undergraduate Philosophy Society, and completed a dissertation titled 'Counterfactual Epistemology without Conceivability Arguments' under the supervision of Dr. Scott Shalkowski. She graduated with an MA (Merit) in 2012.

Following her Masters' degree, Fran earned a place in the Philosophy PhD program at Cornell University. At Cornell, she advanced her studies in metaphysics, but also developed research interests in feminism, philosophy of science, and philosophy of education. During her PhD studies, Fran has presented her work at conferences across the United States, in Europe and in South Africa. Notably, she presented at the Southern Journal of Philosophy workshop on 'The Epistemology of Justice' in 2019. She also achieved a fellowship at the Centre for Ethics and Education at The University of Wisconsin Madison for the year 2016 – 2017.

While studying for her Ph.D, Fran has also developed her passions for both teaching and fine arts. She designed and delivered courses for three years through the Cornell Prison Education Program, and served two years as an instructor for Cornell's Centre for Teaching Innovation. She also taught courses in quilting at local Ithaca store Quilter's Corner, which stocks Fran's quilting patterns designs under her company name 'Beneath the Brambles'. She continued to develop her skills in sewing and fine arts by taking classes at Schweinfurth Art Centre, and at Hog Island Arts and Birding. Fran won a scholarship for the program Quilting by the Lake in 2017, and was asked to provide the illustrations for a class called 'Raptor Rapture' also at Hog Island in 2018.

Fran completed her PhD thesis titled 'Tangled Inferences: An investigation of inference webs in metaphysics, epistemology and the philosophy of science' in 2019 under the supervision of Prof. Karen Bennett (chair), Prof. Dick Boyd, Prof. Ted Sider, Prof. Will Starr, and Prof. Elizabeth Barnes (The University of Virginia, Charlottesville). Following graduation, she has been offered a post-doctoral appointment funded by the journal, *The Philosophical Review*, at Cornell.

For all the many animals who have brightened my life.

ACKNOWLEDGMENTS

Writing a dissertation is hard. I was only able to complete this one with the help, encouragement, and feedback of a great many beautiful people. I will try to give recognition to those people here, though the result of my attempts will undoubtedly be imperfect. I hope that the amazing people in my life who have helped me to get through this will know how valued they are.

First of all, I wish to thank my dissertation committee for their support and influence on my work. I am indebted to my advisor, Karen Bennett, for teaching me to be clear and incisive, both in thought and writing. She has so frequently been exactly the advisor I needed, even when I really didn't know what that was. I also wish to give huge thanks to Dick Boyd, who has shaped my intellectual career in countless ways and who has always believed in my work and philosophical ability. Dick: I always left meetings with you feeling refreshed, inspired, and excited to do philosophy. Ted Sider was a constant support especially through my first years in Grad School when I was still discovering who I was as a philosopher. Elizabeth Barnes is a big part of why I ended up in Grad School in the first place; she was inspiring and encouraging and helped me to see that I really can do philosophy. Will Starr was a first point of contact in Grad School and encouraged me, early on, to develop my work to higher standards.

I am extraordinarily lucky to be surrounded by such an inspiring and supportive network of friends, colleagues, and instructors. Alicia Patterson has been a continuous source of strength and insight for me. I have, in her, both a loving friend and an inspiring

intellectual conversant. Robert Muckle has been a huge source of philosophical inspiration and encouragement and I have benefited from many hours of discussion with him. I'd like to thank Elizabeth Southgate for enduring countless blocks of time working with me in solidarity, and for her unfaltering support. I am indebted to Craig Settle, who has supported me in all the ways I could have possibly asked for and who has encouraged me to have faith in my own strength and abilities. I would also like to recognize Brandon Conley, Yuna Won, Avi Appel, Lucia Munguia, Augie Faller, Bianka Takaoka, Gaile Pohlhaus, Scott MacDonald, Rachana Kamtekar, Tad Brennan, Alexander Kocurek, and Nicole Dular. Special thanks to Michelle Kosch whose generous and thoughtful feedback has had an immeasurable effect on the quality of my work.

My work has benefitted from the feedback of countless participants across numerous conferences and events; too many to number. I'd like to thank, Velislava Mitova, Lubomira Radoilska, Ian Werkheiser, and Joel M. Reynolds. More generally, I'd like to thank participants in the Epistemic Injustice, Reasons, Agency research project, attendees at the Southern Journal of Philosophy's inaugural journal workshop; the Epistemology of Justice, along with attendees at Cornell's discussion club workshop and the graduate editing group. I'd also like to recognize all the many student's I've had the pleasure of doing philosophy with and who have influenced my ideas significantly.

I am lucky enough to enjoy a loving and supportive family. I must express my deep gratitude to my mother, Katharine Fairbairn. She believed in me unfalteringly, even when I didn't believe in myself, and offered me security and reassurance in a very difficult portion of my life. I would like to thank my father, Peter Fairbairn, for teaching me to be

curious and for encouraging me to continue my intellectual development. He is not around to see the end of my time in Grad School, but I hope that his memory lives on in my work. My brothers, Joe Fairbairn, Gus Fairbairn, and Bobb Fairbairn have been a source of inspiration and support throughout my life. For their unconditional love and for helping to keep me grounded, I'd like to thank the many non-human animals who have enriched my life beyond measure. They are; Alastair, Barbara, Cecilia, Douglas, Emily, Hector, Jeremy, Jessica, Jolene, Layla, Lucy, Magnolia, Mary, Penelope, Rhonda, Rita, Robyn, Rosie, Roxanne, Ruby, Ruth, Senna, Shakira, Valerie, and Willow.

TABLE OF CONTENTS

LIST OF FIGURES..... xii
LIST OF ABBREVIATIONS xiii
LIST OF SYMBOLS xv

INTRODUCTION 17

CHAPTER 1.....20
 Introduction 20
 1: The Problem of Advanced Modalizing..... 21
 2: The Role of the Translation Schema..... 26
 3: A Genuine Modal Realist’s Response to the Problem 32
 4: Advanced Propositions 38
 5: Category Mistakes 44
 Conclusion..... 47
 References..... 49

CHAPTER 2.....50
 Introduction 50
 1: The Cases..... 51
 2: Situating the cases 62
 3: Injustice in the Spaces Between Concepts: Inferential Injustice 74
 References..... 89

CHAPTER 3.....95
 Introduction 95
 1: The Literature 96
 2: Reduction..... 104
 3: Realism..... 137
 4: Objectivity 159
 5: Naturalism 174
 5: Conclusion..... 181
 References..... 184

LIST OF FIGURES

- Figure 1 *Lewis Puzzed*
Figure 2 *Carolina Chickadee vs. Black-Capped Chickadee*
Figure 3 *Rattus rattus*
Figure 4 *Pandion haliaetus*
Figure 5 *Ptilonorhynchus violaceus*
Figure 6 *Euplectes progne*
Figure 7 *Cavia porcellus*
Figure 8 *Biston betularia*
Figure 9 *Poecile atricapillus*
Figure 10 *Malurus cyaneus*

LIST OF ABBREVIATIONS

Views:

Modal Realism Genuine Modal Realism

Principles:

Reduction: the account must be able to give a *reductive* analysis of mental illness.

Realism: the account must have a *realist* ontology.

Objectivity: the account must demonstrate that the study of mental illness is or can be *objective*.

Naturalism: the account must employ and advocate a naturalist *empirical* methodology.

N1 the account must be able to give a *reductive* analysis of mental illness where ‘reduced’ is taken to imply *free from values*.

N2 the account must have a *realist* ontology where ‘real’ is taken to imply *human independence*.

N3 the account must demonstrate that the study of mental illness is or can be *objective* where ‘objective’ is taken to imply *free from human judgments*.

N4 the account must employ and advocate a naturalist *empirical* methodology where ‘empirical’ is taken to imply *freedom from presuppositions*.

(Division) The idea that the social contrasts with the scientific.

(Aim) The common social and political aim of those involved in the mental illness literature.

(Legitimate) The thought that psychiatry and the study/treatment of mental illness is only legitimate if naturalism about mental illness can be established.

Logical Rules:

MP modus ponendo ponens

Reductio reductio ad absurdum

(PI) possibility introduction

Sentences:

P^x $\exists y (Wy \ \& \ (\exists z (BSz \ \& \ Izy)))$; there is a y such that y is a world and there is a z such that z is a blue swan and z is in y

- (Swans) $\exists y (BSy)$; there is a y such that y is a blue swan ; there is a blue swan
- (\diamond Swans) $\diamond \exists y (BSy)$; it is possible that there is a y and y is a blue swan; it is possible that there is a blue swan
- (\exists Swans) $\exists x (Wx \ \& \ (\exists y (BSy))^x)$; there is an x such that x is a world and there is a y such that y is a blue swan and y is in x; there is a world in which there is a blue swan
- (wSwans) $(Ww \ \& \ \exists y (BSy \ \& \ Iyw))$
- (Plurality) $\exists y \exists z ((Wy \ \& \ Wz) \ \& \ y \neq z)$; there is a y and there is a z, y is a world and z is a world and y is not identical to z; there are at least two (concrete, spatiotemporally isolated...) possible worlds
- (\diamond Plurality) $\diamond (\exists y \exists z ((Wy \ \& \ Wz) \ \& \ y \neq z))$; it is possible that (there is a y and there is a z, y is a world and z is a world and y is not identical to z); Possibly, there are at least two possible worlds
- (\exists Plurality) $\exists x (Wx \ \& \ (\exists y \exists z ((Iyx \ \& \ Izx) \ \& \ (Wy \ \& \ Wz) \ \& \ y \neq z)))$; There is a w, w is a world, and there is a y and a z, y is a world and z is a world and y is in w and z is in w and y and z are non-identical; There is a world at which there are at least two possible worlds
- (wPlurality) $\exists z ((Ww \ \& \ Wz) \ \& \ y \neq w)$

LIST OF SYMBOLS

Predicates:

Wx	x is a world
BSx	x is a blue swan
Ixy	x is in y

Quantifiers:

$\forall x$	for all x
$\exists x$	there is an x

Modal Operators:

$\diamond P$	possibly, P
$\square P$	necessarily, P

Connectives:

$\sim P$	not P
$\&$	and
\supset	if, then
$x = y$	x is identical with y
$x \neq y$	it is not the case that $x = y$

Terms:

P, Q, Φ	propositional variables
$\beta, \alpha, x, y, z, w$	individual variables
@	the actual world

INTRODUCTION

This dissertation is primarily about inference webs. An inference web is a structured collection of claims and principles in which inferential relationships (of outright inference but also of association and common use) dictate the covert meaning of the individual parts of the structure (Think: Quinean web of belief but for inferential practices (Quine & Ullian, 1978)). Here, I introduce the notion of inference webs and argue that, while they are an essential part of our epistemic practices, they can be problematic and, when they are, they obscure our theorizing and wrong those who operate under them. Inference webs license (often invisible) inferences between views and principles which foreclose discussion and resist examination.

For example, in chapter 3 I argue that the mental illness literature operates under an inference web which adopts certain empiricist conceptions of reduction, realism, objectivity, and naturalism. Four (legitimate) principles are often adopted in that literature: 1. that naturalist accounts of mental illness must be reductive, 2. that naturalist accounts of mental illness must be realist, 3. that naturalist accounts of mental illness must endorse and allow for objectivity, 4. that naturalist accounts of mental illness must be empirical. The problem, I argue, is that these legitimate principles are imbued with empiricist assumptions that cause them to go awry. The most influential covert principle that influences this literature is the idea that the social and the scientific are fundamentally distinct. Most philosophers of science will deny this principle if asked directly about it, but I will show that the same principle is operative in many of the inferences and arguments that are routinely made in the literature.

In chapter 2, I argue that inference webs can be involved in some forms of ‘epistemic injustice.’ An epistemic injustice is an injustice perpetrated against an agent in their capacity as a knower. So far, various varieties of epistemic justice have been suggested in the literature, but I argue that all of these presentations overlook the very structural role that inference webs play in many cases of epistemic injustice. To show this, I introduce a new form of epistemic injustice – inferential injustice – which helps to distill the way in which inference webs can cause injustice. I propose that inference webs tempt people away from seeing certain epistemic agents as falling under certain concepts. For example, women who have heart attacks are frequently misdiagnosed comparative to their male peers. I argue that this is, in part, due to the fact that someone’s being a woman blocks (or provides resistance against) the inference to ‘this person is having a heart attack’ even when there is both i) a good understanding of what heart attacks are and ii) good evidence that the agent is having a heart attack. In these cases, the individual notions and the particularities of the case are well understood but inference webs prevent us from correctly inferring one from the other (or, encourage us to incorrectly infer one from the other).

The first chapter is something of an outlier. It is a paper in modal metaphysics that offers a solution to the problem of advanced modalizing which is levelled against David Lewis’ genuine modal realism. The problem of advanced modalizing is supposed to show that genuine modal realism is unable to accommodate perfectly reasonable modal claims such as ‘possibly, there are many possible worlds.’ I argue that the proposed problem is not a problem at all, since the ontology of Lewis’ account rules out those claim as category

mistakes. Though this paper does not fall squarely under the main theme of the dissertation, it is appropriate to include it in this dissertation because it represents a part of my philosophical expertise that is independent of that project but which I care about very much.

CHAPTER 1

ADVANCED MODALS, ADVANCED QUANTIFIERS, AND REDUCTION

Introduction

Cases of so-called ‘advanced modalizing’ are problematic for genuine modal realists in two big ways: i) they call into question the adequacy of the standard Lewisian translation schema for modal sentences, and (perhaps worse) ii) they indicate that genuine modal realism fails as an analysis of modality. So far, interlocutors in this debate have attempted to aid Lewis by either revising his standard translation schema, or recasting genuine modal realism as a non-reductive account of modality. I think these accounts fail to accurately locate the cause of the problem; rather than arising from the translation schema, the problem of advanced modalizing issues from the ontology of genuine modal realism. Here I argue that the ontology of modal realism requires that we take all advanced modals as false. However, this result should not be taken as any bad-making-feature of Lewis’ theory; it is just one of the necessary and expected features of his reductive realism. For to utter an advanced modal is to make a category mistake; what you say is, strictly speaking, false in spite of the fact that it looks true. The reason it looks true is down to the combination of i) the fact that we are raised as actualists and ii) the fact that advanced modals operate in the same kind of way as metaphors do.

The paper will proceed as follows. In section 1, I present the problem of advanced modalizing. In section 2, I argue that the problem does not issue from the translation schema, but from the ontology of genuine modal realism plus some theoretical constraints. Section 3 is where I argue that we should take all advanced modals as false,

despite the fact that doing so harms the principle of possibility introduction (PI). In section 4, I further support my claim by demonstrating that advanced modalizing-type problems arise for non-modal propositions as well. Finally, in section 5, I argue that advanced modals are like category mistakes. In what follows, ‘genuine modal realism’ and ‘modal realism’ will be taken to mean roughly the view laid out in (Lewis, 1983a, p. 30); that there is a plurality of concrete possible worlds which exist in just the way the actual world does and that these worlds form the reductive base for our modal claims.¹ Where notation is used, I will follow Lewis (Lewis, 1983a): $Wx = x$ is a world, $BSx = x$ is a blue swan, $Ixy = x$ is in possible world y .

1: The Problem of Advanced Modalizing

We all know genuine modal realism as the view that countenances outlandish claims such as (Plurality):

(Plurality) There are at least two (concrete, spatiotemporally isolated...)² possible worlds.

The vast majority of us also take for granted the principle of possibility introduction (PI)³:

(PI) Anything that is true is also possibly true.

¹ This characterization of the target view is rough, since you need not accept *all* of what Lewis says in ‘On the Plurality of Worlds’ in order for advanced modalizing to be a worry for you.

² For brevity, the parenthetical comment will be omitted in what follows.

³ Here I’m using John Divers’ label (Divers, 1999, p. 271).

(Plurality) and (PI) together imply (\diamond Plurality):

(\diamond Plurality) Possibly, there are at least two possible worlds.

So, we'd be within our rights to expect the Modal Realist to take (\diamond Plurality) as true. But the Modal Realist *can't* take (\diamond Plurality) as true. Modal realism is a reductive view; it translates modal sentences of the form 'Possibly, P' into non-modal sentences of the form 'there is a possible world⁴ at which P'. In order for (\diamond Plurality) to be true, its non-modal translation - (\exists Plurality) - must also be true:

(\exists Plurality) There is a world at which there are at least two possible worlds.

(\exists Plurality) is true just in case there is a world which has at least two possible worlds as parts. But according to modal realism, concrete possible worlds are spatiotemporally isolated from one another (Lewis, 1986, pp. 69–78). Hence the truth-conditions for (\exists Plurality) cannot be met and that sentence must be false.

Why does (\diamond Plurality) translate as (\exists Plurality)? A closed sentence of the form ' \diamond F' ('possibly, F') translates into counterpart theory as ' $\exists a(Wa \ \& \ F^a)$ ' ('there is a possible world a and F is true in a '). ' F^a ' is a complex sentence meaning 'F is true in a ' and it is

⁴ For brevity, I will sometimes use 'world' to refer to possible worlds throughout the rest of the paper.

formed by restricting any quantifiers in F to the domain of things in the world denoted by a (Lewis, 1983a, p. 30). So, when dealing with the sentence ‘Possibly, there are blue swans’ we take its formalization in quantified modal logic, which would be ‘ $\diamond \exists x (BSx)$ ’, replace ‘ \diamond ’ with ‘ $\exists w (Ww \ \& \ \dots)$ ’, and restrict all quantifiers in ‘ $\exists x (BSx)$ ’ to the domain of things in the world denoted by w. Altogether, this yields ‘ $\exists w (Ww \ \& \ \exists x (BSx \ \& \ Ixw))$ ’: ‘There is a w, w is a world and there is an x, x is a blue swan and x is in w.’

Take the formalization of (Plurality) as a sentence of Quantified Modal Logic.

(Plurality) $\exists y \exists z ((Wy \ \& \ Wz) \ \& \ y \neq z)$; ‘there is a y and there is a z, y is a world and z is a world and y is not identical to z.’

In quantified modal logic ‘ \diamond ’ is a sentential operator, so we get (\diamond Plurality) by sticking a diamond in front of (Plurality). This yields:

(\diamond Plurality) $\diamond [\exists y \exists z ((Wy \ \& \ Wz) \ \& \ y \neq z)]$; ‘*it is possible that* (there is a y and there is a z, y is a world and z is a world and y is not identical to z).’

To translate (\diamond Plurality) into counterpart theory, we use the procedure detailed above.

First we replace the ‘ \diamond ’ with ‘ $\exists w (Ww \ \& \ \dots)$ ’, then we restrict all of the quantifiers in (Plurality) to the domain of things in the world denoted by w. The result is (\exists Plurality):

(\exists Plurality) ‘ $\exists x (Wx \ \& \ (\exists y \ \exists z ((Iyx \ \& \ Izx) \ \& \ (Wy \ \& \ Wz) \ \& \ y \neq z)))$ ’; ‘There is a w, w is a world, and there is a y and a z, y is a world and z is a world and y is in w and z is in w and y and z are non-identical.’

You might be wondering why the Modal Realist can’t just read (\exists Plurality) as saying ‘there is a world *according to which* there are at least two possible worlds’ or as saying ‘there is a world which *sees* at least two possible worlds.’ John Divers (Divers, 1999) gives a full presentation of the options for understanding the ‘in a world’ relation and shows why none of these help in cases of advanced modalizing. I’ll rehash two of those explanations here: that for ‘being *partly in* a world’ and that for ‘*existing from the standpoint of* a world’ (Lewis, 1983b, p. 40). According to the former, x gets to count as being in world w if it is *partly in* w, that is, if x shares a part with w. If we make this amendment to the translation schema, the problem remains. ‘There is a possible world which shares a part with two possible worlds’ must be false since, on Lewis’ Modal realism, worlds do not overlap. According to the latter, x gets to count as *being in* world w if x exists *from the standpoint of* w. And x exists from the standpoint of w iff ‘it belongs to the least restricted domain that is normally – modal metaphysics being deemed abnormal – appropriate in evaluating the truth at that world of quantifications’ (Lewis, 1983b, p. 40). Even given this treatment, the sentence ‘there is a possible world such that two possible worlds exist from the standpoint of it’ is presumably false since the least restricted normal domain of a world excludes modal metaphysics and, hence, other possible worlds (Divers, 1999, pp. 223–225).

Intuitively, it is bad for the Modal Realist to have to claim that $(\exists\text{Plurality})$ is false. We can bring out the worry more formally by constructing the following *reductio* argument.

1. (Plurality) There are at least two possible worlds
2. (PI) If P, then possibly P
3. Therefore, $(\diamond\text{Plurality})$ possibly there are at least two possible worlds. (from 1, 2, by MP)
4. If $(\diamond\text{Plurality})$ possibly there are at least two possible worlds, then $(\exists\text{Plurality})$ there is a world at which there are at least two possible worlds
5. Therefore, $(\exists\text{Plurality})$ there is a world at which there are at least two possible worlds (from 3, 4, by MP)
6. *It is not the case that* there is a world at which there are at least two possible worlds; possible worlds are spatiotemporally isolated from one another, $\sim(\exists\text{Plurality})$.

From four assumptions (1, 2, 4, and 6), we get a contradiction (5 and 6). What to do? Formally speaking, we have the following options: We might simply reject 6 and allow that at least some worlds have shared parts. This would be radically revisionary to the Modal Realist's project, but not unprecedented. Takashi Yagisawa (Yagisawa, 2009), Kris McDaniel (McDaniel, 2004), and Phillip Bricker (Bricker, 2001) have advanced views that reject 6 but are broadly Lewisian. If those options don't suit, we could bar the route to 3 by denying either of 1 and 2. But the cost of doing so is high; 2 is a logical principle which is both extremely intuitive and almost ubiquitous and 1 is a claim at the very core of modal realism. It's quite possible (perhaps probable) that reader of this paper already

takes 1 to be false, but in the current context of assessing a problem for genuine modal realism, outright rejection of the view should be a last resort. Finally, we might retain 1 and 2, but bar the route from 3 to 5 by denying 4. This would amount to rejecting (or at least heavily revising) Lewis' classic translation scheme for modal sentences.

2: The Role of the Translation Schema

Most of the literature on this problem focuses on the translation schema. I think this is the wrong approach. That said, it is fairly easy to see why it looks as though the translation schema is what's at fault. In translating (\Diamond Plurality), the translation schema takes the quantifiers in (Plurality) and restricts them to the domain of things in some world, w . This action destroys the intended meaning of (Plurality), which was supposed to say something about the space of possibility as a whole, not just a single world. If we start with a claim about the pluriverse (that it contains many possible worlds) and append a modal operator to it, we should end up with a modal claim about the pluriverse. But what we in fact end up with is a modal claim about individual possible worlds. The difference that is made when the quantifiers are 'squished' in this way can be seen intuitively as well as formally. After all, whether it is necessary that my house contains the greatest hits of Wham! is quite a distinct question from whether it is necessary that the *world* does. In a similar fashion, whether it is necessary that there are many worlds in the *pluriverse* (a question familiar from Leibniz (Adams, 1994; Feit, 1998; Hudson, 1997, 1999)), is quite separate from whether there are many worlds in the actual world, which is quite separate from whether there are many worlds in my house or the city of Manchester or what-have-you. Since it is this 'squishing' of the quantifiers that leads to the false claim (\exists Plurality), it would be reasonable to suspect the translation schema of foul play.

Note, however, that the problem is not caused by the mere *manipulation* of the quantifiers.⁵ Quantifier manipulation is part of any use of the Lewisian translation schema and is in fact a part of the underlying metaphysical theory (not just the semantics for that theory). To be possible just *is* to be true at a world, so sentences like ‘possibly, there is a blue swan’ are made true by manipulating the quantifiers within the scope of the modal operator. In this way, in addition to offering a way to formalize modal claims, the translation schema reflects the reductive nature of Lewis’ account. It shows that anything⁶ expressible in quantified modal logic (which has brute modal operators) can be expressed in the language of counterpart theory (which does not). Hence possible and necessary truth can be reduced to truth-at-a-location; to be possibly true is just to be true at some locations; to be necessarily true is just to be true at all locations.

Ordinarily, the result is that a *modal* quantified claim about one world, *w*, becomes a *non-modal* quantified claim about some *other* world, *x*; the modal sentence ‘possibly, there is a blue swan’ is true of *this* world because the non-modal sentence ‘there is a blue swan’ is true of some *other* possible world. The translation schema’s *purpose* is to manipulate the domain of quantifiers that appear within the scope of a modal operator so that ‘possibly, *P*’ may be analyzed as ‘*P* at some world.’ As an example, the modal claim ‘Possibly, there is a guinea pig’ – ‘ $\Diamond \exists y (GP_y)$ ’ - translates as ‘ $\exists w (Ww \ \& \ \exists y (I_{yw} \ \& \ GP_y))$ ’ (Lewis, 1983a,

⁵ Parsons suggests talking of ‘advanced quantification’ rather than ‘advanced modalizing’ (Parsons, manuscript, pp. 14–15).

⁶ The language of counterpart theory can also express some sentences that quantified modal logic cannot (Lewis, 1983a).

p. 30). 'Iyw' explicitly restricts the domain of the second quantifier - $(\exists y (GP_y))$ - to the domain of things in the world denoted by w, resulting in the (English) sentence 'There is a world, x, and *in that world* there is a guinea pig.'

As a more concrete example, let's say that I start with the quantified modal logic interpretation of the sentence 'there are blue swans.'

(Swans) $\exists y (BS_y)$

(Swans) is false; there is nothing in the domain of quantification that is a blue swan. That is, if we were to look into the actual world and inspect all of its many inhabitants, we would not find a single blue swan. That said, (Swans) is at least *possibly* true. So, while (Swans) false, $(\Diamond Swans)$ is true.

($\Diamond Swans$) $\Diamond \exists y (BS_y)$

Now imagine that I want to translate $(\Diamond Swans)$ into the language of counterpart theory. Following the standard procedure, such a translation would yield the following sentence.

($\exists Swans$) $\exists x (W_x \& (\exists y (BS_y))^x)$

$(\exists Swans)$ says something true. It says that there exists some possible world and in that world there exists a blue swan. Now, recall that 'F^a' is a complex sentence meaning 'F is

true in a' and is formed by restricting any quantifiers in F to the domain of things in the world denoted by a (Lewis, 1983a, p. 30). As such the quantifier in (Swans), which started out ranging over the domain of things in the actual world, is manipulated so as to range over the domain of things in the world denoted by ' a .' We already know that ' a ' cannot denote the actual world, since that would make the sentence false. So the effect of the translation schema is to take a sentence about one world, and turn it into a sentence about some other world; a false sentence is possibly true for a world only if it is true simpliciter for some other possible world.

Ordinarily this is not problematic. After all, this is what the translation schema is *supposed* to do: it takes claims about what *might have been* true here, and makes them into claims about what *is* true in some other world. When we try to perform this operation on sentences like (Plurality) however, we run into problems. (Plurality), recall, is formalized in quantified modal logic as follows:

$$\text{(Plurality)} \exists y \exists z ((Wy \ \& \ Wz) \ \& \ y \neq z)$$

On the most appropriate reading of (Plurality), the domain of quantification contains at least two possible worlds. When translating (\Diamond Plurality) into counterpart theory, however, the quantifiers in (Plurality) are distorted in such a way that they only range over the things in a single possible world. That action distorts the meaning of the initial sentence.

For example, say that there are two potential domains of quantification: D1 and D2. D1 contains my house and the house next door as well as the contents of both houses

whereas D2 contains only my house and its contents. The sentence ‘There are at least two houses’ would be true if the domain of quantification were D1 but false if it were D2. In the event that I utter that sentence, charity and context will indicate that we should understand the domain of quantification to include at least two houses. Similarly, charity and context indicate that, in the event that I utter ‘there are many possible worlds’, we should understand the domain of quantification to include at least two possible worlds. When that sentence is appended to a modal operator and then run through the Lewisian translation schema, the result is that the domain of quantification is ‘squished’ so that it contains only a single world. (Plurality) is different from ordinary modal claims like (Swans) because in plurality the quantifiers are not just taken from the domain of one world to the domain of another, but are taken from a domain including more than one world to a domain including a single world.

The problem of advanced modalizing, then, comes about because a claim about several worlds is forced into being a claim about a single world. It is the squishing, and not merely the manipulating of the quantifiers that causes issues. The action of constraining quantifiers in this way is not simply a feature of the translation schema, it is a feature of the metaphysics.

The translation schema is telling us something here though, it is telling us that there is something wrong with the original target sentence (\diamond Plurality). If you submit to the Lewisian theory, then you must think that possibility reduces to a sort of distributional property. That is, for something to be false but possibly true *just is* for it to be (false here but) true at some other possible world. The spread of possible worlds is laid out across

the pluriverse and the spread as a whole determines what the modal facts are. That is, whether some sentence is contingently true depends not just on what is going on in the actual world, but also on what is going on at the other possible worlds. It is important to this reductive picture that the worlds are sharply delineated; if I want to know whether ‘there is a round square’ is true at any world, it is important that I take each world wholly and individually. It will not do for me to look at a handful of worlds together and determine that the fusion of my actual (round) coffee cup with my merely possible (square-shaped) TARDIS coffee cup can make ‘possibly there is a round square’ true.

As a result of this reductive relationship between the pluriverse and the modal facts, I submit that it is wholly inappropriate to ask modal questions of the pluriverse. As a good modal realist, I will agree that there are many worlds, but once you start asking me whether it is possible that there are many worlds, I begin to lose grip of what you want to know. There is something wrong with (\Diamond Plurality); it is an *illegitimate* sentence.

Given the modal realist’s theoretical commitments, it is inappropriate to ask her to offer an analysis of (\Diamond Plurality). At best, (\Diamond Plurality) is something like a category mistake and is always false. In the next section, I will offer some analogies that I hope will bring out this claim as well as arguing that the loss of (PI) is not so big a deal. In section 4, I will offer further support for my claim that it is the ontology - and not the semantics - of modal realism that makes advanced modals problematic, by showing that similar issues arise for non-modal propositions. In section 5, I will offer a deeper account of the truth-values of advanced modals and what it means to say that they are ‘illegitimate.’

3: A Genuine Modal Realist's Response to the Problem

It is widely assumed that, whatever else she says, the modal realist must at least get herself into a position to say that (\Diamond Plurality) is true. For many, this means amending the translation schema so that the translation of (\Diamond Plurality) is true. I think that this is wrong-headed; the genuine modal realist should claim that (\Diamond Plurality) is false. This is not because her translation schema requires it, but because her ontology does. Since, for the modal realist, the modal facts reduce to distributional facts about the pluriverse, it simply *doesn't make sense* to ask what the modal status of these distributional facts are. The best way to demonstrate my point is with the use of analogies. I'll give you two.

3.1 *Advanced Localizing*

Imagine that you and I are working on a puzzle together (see fig. 1). We get to talking philosophically, and I suggest an analysis of what it is for something to be true

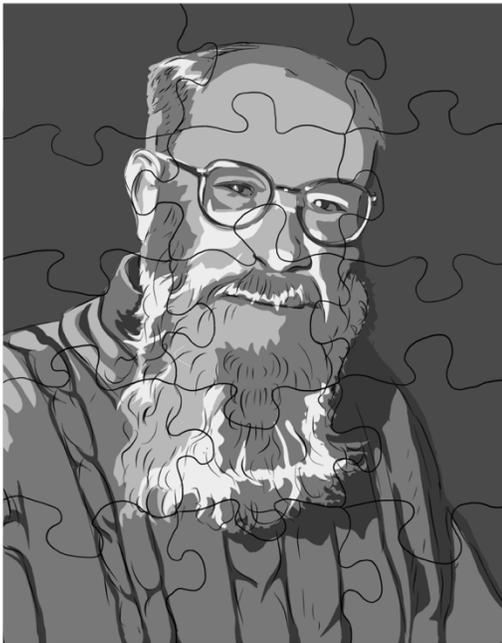


Fig. 1 Lewis Puzzled

everywhere in the puzzle ('everywhere-truth') vs. what it is for something to be true somewhere in the puzzle ('somewhere-truth'). My suggested analysis is as follows:

(Everywhere-Truth): P is everywhere-true iff for any puzzle piece, P is true of that piece.

(Somewhere-Truth): P is somewhere-true iff for some puzzle piece, P is true of that piece.

The sentence ‘there is an eye’ would, on this analysis, be an example of a somewhere-truth, since it is true of some piece of the puzzle that ‘there is an eye’. It would not, on the other hand, be an example of an everywhere-truth because ‘there is an eye’ is not true of all pieces. Contrastingly, ‘there is some pigment’ *would* be an example of an everywhere-truth, since this sentence is true of each individual piece of the puzzle.

Imagine now that you present me with a problem case. You say...

But what of the sentence “there is a beard”? It seems to me that, although this sentence is demonstrably true, it can have neither the status of everywhere-truth (since it is not true of every piece of the puzzle) nor the status of somewhere-truth (since it is not true of any one piece of the puzzle on its own). Surely this is unacceptable since ‘there is a beard’ is demonstrably true. How can a sentence be true, but not somewhere-true? After all, claiming that a sentence is true but not true anywhere sounds dangerously close to saying that it is true and not true... ’

This problem case assumes (LI), the localizing analogue of (PI):

(LI) if P is true, then P is somewhere-true.

But given the analysis I offered, (LI) must be false. Since my account analyses somewhere-truth and everywhere-truth in terms of localized goings-on at particular regions, we should fully expect that ‘there is a beard’ might be true but not somewhere-true. ‘There is a beard’ is *just not the kind of thing* that can be somewhere- or everywhere-

true. That may seem odd given our ordinary usage of ‘somewhere’ and ‘everywhere’, but my analysis can still be helpful in allowing us to sort puzzle pieces, and maybe even more helpful than any analysis which allows ‘there is a beard’ to be somewhere-true.

Similarly, genuine modal realism analyses modal claims in terms of how modal space as a whole plays out; what’s possible in one region depends on what’s going on in the other regions and the modal facts are reflected in the similarities and differences between possible worlds. We can perfectly legitimately make claims about what is true of the overall spread of possible worlds – claims about the *pluriverse as a whole*. But these claims are not apt to be modalized, just as advanced spatialists are not apt to be somewhere- or everywhere-true.

3.2 Advanced Temporalizing

Now imagine that you and I are discussing the B-theory. I suggest to you that, given the B-theory, an important distinction among the truths is between those that are true some of the time (the sometimes-truths) and those that are true all of the time (the always-truths). ‘Fran is sitting’ and ‘something exists’ are both truths, but they are importantly different. ‘Fran is sitting’ is true only when indexed to certain times, whereas ‘something exists’ is true regardless of which time it is indexed to. Because this distinction seems intuitive and important, I decide to offer an analysis for it. That analysis goes as follows:

(Always-Truth) P is always-true iff P is true at every time slice

(Sometimes-Truth) P is sometimes-true iff P is true at some time slice.

‘There are guinea pigs’ is sometimes-true, since it is true of some time-slices (this one) but not true of others (those in the very distant past). Contrastingly, ‘something exists’ is always-true since it is true of every time-slice.

But now we encounter a problem. You may think that, given the meaning of sometimes- and always-truth, any true sentence must also be sometimes-true. After all, intuitively if x is not sometimes-true, then x is *never* true, which means that x is false. This intuition reflects a principle I will call ‘sometimes-truth introduction’ (SI). On my theory, the sentence ‘there are at least two time-slices’ is true. So given the principle (SI), ‘there are at least two times slices’ ought to be sometimes-true. But my analysis cannot allow for that. ‘There are at least two time-slices’ is not true of *any* time-slice, since no single time slice contains two time-slices.

It seems to me a perfectly acceptable response to insist that sentences like ‘there are at least two time-slices’ are just not the kinds of things that get to be sometimes- or always-true. This needn’t have the result that sentences like this are true but never-true; a sentence which is not apt to be sometimes-true is just as unable to be never-true. Consider our conversation progressing as follows:

You: What about the sentence ‘there are swans’; is that sentence always-true or sometimes-true?

Me: well, since that is only true at some time-slices...

You: Oh, don’t misunderstand me. When I say ‘there are swans’ I mean to have my quantifiers range over *all* of the time slices, not just one. My sentence

is something more like ‘there are swans *somewhere in space-time*’ or ‘there are (*ranging over all time-slices*) swans’

Is ‘there are (*ranging over all time-slices*) swans’ something that is true all of the time or something that is true only some of the time? To say that it is sometimes-true feels inappropriate; the overall spread of time-slices does not change, so it would seem odd to suggest that this sentence is sometimes-true and sometimes-false, after all in virtue of what would it be false rather than true? On the other hand, it also feels inappropriate to claim that ‘there are (*ranging over all time-slices*) swans’ is always-true. ‘Something exists’ is always true because there is always, in each time-slice, something to make it true. ‘There are (*ranging over all time-slices*) swans’ is not like that; its truth is not ‘reaffirmed’ by every time-slice. Both of these options feel wrong because it is a mistake to attribute sometimes- or always-truth to this sentence; there are some sentences that just aren’t apt to be sometimes- or always-true.

You might have very legitimate, independent reasons for wanting to reject the above analyses of advanced localizing and advanced temporalizing. But it ought not be on the grounds of their inability to analyze advanced sentences. To do so would be to reject our definition of what it is to be green because it does not accommodate the number 2.

3.3 Advanced Modalizing

Advanced modalizing is unproblematic for genuine modal realism in just the way that advanced temporalizing and advanced spatializing are unproblematic for the above analyses. The modal realist analyzes possibility and necessity in the following way:

(Possibility) P is possible iff P is true at some possible world.

(Necessity) P is necessary iff P is true at every possible world.

As a result of this analysis, there will be some truths that are neither possible nor necessary. This is not a problem. ‘There are at least two possible worlds’ is just *not the kind of thing* that can have the predicates of possibility or necessity attached to it. This has the direct result that (PI) is false, just as (SI) and (LI) are false.

If you are still concerned about losing (PI), or if you are fretting about how you will teach your modal logic proofs next semester, be soothed by the fact that the loss of (PI) in this instance is very well-contained. You may continue to make use of (PI) in all the ways you previously have; it is only when dealing with these very particular, and frankly rather strange, kinds of sentences that (PI) faces any difficulty. This helps, I think, in that our intuitions in favor of these principles are strongest when applied to worldbound claims. (PI) seems so intuitive because if something is true of this world, then surely it’s true that it might have been true – this world is one of the ways the world might have been! Once we start making non-worldbound claims, the same intuitions do not hold as strongly. It is certainly not obvious that, if something is true of the pluriverse as a whole, then it is true of at least one possible world, just as it is not obvious that some property which holds of a jigsaw puzzle must also hold of some individual piece of that puzzle.

A further worry you might have is that by failing to count ‘there are many worlds’ as possible, we are forced to count it as impossible; after all, whatever is not possible is

impossible, right? Wrong. At least for certain cases. When I say that ‘there are many possible worlds’ is not possible, I do not mean to ascribe to it the modal status of being not possible. My claim is that ‘there are many worlds’ is *not-modalizable*. It is not possible, but nor is it *impossible*; it simply lacks a modal status.

4: Advanced Propositions

As already stated, I think it is wrong-headed to think that something has gone wrong in the way that the translation schema deals with advanced modals. I want to demonstrate my point by presenting you with another problem for modal realism. This problem is directly analogous to the problem of advanced modalizing, but it makes no use of the translation schema.

Consider again that troublesome sentence, (Plurality).

$$\text{(Plurality)} \exists y \exists z ((Wy \ \& \ Wz) \ \& \ y \neq z)$$

Now ask yourself about the *proposition* that that sentence expresses. According to Lewis, a proposition is identified with the set of possible worlds at which it holds:

I identify propositions with certain properties - namely, with those that are instantiated only by entire possible worlds. Then if properties generally are the sets of their instances, a proposition is a set of possible worlds. A proposition is said to hold at a world, or to be true at a world. The proposition is the same thing as the property of being a world where that proposition holds; and that is the

same thing as the set of worlds where that proposition holds. A proposition holds at just those worlds that are members of it. (Lewis, 1983a, pp. 53–54)

In order to understand which worlds get into which propositions, we need to know more about what it means for a proposition to *hold* at a world. There are two potential routes one might try to take. I'll address them each individually.

4.1. Restricting using the 'in' relation

In determining whether some proposition holds at a world, we might try to make use of the procedure that Lewis gives us for translating sentences of quantified modal logic into counterpart theory. In 'Counterpart Theory and Quantified Modal Logic' he tells us that we can get the sentence ' Φ holds in world β ' from ' Φ ' by *restricting quantifiers in the sentence to the domain of the world in question*:

To form the sentence Φ^β (Φ holds in world β) from the given sentence Φ , we need only restrict the range of each quantifier in Φ to the domain of things in the worlds denoted by β ; that is, we replace $\forall\alpha$ by $\forall\alpha (I\alpha\beta \supset \dots)$ and $\exists\alpha$ by $\exists\alpha (I\alpha\beta \ \& \ \dots)$ throughout Φ . (Lewis, 1983a, p. 30)

Take (Swans). We formalize this in quantified modal logic as ' $\exists y (BSy)$.' Using the above procedure, we can form the sentence '(Swans) holds in w ' by restricting the domain of the quantifier in (Swans) to things in the world denoted by w . That would give us ' $\exists y (BSy \ \& \ Iyw)$.' In this way, the proposition denoted by (Swans) will be identified with the set of worlds, w , for which ' $\exists y (BSy \ \& \ Iyw)$ ' is true. That is to say, the proposition denoted by

(Swans) will be identified with the set of worlds which contain blue swans. Necessary propositions (which hold at every world) are identified with the set of all worlds. Impossible propositions (which don't hold at any world) are identified with the empty set. Two propositions are identical when they are identified with the same set.

With which set, we might ask, is (Plurality) identified? Following the above analysis, this proposition should be identified with the set that includes any world, w , for which ' $\exists z ((Ww \ \& \ Wz) \ \& \ Izw)$ ' is true. But this sentence must be false for any possible world denoted by ' w ' because no possible world shares a part with any other. As a result, the proposition that there are at least two possible worlds (on this procedure) is identified with the empty set.⁷ That looks bad for the Modal Realist; no view should be forced, by its own lights, to claim that it is itself impossible.⁸ I take it that anyone who is concerned by the fact that (\exists Plurality) is false ought to be doubly concerned by the fact that (Plurality) itself is not just false but *impossible*. But this problem arises without making any use of the Lewisian translation schema. Since this problem shares so much in common with the problem of advanced modalizing, we ought to suspect that the translation schema is not at the root of the problem in either case.

4.2. *Anchoring to a world*

⁷ Hud Hudson (Hudson, 1997, p. 81) makes a similar argument as part of a debate that moves between him and Neil Feit (Feit, 1998; Hudson, 1997, 1999) over whether the existence of contingent facts undermines the principle of sufficient reason.

⁸ Well, perhaps Dialetheism would, but you take my point.

But perhaps the procedure I use above is not the right one. After all, (Plurality) is a sentence of counterpart theory and Lewis is not exactly clear on how the semantics for sentences of counterpart theory should work. Perhaps a more charitable process to employ would be one which anchors the sentence to a named world.⁹ In this connection, (Swans) would confuse us since (Swans) is a sentence of quantified modal logic rather than of counterpart theory, so take instead the counterpart theory formalization of ‘possibly, there are blue swans,’ (\exists Swans).

$$(\exists\text{Swans}) \exists x (Wx \ \& \ \exists y (BSy \ \& \ Iyx)).$$

The new procedure tells us that the proposition denoted by (\exists Swans) holds in the world w if the following is true:

$$(\mathbf{w}\text{Swans}) (Ww \ \& \ \exists y (BSy \ \& \ Iyw))$$

Here the domain of quantification is left untouched and includes all of the possible worlds but we are able to learn about what holds at an individual world, w , by naming that world.

What of (Plurality)? As above, the proposition denoted by (Plurality) will hold in some world w if the following is true:

⁹ Thanks to Will Starr for helping me to see this.

(wPlurality) $\exists z ((Ww \ \& \ Wz) \ \& \ y \neq w)$

(wPlurality) says something like (when spoken by world w) 'I am a world and there are many other worlds that are not me.' The idea being that we can approach each possible world, x, y, z, etcetera and ask whether the equivalent sentence - (xPlurality), (yPlurality), (zPlurality) - is true. That procedure will tell us which worlds (Plurality) holds in, and hence which worlds get into the set with which (Plurality) is identified. Contrary to the first procedure, this procedure would have the result that the proposition denoted by (Plurality) holds in every possible world since, no matter which world we name, 'I am a world and there are many other worlds that are not me' will be true for that world. Hence, using this procedure, (Plurality) is both true and necessarily true.

I maintain that this procedure is also problematic, in spite of the fact that it avoids making (Plurality) an impossible proposition. Compare (wPlurality) with (wSwans). (wSwans) is true only for worlds containing blue swans because it says something like 'I am a world and there is a blue swan which is in me.' The quantifier ranges over the entire pluriverse but the 'in' relation constrains it to the domain of things within w. In this way, it matters which world we are talking to when we ask whether 'I am a world and there is a blue swan which is in me' is true. (Plurality) on the other hand lacks an 'in' relation, and as such its truth really isn't affected by which world we are talking to. This procedure makes (Plurality) trivially true *and* trivially necessarily true.

Perhaps that doesn't trouble you. But consider this; any proposition that, like (Plurality), lacks an 'in' relation will come out trivially true (and trivially necessarily true) in the same

way. Imagine a robot, Bob, who speaks only the language of counterpart theory. I tell Bob that ‘there are blue swans’ is true and he formalizes that sentence in the only way he knows how; into counterpart theory. What will this sentence look like? You might be tempted to say ‘ $\exists x \exists y (Wx \ \& \ BSy)$ ’, but that’s not quite right. I didn’t say anything about worlds to Bob, I only told him about swans. There is no reason for him to import any information about worlds into his translation of the target sentence. Perhaps if he knew Quantified Modal Logic, he would be tempted to interpret my claim charitably as being one about what exists in the actual world. Then he would form it in quantified modal logic first and then translate *that* sentence into counterpart theory. If Bob were to do that, then he’d likely end up with the formalization ‘ $\exists x \exists y (Wx \ \& \ BSy)$.’ *But Bob doesn’t know quantified modal logic*, all he can do is take my exact words and form a corresponding sentence of counterpart theory.

So, Bob translates ‘there are blue swans’ into counterpart theory as ‘ $\exists x (BSx)$.’ Now we might ask ‘which set of worlds is denoted by the sentence Bob gave us?’ Using the second procedure above, we should say that the proposition denoted by ‘ $\exists x (BSx)$ ’ includes any world, w , for which BSx is true. Since the domain of quantification includes all of the possible worlds, this sentence must be true no matter which world we wish to assess; the world itself really never comes into it. This has the result that any proposition like that denoted by (Plurality) will be trivially true by the lights of the modal realist’s reductive ontology. But here’s the really big problem; due to Lewis’ account of propositions, the proposition denoted by (Plurality) and that denoted by ‘ $\exists x (BSx)$ ’ and that denoted by any similar claim will be the same proposition; they will all be identified with the set of all worlds.

One could bite the bullet at this point and accept that any proposition which makes a claim about the pluriverse at large (rather than an individual world) is trivially true, and necessary, and identified with the set of all worlds. But this state of affairs does indicate that facts about the pluriverse at large are vacuous in a certain sense. They don't really tell us about what's possible or necessary because they don't tell us anything about individual worlds. For that reason, we should be suspicious of attempts to modalize facts about the pluriverse (such as (Plurality)). (Plurality) and $\exists x (BSx)$ are just not modalizable; they are not apt to be modalized.¹⁰

5: Category Mistakes

At this point, you might be wondering two things; i) what exactly do I mean when I say that certain truths are 'not-modalizable?'; and ii) how can these sentences be illegitimate when they look so respectable? My answer to both is to suggest that 'possibly, there are many possible worlds' constitutes a category mistake.

A category mistake is a particular kind of infelicitous sentence. A classic example is 'the theory of relativity is sleeping.' It's immediately obvious that this is a weird sentence; it's tempting to call it 'false,' but very counterintuitive to say that it's negation is true; it seems inaccurate to claim that it is complete nonsense, but remains unclear how exactly to make sense of it. Just how we should understand these sentences, and what goes wrong with them, is extremely controversial. An intuitive explanation is that some predicate or property is being assigned to an object that *just isn't the kind of thing to have that property*. So,

¹⁰ Thanks to Will Starr for discussion here.

for example, ‘George Michael is from North Manchester’ is false because a certain property - being from North Manchester - is attributed to an object - George Michael - which does not have that property. In contrast, ‘the number three is worried about completing its dissertation’ is a category mistake because not only does the object - the number three - fail to have the property - being worried about finishing it’s dissertation - but it *fails to even be a candidate* for having that kind of property; numbers are just not the kinds of things that can be worried. This can help us make sense of the fact that, while ‘the number three is worried about completing its dissertation’ looks false, that is not to say that its negation is true; the sentence is in some sense broken.

I think something similar is going on in the case of advanced modalizing: advanced modals are false, but this is not to say that their negations are true. ‘Possibly, there are at least two possible worlds’ is false, but so is ‘*it is not the case that, possibly there are at least two possible worlds*’. That is to say, while it is not possible that there are at least two possible worlds, this fact does not entail that it is impossible that there are at least two possible worlds, just as its being false that the theory of relativity is sleeping does not entail that the theory of relativity is not sleeping, or that the theory of relativity is awake.

What truth-value are we to assign to a category mistakes? Ofra Magidor suggests that, metaphor being a variety of category mistake, one might employ the Gricean picture of metaphor to understand category mistakes (Magidor, 2015).¹¹ That account has it that metaphorical sentences such as ‘one must walk the line between productivity and procrastination,’ take their meaning from the literal expression of the sentence (Grice,

¹¹ Though she does not endorse this picture herself (Magidor, 2013).

1991). In this way, all metaphors are strictly speaking false; there is no ‘line’ between productivity and procrastination that one could literally walk along. However, metaphors can *gain* meaning through the negotiation of conversational implicatures through which certain maxims are disregarded. So for example, imagine that I utter the sentence ‘my paper is angry that I didn’t work on it yesterday.’ The literal meaning of that sentence is false; my paper is not angry because philosophical papers are just not the kinds of things that can be angry (thank God). But more than merely uttering a falsehood, in uttering this sentence I break one of the maxims of conversation. For example, I trigger the presupposition that my paper experiences emotions. But most conversational contexts cannot accommodate that presupposition since in most cases we take it for granted that papers do not experience emotions. Triggering a presupposition which cannot be accommodated in this way results in an infelicitous sentence (Magidor, 2013, pp. 131–132). However, since you my conversational partner, are a reasonable interlocutor you disregard that maxim, thus giving rise to the relevant metaphorical meaning *making use of* the literal content of the sentence (Magidor, 2015).

Making use of this account of metaphor (as Magidor suggests that we do), we get an intuitive diagnosis of category mistakes according to which they are infelicitous because they trigger presuppositions that are not acceptable in the conversational context. If I’m right that advanced modals are category mistakes then, on this picture, (Plurality) and its kin are (strictly speaking) false although this does not entail that their negations are true.

But why even think that advanced modals are category mistakes? ‘My computer desk is a naturalist about biological species’ is obviously weird; it’s immediately clear that

something has gone wrong with this sentence. But ‘possibly, there are many possible worlds’ looks pretty good - it at least doesn’t look any weirder than other sentence about possible worlds. Why one earth should we think that there’s anything wrong with it? The burden is on the modal realist to explain why.

I think that (Plurality) masquerades as a felicitous sentence because it has a close analogue that is felicitous in certain (common) contexts. Most of us would agree that ordinary conversational contexts are actualist; they presuppose that quantifiers range over only objects in the actual world. As such, in those ordinary conversational contexts modal predicates such as ‘necessary’ invoke the presupposition that our quantifiers are *worldbound*. Advanced modals suffer from *presupposition failure* because they presuppose absolutely unrestricted quantifiers, and this presupposition is not accommodated by most conversational contexts. The presupposition trigger for sentences of the form ‘x is necessary’ will always be accommodated as long as we are actualists; the actualist simply does not have the ontological resources to make advanced quantifications, so her quantifiers will always be worldbound and modal presuppositions always met. It is only when we embrace the genuine modal realist’s ontology that we can start to talk in advanced terms. So the presuppositional failure only kicks in when we fully commit to the ontology of the genuine modal realist. This can explain why some advanced modals appear to be benign; we are so used to assessing quantified sentences in terms of our actualist sensibilities that all quantifications are read as being worldbound, even after we make the shift to genuine modal realism.

Conclusion

Because of the modal realist's expanded ontology, she is able to utter sentences that others can't; namely 'advanced' sentences like (Plurality). The problem of advanced modalizing results from a misunderstanding of the status of those kinds of sentences. Advanced modals come out as 'unanalyzable' given the modal realist's ontology, but that is no threat to the reductive status of her account. In addition, while it is true that advanced modals come out false given the Lewisian translation scheme, this is not the disaster that most have taken it to be. We should regard advanced modals as committing category mistakes – they attempt to assign the properties of possibility and necessity to things that are *just not apt to be modalized*. This has the result that some truths – such as (Plurality) – are neither necessary nor possible nor even impossible. That is to say, the sentences 'possibly, there are at least two possible worlds', 'necessarily, there are at least two possible worlds' and 'it is impossible that there are at least two possible worlds' are false, even though 'there are at least two possible worlds' is true. This, of course, has the result that (PI) is false: not all true sentences are possible. But this fact does no serious violence to our original intuitions about modality. Modal realism can keep its translation schema, and its reductive ontology, even in the face of advanced modalizing.

References

- Adams, R. M. (1994). *Leibniz: Determinist, Theist, Idealist*. Oup Usa.
- Bricker, P. (2001). Island Universes and the Analysis of Modality. In G. Preyer & F. Siebelt (Eds.), *Reality and Humean Supervenience: Essays on the Philosophy of David Lewis*. Rowman & Littlefield.
- Divers, J. (1999). A genuine realist theory of advanced modalizing. *Mind*, 108(430), 217–239.
- Feit, N. (1998). More on brute facts. *Australasian Journal of Philosophy*, 76(4), 625 – 630.
- Grice, P. (1991). *Studies in the Way of Words*. Cambridge: Harvard University Press.
- Hudson, H. (1997). Brute facts. *Australasian Journal of Philosophy*, 75(1), 77 – 82.
- Hudson, H. (1999). A true, necessary falsehood. *Australasian Journal of Philosophy*, 77(1), 89 – 91.
- Lewis, D. K. (1983a). Counterpart Theory and Quantified Modal Logic. In *Philosophical Papers, Volume I* (pp. 26–39). New York: Oxford University Press.
- Lewis, D. K. (1983b). Postscripts to “Counterpart Theory and Quantified Modal Logic.” In *Philosophical Papers, Volume I* (pp. 39–46). New York: Oxford University Press.
- Lewis, D. K. (1986). *On the Plurality of Worlds*. Wiley-Blackwell.
- Magidor, O. (2013). *Category Mistakes*. Oxford University Press.
- Magidor, O. (2015). Category mistakes and figurative language. *Philosophical Studies*, (1), 1–14.
- McDaniel, K. (2004). Modal realism with overlap. *Australasian Journal of Philosophy*, 82(1), 137 – 152.
- Parsons, J. (manuscript). *Against advanced modalizing*.
- Yagisawa, T. (2009). *Worlds and Individuals, Possible and Otherwise*. Oxford University Press.

CHAPTER 2

INJUSTICE IN THE SPACES BETWEEN CONCEPTS

Introduction

I argue that epistemic injustice manifests not only in the *content* of our concepts, but in the *spaces* between them. Others have shown that epistemic injustice arises in the form of ‘testimonial injustice’ - where an agent is harmed because her credibility is undervalued - and ‘hermeneutical injustice’ - where an agent is harmed because some community lacks the conceptual resources that would allow her to render her experience intelligible. I think that epistemic injustice also arises as a result of prejudiced and harmful defects in the inferential architecture of both scientific practice and everyday thinking. Drawing on lessons from the philosophy of science, I argue that the inferential architecture of our epistemic practices can be prejudiced and wrongful, leading to a variety of epistemic injustice that I am calling ‘inferential injustice.’ This type of injustice is fully structural; it inheres in our epistemic practices themselves rather than as a direct result of an individual’s action. For this reason, cases of inferential injustice are importantly different from extant cases of epistemic injustice, and are especially hard to track. We need a better understanding of inferential injustice so that we can avoid and ameliorate cases such as the ones I present here.

The paper will proceed as follows. In section 1, I present several cases of epistemic injustice; two cases which are familiar from the literature and three novel cases that are the target of this paper. In section 2, I situate those cases within the broader context of

Miranda Fricker's account of epistemic injustice. This groundwork will allow me, in section 3, to offer a full analysis of inferential injustice.

1: The Cases

An epistemic injustice is an injustice that wrongs someone *specifically in their capacity as an epistemic agent*.¹² As a good first approximation, we can say that an epistemic injustice consists of the following three components:

(Harm component) Epistemic agent, A, or group, G, suffers an epistemic harm, H.

(Wrong Component) H is *wrongful*.

(Injustice Component) prejudice forms part of the causal story for H.

Miranda Fricker seems to require that each of these components be present in any instance of epistemic injustice; to suffer epistemic harm is sufficient to count as epistemic injustice as only some epistemic harms are wrongful (Fricker, 2007, pp. 19–20) and to suffer epistemic wrong is not enough, since some epistemic wrongs occur by accident or simply as a result of misfortune and are hence not unjust (Fricker, 2007, pp. 41–60, 161–176).¹³ In her 2007 book, *Epistemic Injustice*, Fricker offers up two varieties of epistemic

¹² The notion of a distinctly *epistemic* harm is pivotal for Fricker. According to her, epistemic injustices must involve harms that are epistemic in nature. For what it's worth, I am undecided on the matter of whether there is a uniquely *epistemic* type of harm in the sense that Fricker seems to want. This is not the place for that discussion, however, since the arguments I make here will stand even if it turns out that Fricker's cases of epistemic injustice do not involve any distinctly epistemic harm.

¹³ It remains opaque to me what the distinction is between just and unjust epistemic wrong. According to Fricker, epistemic wrong is distinct from other types of wrong insofar as it insults some agent's epistemic

injustice: testimonial injustice and hermeneutical injustice. I'll present the standard cases of those types of epistemic injustice in subsections 1.1 and 1.2. With this grounding, I'll be able to move forward with the presentation of my three novel cases (subsections 1.3-1.5).

1.1 Warning about Murder: Marge Sherwood

Marge¹⁴ is worried about her husband, Dickie. Dickie left for a boat trip several days ago with his mysterious new friend, Tom Ripley, and he hasn't been seen since. Marge doesn't trust Tom, and suspects he has done something terrible to Dickie. In her fear and distress, she confides in her father-in-law, Herbert Greenleaf, who dismisses her concerns and simply responds 'Marge, there's female intuition, and then there are facts' (Fricker, 2007, p. 88; Minghella, 2000). As it turns out, Marge's concerns are well-founded - Ripley had murdered Dickie - but she is not taken seriously by Greenleaf despite the fact that he has no obvious reason to reject Marge's testimony.

Marge suffers an epistemic harm when Greenleaf fails to correctly estimate her credibility. This harm is also *wrongful*, because Greenleaf *underestimates* Marge's credibility (rather than

autonomy, for example by failing to treat them as a full epistemic agent (Fricker, 2007, pp. 129–147). She uses a Kantian approach here; in cases of testimonial injustice, this wrong takes the form of epistemic objectification in which a testifier is regarded as an object in epistemic terms, as opposed to a full epistemic subject capable of generating and managing knowledge. It is not clear to me, given this framework, how one might wrong someone as an epistemic agent (in this Kantian sense) without also perpetrating an injustice.

¹⁴ This example (from Fricker) is taken from Anthony Minghella's screenplay of *The Talented Mr. Ripley*, which is itself based on Patricia Highsmith's novel of the same name (Highsmith, 2008).

overestimating it).¹⁵ What's more, the fact that this happens to Marge is not merely the result of some harmless mistake (as it would be if Greenleaf just didn't realize the extent of Marge's expertise on the subject). Nor is it the result of simple epistemic bad luck (as it would be if Marge were unlucky enough to exhibit a facial tic which is otherwise a good indicator of untrustworthiness).¹⁶ Rather, Greenleaf rejects Marge's testimony due to the influence of prejudice; that women are 'over-emotional,' or 'irrational' or some such.

Marge suffers an epistemic injustice in virtue of the fact that i) she is epistemically harmed, ii) that epistemic harm is wrongful, and iii) that wrong comes about as a result of underlying prejudices (Fricker, 2007, pp. 14–20). More precisely, she suffers a *'testimonial injustice,'* which is a variety of epistemic injustice in which a testifier's credibility is undervalued on the basis of prejudice (Fricker, 2007, p. 4).

1.2 Sexual Harassment: Carmita Wood

¹⁵ On Fricker's view, I cause someone epistemic harm whenever I incorrectly evaluate their credibility, but I only *wrong* them if I *undervalue* it (Fricker, 2007, pp. 19–21). I'm not entirely convinced by this claim. It seems to me that one might be epistemically wronged as a result of their credibility being *overestimated*. Imagine, for example, that Greenleaf were to take Marge's testimony *too* seriously on the basis of a belief that women are inherently intuitive and can 'sense' danger. This harms Marge as an epistemic agent because it does not take seriously her rationality and treats her merely as a passive 'vessel' for knowledge. This issue will not be important for the purposes of this paper, however, and hence will not be discussed further here.

¹⁶ For what it's worth, I am personally very skeptical of the idea that 'bad luck' is so straightforwardly a bar to epistemic injustice. However, this is not the place for a discussion of that issue and so it will be set aside for the remainder of the paper.

Carmita Wood, an employee at Cornell University, became a victim of sexual harassment before that term had been introduced into the lexicon (Brownmiller, 1990, pp. 280–281). As a result, Wood had an experience which was extremely detrimental to her, but which she lacked the proper conceptual machinery to process. This leaves her unable to approach her employers with any recognizable complaint and even unable to fully articulate the nature of her experience to herself. She subsequently becomes quite ill, is unable to find support or relief in her job and is forced to quit.

According to Fricker, Wood suffers an epistemic *harm* because, at the time of the assault, the epistemic community literally lacked the hermeneutical resources necessary for her to understand her experience.¹⁷ As a result, Wood is unable to render that experience intelligible - either to herself or to others - and that harms her epistemically. That harm is also *wrongful* because the experience is one that it is strongly in her interests to understand (Fricker, 2007, p. 162).¹⁸ Finally, the wrong she suffers is also *unjust* because it is not merely the result of epistemic bad luck (as it would be if Wood were unlucky enough to have an experience which we haven't gotten around to conceptualizing yet) but of the

¹⁷ This is controversial. As we will see in more detail later, there are disagreements both over whether the resources were lacked and over who lacked them.

¹⁸ Fricker argues that Wood's harasser is also subject to a lack of intelligibility and hence suffers epistemic harm, but in contrast to Wood herself, the harm suffered by Wood's harasser is not of immediate, individual importance to him. So, whereas Wood's harasser suffers a hermeneutical harm by being cognitively disadvantaged by a lack of conceptual resources, Wood herself suffers a hermeneutical *wrong* because she is unable to render intelligible an experience that it is uniquely important for her as an individual to understand (Fricker, 2007, p. 151).

systematic marginalization of others who have had the same kind of experience (Fricker, 2007).

Like Marge, Wood suffers an epistemic injustice in virtue of the fact that i) she is epistemically harmed, ii) that epistemic harm is wrongful, and iii) that wrong comes about as a result of underlying prejudices. In contrast with Marge, Wood suffers a *'hermeneutical injustice'*; a variety of epistemic injustice in which a gap in the conceptual resources of an epistemic community causes harm to an epistemic agent (Fricker, 2007, p. 155).

With these two basic cases on the table, I will now go on to present my three novel cases. My aim in the rest of this section will be to show that these cases are potential cases of epistemic injustice. In section 2, I will turn my attention more towards demonstrating how my cases differ from the extant cases.

1.3 Heart Attacks in Women: Alison Fillingham

Alison Fillingham, a nurse with 24 years of experience, was 49 years old when she began experiencing pain in her collarbone ("Misdiagnosis of heart attacks," 2019). Thinking that the pain was merely the result of a hectic few days, she continued her life as usual with the hope that the pain would disappear quickly. Two days later, the pain had spread to her jaw and intensified substantially and a friend finally convinced Fillingham to call an ambulance. Upon arrival, the ambulance crew informed her that she was merely having a panic attack, and advised that she be taken to the hospital without urgency. Hours later, she was seen by a doctor and diagnosed as having had a heart attack ("Misdiagnosis of heart attacks," 2019).

Both Fillingham and the paramedics that assess her fail to see that she is having a heart attack. Because of this, Fillingham suffers an *epistemic wrong* akin to that suffered by Carmita Wood; she is unable to fully conceptualize her experience and as such is denied epistemic recourse to others regarding this experience. This wrong is also *unjust*. There is extremely good reason to believe that Fillingham's being a woman plays a big role in the failure, both on her part and on the part of medical professionals who treat her, to correctly diagnose her condition. It is well-documented that women suffering heart attacks are frequently misdiagnosed by medical professionals (Mikhail, 2005; "Misdiagnosis of heart attacks," 2019; Publishing, n.d.). Yet our collective understanding of heart attacks, their symptoms, their signs etc. is, in the abstract, very good and we have a high success-rate of diagnosing them in men (Publishing, n.d.). Indeed, research conducted out of the University of Leeds shows that women are 59 percent more likely than men to receive a misdiagnosis (Alabas et al., 2017) and there is an increasing literature which studies the influence that gender has on these misdiagnoses.¹⁹ This gives evidence that the epistemic wrong suffered by Fillingham is associated with latent prejudices and injustice in the practice and production of research, treatment, and diagnosis.

1.4 Endometriosis: Lindsay Murphy

¹⁹ The study used the SWEDEHEART registry to assess 180,368 patients who suffered acute myocardial infarction or STEMI between 2003 and 2013. STEMI occurs when there's a total blockage of a coronary artery and is contrasted with Non-ST-elevation myocardial infarction or NSTEMI.

Lindsay Murphy, a 36-year-old from San Francisco, had been experiencing intense, debilitating pain in her abdomen, usually accompanying her periods (Scott, 2015). She had seen numerous doctors about the pain and invariably been sent away with a presumptive diagnosis of ovarian cysts (though no cysts could be found) or with ‘non-pathological’ period pain. Her symptoms gradually expanded and worsened until, in August of 2011, she fainted attempting to load the washing machine. She was rushed into hospital with dangerously low oxygen levels and underwent nine hours of surgery. As a result, she was diagnosed with endometriosis: a condition in which tissue similar to the uterine-lining develops outside of the uterus. Typical symptoms of endometriosis include severe abdominal pain and pelvic cramping especially around the time of one’s period. In severe cases like Murphy’s, the endometrial lesions can extend beyond the abdominal cavity up to the lungs, throat, urethra, bladder, kidney, liver, and sciatic nerve (Scott, 2015).

Lindsay Murphy suffers an *epistemic wrong* akin to that suffered by Carmita Wood; in failing to be diagnosed appropriately she is prevented from understanding a portion of her life that it is strongly in her interests to understand. This both robs her of important self-knowledge and prevents her from having epistemic recourse to others. Like Wood, Murphy also enjoys an epistemic relief when the correct diagnosis of endometriosis is finally handed to her. Ballard, Lowton, and Wright report:

Women benefited from a diagnosis, because it provided a language in which to discuss their condition, offered possible management strategies to control symptoms, and provided reassurance that symptoms were not due to cancer. Diagnosis also sanctioned women’s access to social support and legitimized

absences from social and work obligations. (Ballard, Lowton, & Wright, 2006, p. 1296)

This wrong is also *unjust*. There is extremely good reason to believe that Murphy's being a woman plays a big role in the failure, both on her part and on the part of medical professionals who treat her, to correctly diagnose her condition. It takes an average of 8.7 years to be correctly diagnosed with endometriosis (Barbieri, 2017) and misdiagnoses are extremely common, despite the fact that approximately 1 in 10 women have the condition (Scott, 2015). The Endometriosis Foundation of America estimates that, worldwide, more women suffer from endometriosis than suffer from type 1 diabetes, multiple sclerosis, lupus, ALS, heart disease, breast cancer, and ovarian cancer *combined* ("Endometriosis Foundation of America," 2019).

Though the symptoms of endometriosis can be obscure and overlap with other conditions, the evidence suggests that this overlap is not enough to account for the typical delay in diagnosis. Ballard et al. suggest that diagnosis is often delayed due to stereotyping and stigma on the part of the doctor, on the part of the patient, or on the part of family and friends (Ballard et al., 2006; Barbieri, 2017; Scott, 2015; Seear, 2009). They suggest that doctors often assume (implicitly or explicitly) that pain associated with the menstrual cycle is normal and trivial. For example, a study that interviewed women with endometriosis reports that their 'doctors trivialized or normalized their pain complaints by suggesting that menstrual pain was a normal and non-pathological process.' (Seear, 2009, p. 1221). The evidence also indicates that it is extremely difficult to debunk the presumption that pain associated with the menstrual cycle is normal; even

when women with endometriosis asked explicitly whether their symptoms could be due to endometriosis many clinicians said ‘no’ (Barbieri, 2017, p. 8).

This lends credence to the thought that the wrong suffered by Murphy comes about due to harmful, prejudicial stereotypes that make both doctors and patients more likely to disregard the symptoms of endometriosis or attribute them to non-pathological ‘period pain.’ Despite the fact that both the symptoms and the prevalence of endometriosis is known to medical professionals, Murphy is consistently misdiagnosed. This might reasonably be traced to the normalization of abdominal pain in women and due to the trivializing of pain reports (Scott, 2015; Seear, 2009) as well as sexist practices in the medical world and the world of research (Bell, 2016; Beusman & Stoner, 2018; Hoffman & Tarzian, 2001; Norman, 2018; Pasha-Robinson, 2017; Scott, 2015).

1.5 ‘High-Functioning’ Bipolar Disorder: M²⁰

M is an individual who has bipolar disorder. Before being diagnosed, M had spent years searching for a diagnosis, undergoing trials of antidepressants and benzodiazepines, and being repeatedly misdiagnosed (“Invalidating Bipolar Disorder,” 2019). During this time, M suffers an *epistemic wrong* akin to that suffered by Carmita Wood; M is prevented from rendering intelligible an experience which it was strongly in their interests to understand and as a result is denied epistemic recourse to others. This fact becomes especially salient when M talks about the relief (not only medical but epistemic too) of finally receiving the diagnosis they needed:

²⁰ ‘M’ refers to an anonymous individual who authored an article on the webpage for the National Institute of Mental Illness (“Invalidating Bipolar Disorder,” 2019).

After years of being misdiagnosed and going on and off antidepressants, I was finally given the diagnosis of bipolar disorder. That was a huge breakthrough for me. It made the way I felt and the severe mood swings I would experience feel validated. There was a reason. I now had words to explain what I was going through: mania, depression, hypomania. (“Invalidating Bipolar Disorder,” 2019)

Once diagnosed, things improved significantly; M’s feelings were validated, they were able to find medication to help temper their symptoms, and they were able to make better sense of their experiences. However, M reports that the epistemic relief they experienced was stultified by a continued skepticism from their community. Even after diagnosis, M continued to be prevented from having epistemic recourse to others:

I found that the validation I felt, or the acceptance of this diagnosis, was not felt by everyone. There are many reasons for this: lack of knowledge, bias, misconceptions, etc. Below are some of the responses I have received after telling people about my bipolar diagnosis. Some people have been supportive, some well-intentioned, others ignorant, or just plain hurtful. A few of the responses I have received are listed below.

“You don’t have bipolar.” “You seem normal.” “You don’t seem crazy.”

I’m not crazy. I have a mental illness. I don’t announce it to the world when I can’t get out of bed for 48+ hours or that the reason I have recently taken up so

many hobbies or work so many hours is actually one of the many, many symptoms of a manic episode.

“You didn’t seem like you had bipolar until you were diagnosed.”

This one hurts a lot. I have finally, for once in my life, had my feelings and emotions validated. I understand better why I am the way I am, and for the first time, I can actually work towards a proper plan to treat it, or minimize it. I was also very good at hiding it most of the time. This response completely crushes that feeling. (“Invalidating Bipolar Disorder,” 2019)

The epistemic wrong suffered by M is also *unjust*. There is extremely good reason to believe that M’s experience (both of struggling to get a diagnosis, and of widespread skepticism once a diagnosis is achieved) is a result of prejudice regarding the lived experience of people with bipolar disorder. It takes on average 5-10 years to be diagnosed with bipolar I (Cha, Kim, Ha, Chang, & Ha, 2009; Ghaemi, Boiman, & Goodwin, 2000; Hirschfeld, Lewis, & Vornik, 2003), and bipolar II is expected to take much longer (Cha et al., 2009, p. 99).²¹ It is not uncommon for bipolar disorder to go unrecognized, both in the medical community and in the wider social community, due to misconceptions about

²¹ Bipolar I and Bipolar II are differentiated by the severity of the manic episode suffered. Bipolar I is characterized by manic episodes, either with or without depressive episodes. Bipolar II is characterized by having at least one depressive episode lasting two weeks or more and at least one period of hypomania. A manic episode is similar to but more extreme than a hypomanic episode (Cha, Kim, Ha, Chang, & Ha, 2009; Hirschfeld, Lewis, & Vornik, 2003; Roland, 2016).

the ways in which it presents; it is assumed (implicitly or explicitly) to manifest in ways that are publicly very obvious, yet many people with bipolar disorder are ‘high-functioning’ (Fiala, 2004; Yu, 2017, p.); they are doctors, lawyers, accountants, and judges; they meet deadlines and produce creative work; they engage with others and are well-presented.

Though the symptoms of bipolar disorder overlap with other conditions such as unipolar depression, overlapping symptoms cannot fully account for the prevalence of misdiagnosis or the ignorance in the community. Bipolar disorder affects around 5% of the community at large, and around 50% of depressed outpatients (Benazzi, 2007, p. 935). Yet of patients with bipolar disorder, 70% don’t receive diagnosis until they have seen upwards of four physicians (Lohano, Loganathan, Roberts, & Gao, 2010). This might reasonably be traced to the ongoing social ignorance surrounding the ways in which bipolar disorder manifests.

2: Situating the cases

I’d now like to spend some time differentiating my cases (1.3-1.5) from the classic cases of testimonial and hermeneutical injustice (1.1 and 1.2). I’ll begin (2.1) by providing an overview of the notion of testimonial injustice, drawing out the distinctive features of this phenomenon and showing how my cases differ. In 2.2, I’ll look more carefully at the central cases of hermeneutical injustice that have been discussed in the literature and argue that my cases are importantly different in several ways. In section 3, I’ll unify my cases under the heading ‘inferential injustice’ and offer a general description of what goes wrong in these cases.

2.1 Comparing cases of testimonial injustice

While cases 1.3-1.5 are cases of *epistemic* injustice, they are importantly different from the typical cases of *testimonial* injustice. As we have seen, testimonial injustice occurs when prejudice cause a speaker's credibility to be undervalued. In the typical case of testimonial injustice, a protagonist (Marge Sherwood) makes a testimony (that Ripley is sinister) which is not taken seriously by her interlocutor (Herbert Greenleaf) due to prejudicial stereotypes (that women are irrational) (Fricker, 2007, pp. 88–90). What is characteristic about testimonial injustice is that it involves testimony; it arises when one testifies about x, and is credible in their testimony about x, but is not taken seriously in that testimony. Fillingham's case cannot be a case of testimonial injustice because she makes no relevant testimony; she does not claim 'I am having a heart attack' because she doesn't know that she is having a heart attack. She does testify to her *symptoms* and I am open to the idea that testimony may be made without literal words; perhaps one can testify about the pain one is in by making certain facial expressions. But in any case, Fillingham's testimony regarding her pain and her symptoms is taken seriously, it is what's inferred on the basis of those symptoms that is faulty.

What about Murphy's case? Similarly to Fillingham, Murphy doesn't testify that she has endometriosis because she doesn't *know* that she has endometriosis. She *does* testify to her symptoms, and this testimony is undervalued to some extent, because her doctor takes her to be overestimating her pain levels. This is probably a case of testimonial injustice; the sufferer is not taken to be credible in her testimony. However, the case can't be fully explained by testimonial injustice. Even when women are taken seriously with regard to

their pain levels, doctors often misdiagnose them as having especially bad menstrual pain, or as having ovarian cysts. Since these cases are ones where their credibility is *not* undervalued, they cannot be cases of testimonial injustice. Yet I would submit that these cases are still cases of epistemic injustice because the sufferer is epistemically harmed as a result of unable to conceptualize her experience and unable to seek recourse to others just as in Carmita Wood's case.

What of M? When M's friends undervalue M's credibility regarding the bipolar diagnosis, they wrongfully cause M 'testimonial harm,' but they do not perpetrate testimonial *injustice*. Cases of testimonial injustice are ones in which a speaker is taken to not be credible *about a certain matter* due to their perceived membership *in a certain group*. For example, a woman - Delina - who testifies on a matter about which she is exceedingly credible - video games - might suffer epistemic *wrong* if her credibility is undervalued by those listening to her. She suffers testimonial *injustice* insofar as that credibility deficit is caused by the listeners thinking, on the basis of prejudicial stereotypes, that Delina is not *the kind of person* who would know about video games. This is an essential part of testimonial injustice; the credibility deficit must be caused by the listener's assumption that the speaker is not *the kind of person* to know about *the kind of things* that are under discussion. When M testifies to having bipolar disorder M's credibility is undervalued; M is taken to be wrong about their diagnosis when in fact they are not. But it is not the case that M's friends think M is not *the kind of person* who is likely to know about *this kind of thing*. They don't think, due to M's social group, that M isn't a knower with regard to bipolar disorder. After all, what kind of group membership could motivate such a belief? Certainly not membership in the group 'sufferers of bipolar disorder,' because M's friends

don't think that M has bipolar disorder in the first place. Presumably not race or gender since the case is equally familiar regardless of what race or gender we take M to have. Hence, while M's credibility is undervalued, testimonial injustice cannot explain what goes wrong in their case.

2.2 Comparing cases of hermeneutical injustice

Cases 1.3-1.5 involve some of the same epistemic wrongs as occur in the Carmita Wood case. As we saw above, Wood is epistemically harmed because a lacuna or paucity in the collective hermeneutical resource prevents her from rendering her experience fully intelligible either to herself or to her epistemic community. As Fricker presents the case, Wood really suffers two separate epistemic wrongs.²² On the one hand, we are told she is denied self-knowledge about an experience it is strongly in her interests to understand (Fricker, 2007, p. 151). On the other hand, we are told that she is denied recourse to others regarding that same experience (Fricker, 2007, p. 6). Both wrongs are also present in the three cases I give. Prior to their respective diagnoses, Fillingham, Murphy, and M were unable to conceptualize their experiences and unable to have recourse to others regarding their experiences, including being prevented from sharing in an epistemic community which could offer substantial support; they can't speak with other sufferers of

²² It is not entirely clear that Fricker distinguishes between these two harms. She mentions both (Fricker, 2007, pp. 6–8 & 147–169, 2016) but is not explicit about whether i) both are necessary for hermeneutical injustice, ii) the former is necessary and the latter not, iii) the latter is necessary and the former not, or iv) any combination is sufficient. I think that the right way to understand hermeneutical injustice fits option iv) but this is not the right place to discuss that issue. The remainder of the paper will proceed on the assumption that iv) is true, but not much will turn on that assumption.

bipolar disorder or endometriosis (for example), they can't share in the experiences of others, they can't compare difficulties or work through bewildering experiences. Since these elements are what constitute the epistemic wrong in Wood's case, we ought to conclude that Fillingham, Murphy, and M suffer epistemic wrongs as well.

But cases 1.3-1.5 differ from the Carmita Wood case in some important respects. Fricker presents the Carmita Wood case as one in which the concept of sexual harassment was lacking in an important way. To bring this out, she emphasizes the effect that introducing an appropriate concept will have on a sufferer of hermeneutical injustice. On Fricker's reckoning, when the term 'sexual harassment' is introduced into the lexicon, it serves 'not simply [as] a hermeneutical breakthrough for her [Wood] and for the other women present, but also *a moment in which some kind of epistemic injustice is overcome*' (Fricker, 2007, p. 149, emphasis mine). The same is not true for Fillingham, Murphy, and M. They continue to face many of these difficulties even once diagnosed. Having the diagnosis – the concept, the label – does not seem to overcome the epistemic injustice. This is particularly salient in M's case. Those M confides in struggle to believe M, or they play down M's symptoms, the severity of M's condition, or the severity of bipolar disorder as a whole. Fillingham, Murphy, and M all suffer because their communities don't take them to 'count' as members of a group to which they in fact belong *even after diagnosis is achieved*.²³

²³ Wood's case is a little more complicated than this gloss suggests as she may continue to be denied recourse to others if those others are of a different hermeneutical community. For example, if Wood leaves the consciousness-raising group and enters a community into which the notion of sexual harassment has

Additionally, while Wood begins with a lack of self-knowledge, then gains self-knowledge as she and other group-members forge towards a better understanding of the experience that they share, Murphy and M may start with self-knowledge and then lose it.²⁴ As reported above, some women with endometriosis suggest to their doctors that their pain may be due to endometriosis and yet their doctors tell them that this is not the case (Barbieri, 2017, p. 8). These women enter the Doctor's office with the knowledge (or suspicion) that they have endometriosis and leave having lost that knowledge. They may also leave questioning the severity of their symptoms or wondering whether they are exaggerating the pain they suffer, or are failing to deal with a level of pain that is 'normal' (Denny, 2004, p. 644).

M is also in danger of losing self-knowledge. In a recent post called 'I am a good bipolar' (KallemWhitman, 2017), Dr. Rachel KallemWhitman describes how difficult it can be to learn to recognize the signs of a manic or depressive episode. She explains how incredibly subtle and nuanced those signs are, and emphasizes the propensity of others to disregard them: 'They all start out so infuriatingly under the radar. Cloaked in self-sabotage.

not yet permeated then she will continue to suffer epistemic isolation of the sort she suffered before being introduced to the notion herself. Fricker also agrees that this is the case (Fricker, 2016).

²⁴ There is a lot of discussion on this claim from Fricker. Rebecca Mason (Mason, 2011) suggests that Wood begins with self-knowledge and the nature of wrong done to her is the failure of her community to recognize or take seriously her testimony about her experience. Kristie Dotson (Dotson, 2011), Jose Medina (Medina, 2012), and Gaile Pohlhaus (Pohlhaus, 2012) stress that it is overly simplistic and even harmful to regard Wood as entirely lacking in self-knowledge. Ian Werkheiser (Werkheiser, 2014) has argued that one may lose self-knowledge as a result of questioning by others.

Designed to go unnoticed, fatally subtle. Mental illness is an abusive dark shadow' (KalleWhitman, 2017). KalleWhitman has worked extremely hard on learning to diagnose and spot these subtle signs which are often invisible to others. An ability like this takes a lot of work, and is undoubtedly significantly derailed by skepticism from others ("Invalidating Bipolar Disorder," 2019; KalleWhitman, 2017). From this we can see very clearly the ways in which the more subtle symptoms of bipolar can be trivialized by others. One effect of constant skepticism of this kind is that one begins to doubt their initial interpretation; perhaps my symptoms are not as serious as I anticipated, perhaps I don't have 'real' or 'proper' bipolar (Fiala, 2004), perhaps the indicators I'm seeing are not indicators but just normal feelings ("Invalidating Bipolar Disorder," 2019). In support of this, Ian Werkheiser's recent work suggests that asking someone to provide reasons in support of a claim often reflects oppressive and prejudiced assumptions which cause substantial harm and loss of knowledge (Werkheiser, 2014).

2.3 On Hermeneutical Paucity

Characteristic of hermeneutical injustice, Fricker tells us, is that there be a 'paucity of concepts' (Fricker, 2016, p. 171) which gives rise to harms of the type Carmita Wood suffers. Carmita Wood, in Fricker's version of the case, suffers because the concept she needs isn't there when she needs it; she literally lacks name or notion that she can apply to this experience that will allow her to make sense of it in her own mind and talk with others about it. As Fricker describes it, this becomes particularly salient when the notion of sexual harassment is finally developed and light is finally thrown on Wood's experience.

What exactly this ‘paucity’ consists in, however, is not entirely clear. Fricker sometimes describes it as a ‘conceptual paucity’ (Fricker, 2016, p. 171), but elsewhere describes it as involving a more general paucity in the ‘collective hermeneutical resources’ (Fricker, 2007, pp. 151–152, 2016, p. 161). A natural reading of the Carmita Wood case is that the conceptual paucity at issue was a literal lack of existence; the concept that she needed literally failed to exist. But, as has been discussed in the literature (Dotson, 2011; Fricker, 2016; Mason, 2011; Medina, 2012; Pohlhaus, 2012) this may not be the most accurate account of what was going on in the Carmita Wood case and it is certainly not the only way that case could possibly have played out. For example, imagine a contemporary of Carmita Wood’s, call her ‘Alternate-Wood’, who suffered the same experience as Wood but did not have the good fortune of attending the same consciousness-raising group. Alternate-Wood continues to suffer, unable to conceptualize her experience, even after Wood and her discussion group introduce the notion of sexual harassment. As far as Alternate-Wood is concerned, nothing has changed and yet, after the consciousness-raising group meets, it is no longer true that the requisite concept doesn’t exist.

As this example makes clear, it seems as though *access* to the requisite concept is more important than whether the concept exists; even if a concept strictly speaking exists one may still suffer hermeneutical injustice if one is unable to *access* that concept. That said, given the relational nature of knowledge, even having *access* to the requisite concepts won’t necessarily be enough to avoid hermeneutical harm; one also needs to achieve conceptual ‘match’ with the relevant epistemic community. For example, Wood may gain access to the concept of ‘sexual harassment’ but if the relevant epistemic community continues to lack that concept, she will continue to be harmed by an inability to have her

experience collectively understood. This is because whether I really have access to some concept depends on where I am and what I am doing. For example, the conceptual resources available to me when talking with family members about how to deal with the obstructive next-door-neighbor are very different from those available to me when discussing modality in my graduate seminar. Indeed, part of becoming a capable teacher involves working out what hermeneutical resources you and your students share in common, and operating within those bounds. This brings out the important fact that, as epistemic agents, each of us is a member of numerous distinct pools of hermeneutical resources (Dotson, 2011, pp. 31–32; Mason, 2011, pp. 300–306; Medina, 2012, pp. 103–107; Pohlhaus, 2012, pp. 715–717 esp.). As such, concepts do not exist or fail to exist simpliciter, nor are they accessible or inaccessible simpliciter. Rather, a concept’s existence and accessibility is relative to an epistemic community. This has the result that an epistemic agent may in fact possess the conceptual resources necessary to make sense of her own experience (and hence not suffer any lack of self-knowledge), and yet still be hermeneutically stultified because the dominant epistemic community lacks said resources.²⁵

This leads me to believe that it is not a requirement of hermeneutical injustice that an agent lack self-knowledge; it is enough that the agent is unable to share their experience with the relevant communities. In the face of epistemic oppression, epistemic counter-communities may spring up in which there are conceptual resources apt for explaining the experiences of marginalized individuals. This idea is already present in the literature; it

²⁵ Rebecca Mason, in fact, argues that Fricker’s central case of hermeneutical injustice - the case of Carmita Wood - is one such case (Mason, 2011, p. 297).

is present in Pohlhaus' account of 'willful hermeneutical ignorance' (Pohlhaus, 2012, p. 722), in Dotson's analysis of 'contributory injustice' (Dotson, 2011, pp. 31–35) and in Mason and Medina's claims that Fricker's account of hermeneutical injustice itself perpetrates hermeneutical injustice (Mason, 2011; Medina, 2012).

I think that our hermeneutical resources can be harmful in a further way, namely in their *inferential structure*. In the remainder of the paper, I will argue for a variety of epistemic injustice that I call 'inferential injustice.' This is a species of epistemic injustice that arises from prejudice in the inferential structure of our hermeneutical resources. The particular cases I have presented in this paper are symptomatic of a more general problem and help to bring out the nature of inferential injustice.

I should note here that there is both a broad and a narrow reading of Fricker's notion of hermeneutical paucity.²⁶ The narrow reading takes conceptual resources to refer narrowly to concepts. A more broad reading would take 'conceptual/hermeneutical resources' to capture any and all of the components in our conceptual machinery, including things like inferential patterns and access relations. Either reading is perfectly reasonable.²⁷ I take it, however, that the difference is rather terminological in this context. If one wishes to adopt the former, more narrow interpretation of Fricker (2007), then one should take inferential injustice to be a species of epistemic injustice which is distinct from testimonial and hermeneutical injustice. If one adopts the latter, more broad conception of

²⁶ Thanks to Amy Flowerree and Elizabeth Anderson for encouraging me to bring this out more clearly.

²⁷ Arguably, the narrower conception is employed by Pohlhaus (Pohlhaus, 2012) and Dotson (Dotson, 2011), while the broader interpretation is employed by Goetze (Goetze, 2018) and Fricker (Fricker, 2016).

hermeneutical injustice, however, then one should take inferential injustice to be a species of *hermeneutical injustice*. I'll proceed in the spirit of the narrower reading, though I maintain that inferential injustice remains important either way for three main reasons.

The first reason is that inferential injustice requires a different solution from the classic cases of hermeneutical injustice. Whichever interpretation one adopts, it remains the case that the right way to respond in Wood's case was to develop new concepts and new ways of discussing her experience; new resources needed to be introduced. In cases 1.3-1.5 this is not the case. In fact, in those cases it is pivotal that we *not* introduce new concepts; what needs to happen is that Fillingham (e.g.) is seen as having a *heart attack*, the *same* condition that many hundreds of men have. It would be exactly the wrong thing to do to introduce a new notion to account for Fillingham's experience. Similarly, it is important that M be seen as having *bipolar disorder*, and that M not be attributed some new, different condition. For these cases, what they need is to be inducted into a group that *already exists*.

The second reason that inferential injustice remains important even if we adopt the broad reading of 'hermeneutical paucity' is that it highlights *particular locations* where prejudice can gather; namely, in the inferential relationships we employ when interacting with the world. The harm suffered by Wood has the same flavor as that suffered by Murphy, Fillingham, and M, but a different cause. While all are prevented from rendering their experiences fully intelligible, in Wood's case there simply isn't an accessible concept which accurately represents Wood's experience. As such, one can look for varieties of experience which don't seem amenable to any of our current hermeneutical resources. But in the case of Fillingham, Murphy, and M, it appears as though we have just the

concepts necessary to understand their experiences; Murphy suffers from normal period pain, Fillingham is having a panic attack, M is feeling sad. Their experiences are overridden in the kind of way that Dotson (following Audrey Lorde (Lorde, 1990)) describes as ‘scrambling or distorting’ (Dotson, 2011, p. 33). The issue is that by continuing to focus on the generation of concepts we will miss important cases like these which seem to include all the right concepts . M is not harmed because there is some concept we do not have, rather she is harmed because *the concepts that we do have* are taken to relate in inappropriate ways.

Finally, inferential injustice remains important even if we adopt the broad reading of ‘hermeneutical paucity’ because it exposes the looping quality of cases 1.3-1.5. Our inferential practices are not only influenced by abstract and detached theories; they respond to the feedback we get when we use them. As such, an inference that appears to be confirmed by the world will become more entrenched. This is akin to Quine’s web of belief (Quine & Ullian, 1978) but applied to inferential practices, and to Hume’s notion of custom and habit (Hume, 2007, pp. 43–46). As inferential patterns become more widely adopted and more reliable, they become harder to shift; like tracks in the mud, the more you traverse them the harder they become to escape. For example, M will encounter significant resistance the next time they attempt to talk to their friends about Bipolar.²⁸ What of the patient with endometriosis who has suggested to her doctor that she might have the disease but has then been told that she doesn’t? If she suggests it again, it will be even harder for her to get that idea off the ground; her doctor will have an implicit backdrop which guards against her claim.

²⁸ Thanks to Alicia Patterson for this example.

3: Injustice in the Spaces Between Concepts: Inferential Injustice

The cases I have presented in this paper bring out a way in which hermeneutical resources can be oppressively impoverished which has not yet been discussed in the literature. In our quest to uncover epistemic injustice, we must become more aware of the ways in which the framework of our epistemic resources might be lacking, harmful, and unjust. That is the moral that I take to be brought out by cases 1.3-1.5, and that I hope to capture in my notion of inferential injustice. If I am successful, you will see that epistemic injustice can occur explicitly as a result of the structural, inferential practices that we employ as part of the framework or backdrop of our epistemic lives.

Work in the philosophy of science has shown us that i) given any empirical evidence, there will always be an infinite number of theories that equally well fit that evidence (Duhem, 1954, p. 189) and ii) only ‘whole sciences’²⁹ - and not lone hypotheses - can furnish us with testable predictions (Duhem, 1954, pp. 185–187; Quine, 1963, p. 42). For this reason, we must make use of what is sometimes called ‘extra-experiential criteria’ to allow us to tell between empirically equivalent theories. Extra-experimental criteria allow us to narrow down the field of infinitely many empirically equivalent theories to a manageable number of theories that can be tested against one another. In essence, they let us kick out all of the outlandish theories and focus on the theories that are more likely to be true. How do we decide which theories to kick out and which to keep? On the basis

²⁹ A ‘whole science’ in this sense is a huge collection of posits consisting of the hypothesis under consideration *plus* all the background assumptions that support that hypothesis (Quine, 1963).

of ‘projectability judgments’ (Goodman, 1983); judgments guided by what we think a good theory looks like in the context under scrutiny.

In order for projectability judgments to be at all truth-tracking - in order for them to help us traverse the world - they must be collective, relational, and have some cross-temporal stability. This reflects the Kuhnian idea that sciences operate within *paradigms* (Kuhn, 1996; Lakatos, 1976; Laudan, 1980). A paradigm is a sort of structure or framework that forms the backdrop of our epistemic practice. Essentially it offers us rules which dictate and constrain the extra-experimental criteria mentioned above. It guides us with regard to; what counts as evidence; what counts as *good* evidence; how to identify and appraise evidence; what information is relevant or irrelevant; when to disregard evidence; what a good argument or inference looks like; when a hypothesis can be considered confirmed; etcetera. It is helpful, I think, to regard the framework as its own structure, that exists somewhat independently as part of our epistemic resources. In this way, we can extract, to a certain extent, the framework itself from the ‘contents’ of that framework.

3.1 Prejudice in the Inferential Structure of Hermeneutical Communities

I have argued that our epistemic resources have a structured backdrop or framework that can be thought of as, to a certain extent, separate³⁰ from the ‘contents’ of that framework.

³⁰ I’m speaking metaphorically here. What I call the ‘framework’ and what I call the ‘contents’ doesn’t map perfectly onto any two categories, but is a general approximation of how we might think about the inferential relationships (on the one hand) and the concepts (on the other). That said, I wouldn’t want to say that the framework and the contents, or the inferential relationships and the concepts are completely separable.

If that is true, then we should expect that the framework itself (as well as its contents) can be unjust and/or wrongful. This is what makes for inferential injustice. It is a very *structural* variety of injustice; it doesn't originate in any individual action, but comes about due to the overall structure of our epistemic resources. It is this *structure* that is wrongful and unjust in cases of inferential injustice.

Some previous criticisms of Fricker have likewise argued that her account fails to fully appreciate the extent to which our epistemic resources are structured, relational, and shared (Dotson, 2011; Mason, 2011; Medina, 2012; Pohlhaus, 2012, 2014). A particular focus for these criticisms has been the fact that, for example, my concept of x depends in many ways on what *your* concept of x is, as well as on what you do with that concept, and on whether you recognize us as sharing a concept. These kinds of concerns give rise to Gaile Pohlhaus' notion of 'willful hermeneutical ignorance' which occurs when 'marginally situated knowers actively resist epistemic domination through interaction with other resistant knowers, while dominantly situated knowers nonetheless continue to misunderstand and misinterpret the world' (Pohlhaus, 2012, p. 717). Pohlhaus' notion is predicated on the fact that knowing is a group activity; introducing the notion of sexual harassment is not enough to overcome the wrong suffered by Carmita Wood because the social world must also 'accept' that notion and induct it into the generally accepted stock of resources. Kristie Dotson's notion of 'uptake' - the act of recognizing epistemic resources as such - also reflects this concern (Dotson, 2011). Her notion of 'contributory injustice' accounts for a kind of epistemic injustice which one perpetrates by 'maintaining and utilizing structurally prejudiced hermeneutical resources that result in epistemic harm to the epistemic agency of a knower' (Dotson, 2011, p. 31).

There is a further, previously overlooked, way in which our epistemic resources are relational: in the structural rules that guide our projectability judgments. Though the term is technical, projectability judgments themselves are extremely familiar to us all. When assessing a situation, for example making a diagnosis, we are presented with information. We must process that information and differentiate the relevant from the irrelevant, the more useful from the less useful, and identify which inferential patterns are open to us in the context at hand. This is a somewhat holistic process. For example, I might have a pretty good understanding of what a Black-Capped Chickadee is. Perhaps I can even provide you with a rough-and-ready list of characteristics to look for. Say I now go out into the world and encounter a bird. I have to determine what species this bird is. In order to do this I do not simply run through a checklist of characteristics, rather I must draw on an entire network of ideas and concepts. Even if I have access to a list of characteristics that is unusually comprehensive, I at least need to determine whether, for example, the black markings that I observe are really black and not just cast in shadow, whether the bird really is small or whether it just looks so from this angle, and whether I am in an area where the very similar-looking Carolina Chickadee



Carolina Chickadee
(*Poecile carolinensis*)



Black-Capped Chickadee
(*Poecile atricapillus*)

Fig. 2

can also be found (fig. 2). The notion of projectability reflects the fact that these assessments necessarily involve judgments (*projectability* judgments) about what a reasonable, likely-to-be-true theory looks like. Hence, whether an agent arrives at one conclusion of another depends on the which factors that agent takes to be operative in determining which theory a set of data supports.

Unsurprisingly, the process of arriving at projectability judgments can go awry in a number of ways. My claim is that misguided projectability judgments can sometimes lead to epistemic injustice. To be clear, misguided projectability judgments can derail investigation and cause epistemic harm in lots of ways that aren't necessarily unjust. Where they become unjust is when they are motivated by background theories that are prejudiced. This happens often both within and outside of science. Take the well-worn case of the race-IQ controversy from the 1960s. In that case, research produced by well-regarded scientists and published in well-regarded science journals seemed to indicate that race partly determined intelligence. This is false, of course, but its falsehood managed to slip past many intelligent and interrogative minds. Though counterevidence was readily available, that evidence was unnoticed by many because it wasn't as projectible as the (prejudiced and false) hypothesis that African Americans *are biologically determined to have a lower IQ than whites*.³¹

Here's a less well-known example: As recently as 1960 there have been scientific studies, published in respectable journals by very able scientists that posit a relationship between the menstrual cycle and things like airplane accidents (Dalton, 1960b), criminal activity

³¹ For more on this, see (N. Block, 1995; N. J. Block & Dworkin, 1976; Gould & Gould, 1996).

(Dalton, 1961), deterioration in schoolgirls' work (Dalton, 1960a), and psychiatric illness (Dalton, 1959). We might well ask what these studies were taken to conclude. The answer is problematic³² but even before we ask that question we should ask why these studies went forward in the first place. Studies like this one are designed to test hypotheses. In order for a researcher to invest the time, energy, and money into conducting a scientific study, there must be some hypothesis on the table which they take to be projectible enough to warrant this kind of work. In the case under discussion, the hypothesis appears to be something like the following:

H: The menstrual cycle has an effect on crime, accidents, schoolwork, ...
etcetera.

But why think that H is *projectible*. In the opening to a paper on menstruation and accidents, Katherina Dalton tells us:

Whitehead (1934) observed that in some air accidents involving women air pilots, in which the cause of the accident could not be found, the pilots were in the

³² For those of you who are just too curious, here's an excerpt from one of Dalton's papers:

It has been shown that during menstruation a deterioration occurs both in a schoolgirl's work and in her behaviour (Dalton, 1960a. 1960c), and it is also at this time that women are most liable to be involved in accidents (Dalton, 1960b) or to be admitted to hospital with an acute illness (Dalton, 1959). This gradual psychiatric recognition of the social significance of menstruation in the various aspects of a woman's life has led to an investigation of the menstrual factor importance in crime.

(Dalton, 1961, p. 1752).

menstrual phase of the cycle. Although there have been many studies in other fields of accident proneness, little attention appears to have been given to the part played by menstruation. During work on the diseases of menstruation it was observed that premenstrual lethargy and irritability result in slow reaction time and loss of judgment. As both judgment and reaction time are important factors in the avoidance of accidents, it was felt that a study of women involved in accidents might reveal a correlation between the accident and menstruation. (Dalton, 1960b, p. 1425)

This paints a pretty clear picture of what factors *Dalton takes to be operative* in her projectability judgment. Namely:

1. The belief that ‘premenstrual lethargy and irritability result in slow reaction time and loss of judgment’
2. The data from Whitehead that, within his study, *some* of the air accidents that both involved women air pilots *and* had unidentified causes were ones in which the pilots were ‘in the menstrual phase of the cycle’
3. The belief that both judgment and reaction time are important factors in the avoidance of accidents.

But these factors do not make H projectible unless we also assume a bunch of other stuff such as; that the ‘data’ recorded in 2 is both relevant and significant; that women truly do experience increased lethargy and irritability during their period; that judgment and reaction time in the context of piloting a plane are meaningfully related to lethargy and

irritability; etcetera. In the background of this literature is the inclination to believe old stereotypes about women and the menstrual cycle such as that women behave irrationally, emotionally, and even dangerously during certain stages of the menstrual cycle. Take the following concluding remark, also from Dalton:

One wonders if awareness by women that they are more accident prone at certain times of the month will enable them to use greater care in avoiding accidents, or will it merely induce a menstrual neurosis? (Dalton, 1960b, p. 1426)

The case presented here is an example of how inferential practices and projectability judgments can become infected with stereotypes and prejudices. This is exemplified in the fact that Dalton takes the most the most projectible explanation for the data to be H.

3.2 Inferential Injustice

We need a term to be able to talk about cases where our inferential practices themselves are harmful, wrongful, and unjust. I suggest that we use the term ‘inferential injustice’ for this phenomenon. In inferential injustice, biases, prejudices, and stereotypes operate as part of the projectability judgments made in the community. Hence, while concepts may be problematic in themselves, the relational webs in which those concepts are situated can be just as problematic. The type of cases I bring attention to here are prompted by injustice in the hermeneutical resource itself. In this way, they are properly and wholly structural in their nature. That being the case, it is most accurate to say that the ‘wrongdoer’ in these cases is the structure or the epistemic resources themselves. Rather than there being an individual who acts in such a way as to wrong M, M is wronged by

the epistemic set up of our collective lives. Recall M and their doctor, who is reticent to diagnose M as bipolar. In this case, rather than there being an individual who acts in such a way as to wrong M, M is wronged by the epistemic set up of *our collective lives*. If it seems to you that the Doctor does everything right, then this is probably why. As far as the internal rules of the framework go, the Doctor is using the framework correctly; it's somewhat inadequate to say that M's Doctor individually did something wrong because that obscures the real source of the problem, which is the framework that both M's Doctor and the rest of us operate against.³³

Inferential injustice offers an account of what goes wrong in cases 1.3-1.5. Take Allison Fillingham to start: Fillingham suffers epistemic harm because her community (herself included) fails to recognize her experience as being that of a heart attack. This isn't because we don't know what heart attacks are or even because our understanding of heart attacks is in its infancy. On the contrary, we know quite a lot about heart attacks, what they are, what causes them, and what they look like and we are generally quite good at identifying heart attacks in men. So, clearly, the failure of diagnosis does not result from an accidental lack of knowledge, as it might do in the case of some very rare, poorly understood disease. What's more, Fillingham displays the symptoms of a heart attack in a fairly straightforward way; had those symptoms been displayed by a man, we could expect them to have been diagnosed much earlier (Maas & Appleman, 2010). The problem is that her community (again, herself included) is distracted by other factors, such as the fact

³³ I should be clear that I am not suggesting that a single framework must necessarily be employed by everyone. There will be a number of different epistemic frameworks just as there are a number of differing hermeneutical communities.

that she is a woman. It is historical injustices which lead paramedics to fail to see a heart attack when it is in front of them.

My claim is that the wrong suffered by Fillingham is in part a result of prejudice in prevailing projectability judgments. The symptoms Fillingham experiences (the data) fit the diagnosis (the hypothesis) 'Alison Fillingham is having a heart attack.' Indeed, in some contexts - say in the case of Fillingham's contemporary, Mark - the same symptoms would be taken as evidence for the equivalent diagnosis: 'Mark is having a heart attack.' In Fillingham's case, however, those same symptoms are taken as evidence for a different diagnosis; that of a panic attack. Why is that? Not because there is some symptom that is consistent with a panic attack but not consistent with a heart attack (in fact, necessarily this must not be the case since any symptom Fillingham has is consistent with her having a heart attack); not because the paramedics who tend her simply fail to notice major symptoms (they are aware of and log all the relevant symptoms that may have led them to a diagnosis of a heart attack); and not because the paramedics who tend her just radically misunderstand what a heart attack is - they know (as does Fillingham) that the symptoms she is displaying are consistent with a heart attack. Rather, the problem seems to be a problem in inference. Upon being presented with the data (her symptoms) the hypothesis that she is having a heart attack is not among the most projectible hypotheses.

This demonstrates that one can possess and understand a concept and yet fail to 'see' it when encountered with it. Consider this quote from Dr. Chris Gale, Associate Professor of Cardiovascular Health Sciences and Honorary Consultant Cardiologist at the University of Leeds:

We need to work harder to shift the perception that heart attacks only affect a certain type of person. Typically, when we think of a person with a heart attack, we envisage a middle aged man (*sic.*) who is overweight, has diabetes and smokes. This is not always the case; heart attacks affect the wider spectrum of the population – including women. (“Misdiagnosis of heart attacks,” 2019)

Gale’s suggestion is not so much that we fail to understand the symptoms of heart attacks in women, but that other factors make that diagnosis less projectible. My suggestion is that prejudice plays a role in lowering the projectability in cases like Alison’s.

Earlier, I said that a good first approximation of the idea of epistemic injustice is to characterize it in terms of three components; a harm component, a wrong component, and an injustice component. An approximation of Fillingham’s case, on this model, looks like this:

(Harm component) Fillingham is not counted (by herself or her doctors) as having had a heart attack and hence i) she is unable to make sense of her experiences and ii) she is unable to seek epistemic recourse to others with regard to these experiences.

(Wrong Component) That she is not counted as such is harmful for Fillingham in particular because she has a great personal need to understand and share her experience.

(Injustice Component) That she fails to be counted is the result of faulty projectability judgments fueled by prejudicial assumptions about what a heart attack looks like

Murphy's case is similar. Her symptoms fit the symptoms of endometriosis but they also fit the symptoms of a number of other conditions. In her case, as in a great many others, her doctors diagnose her (incorrectly) with those other conditions. Why? Because, given certain background assumptions, the diagnosis of ovarian cysts or simple 'non-pathological' menstrual cramping is more projectible than the diagnosis of endometriosis. The problem for Murphy is that, in general, prejudices affect how (comparatively) likely her doctors take the endometriosis diagnosis to be. That is, the concept³⁴ 'sufferer of endometriosis' is trumped by the concept 'sufferer of menstrual cramping' because the latter is so ready to hand.

Finally, M's case shares the same characteristics. M talks about their diagnosis with friends and colleagues who respond with disbelief. Why? Because to them the explanation of 'just having a bad day' or 'needing to relax more' or 'reacting normally to stressful life situations' is more projectible than the explanation that M has bipolar disorder despite the

³⁴ Here I am working with the understanding of 'concept' that I have seen used in the literature generally. It is controversial what the right metaphysics of concepts is, and a full discussion of that point would take us too far afield here. But I take it that it is safe enough to assume a loose notion of concept on which you and I can share a concept of 'pop music' without either using it is all the same inferences or assigning to it the same extension. This is the notion of concept we use when we spend a semester-long seminar on the project of trying to refine the notion of 'naturalism.'

fact that they have extremely good evidence for her claim and is an extremely credible testifier. Indeed, Dr. Suzanne Fiala reports similar experiences despite being a practicing physician and hence, presumably, very credible (Fiala, 2004). This is due to prejudicial stereotypes about the experience of bipolar disorder; M's experience doesn't fit with our stereotype of the experience of bipolar disorder. All considered, M's friends take it as more likely (given M's presentation plus their background assumptions) that M's doctor is mistaken than that M has bipolar disorder.

In each of these cases, biases, prejudices, and stereotypes operate as part of the projectability judgments made in the community. Hence, while concepts may be problematic in themselves, the relational webs in which those concepts are situated can be just as problematic.

The type of cases I bring attention to here are prompted by injustice in the hermeneutical resource itself. In this way, they are properly and wholly structural in their nature. That being the case, it is most accurate to say that the 'wrongdoer' in these cases is the structure or the epistemic resources themselves. If it seems to you that the Doctor does everything right, then this is probably why. It's somewhat inadequate to say that M's Doctor did something wrong *as an individual*; as far as the internal rules of the framework go, he is using the framework correctly. And indeed, M is not wronged by the Doctor *as an individual*, M is wronged by the epistemic set up of our collective lives.

What makes these cases of injustice as opposed to mere bad luck? Two things. First is the fact that, operative in the wrongs perpetrated by that epistemic structure, are prejudiced

and harmful assumptions and stereotypes. Second is the fact that the problematic state of the framework is itself caused by the fact that individuals like M are wrongfully not counted as falling under concepts that they in fact fall under. This is somewhat like Fricker's notion of hermeneutical marginalization. projectability judgments are influenced by prejudices, hence prejudicial (stereotypes form part of projectability judgments). Individuals with (e.g.) bipolar disorder are prevented from contributing to the discourse because they are not recognized as members of the group. Prejudice influences the ways in which we make inferences. This leads to people who do have (e.g.) bipolar disorder not being counted as having bipolar disorder. *This* is what marginalizes them: they are unable to contribute to the communal understanding of bipolar disorder because they fail to be 'seen' as sufferers of bipolar disorder.

Recognizing inferential injustice also helps bring to light the looping quality of cases 1.3-1.5. Our inferential practices, frameworks, projectability judgments are not only influenced by abstract and detached theories; they respond to the feedback we get when we use them. As such, an inference that appears to be confirmed by the world will become more entrenched. This is akin to Quine's web of belief (Quine & Ullian, 1978) but applied to inferential practices, and to (Hume, 2007, pp. 43–46) notion of custom and habit. As inferential patterns become more widely adopted and more reliable, they become harder to shift; like tracks in the mud, the more you traverse them the harder they become to escape. For example, M will encounter even greater obstacles the next time they want to talk to their friends about bipolar.³⁵ The woman with endometriosis who suggests to her doctor that she might have the disease will be taken even less

³⁵ Thanks to Alicia Patterson for this example.

seriously upon subsequent visits. Once her suggestion has been rejected by her doctor, it will be even harder for her to get that idea off the ground; her doctor will have an implicit backdrop which guards against her claim.

Conclusion

As classic cases of hermeneutical injustice demonstrate, it's important that we have the right concepts. But it is just as important that the concepts we do have are assigned in the right ways. Cases of hermeneutical injustice call for us to add new conceptual resources to the communal stock; an outright lack of concepts can be amended by introducing conceptual resources that accommodated the experience, object or what have you. But, crucially, this will only work if the new resources are used in the correct ways. Once a conceptual resource exists in the relevant community, there is room for it to be misused, and this is what leads to the kinds of cases I have introduced here.

My aim here was to demonstrate that injustice can be present in the spaces between our concepts. I argued, via the use of some exemplary cases, that epistemic injustice can result from the ways in which we make projectability judgments and from the standard patterns of inference that we require of others and make us of ourselves. If our inferential practices are prejudiced in this way, that constitutes a thoroughly structural way in which our hermeneutical resources might cause epistemic harm.

References

- Alabas, O. A., Gale, C. P., Hall, M., Rutherford, M. J., Szummer, K., Lawesson, S. S., ... Jernberg, T. (2017). Sex Differences in Treatments, Relative Survival, and Excess Mortality Following Acute Myocardial Infarction: National Cohort Study Using the SWEDHEART Registry. *Journal of the American Heart Association*, *6*(12), 1–12. <https://doi.org/10.1161/JAHA.117.007123>
- Ballard, K., Lowton, K., & Wright, J. (2006). What's the delay? A qualitative study of women's experiences of reaching a diagnosis of endometriosis. *Fertility and Sterility*, *86*(5), 1296–1301. <https://doi.org/10.1016/j.fertnstert.2006.04.054>
- Barbieri, R. (2017). Why are there delays in the diagnosis of endometriosis? *OBG Management*, *29*(3), 8–16.
- Bell, T. (2016, February 17). Here's Why Women Get Called "Hysterical" When They're Actually in Pain. Retrieved May 19, 2019, from ATTN website: <https://archive.attn.com/>
- Benazzi, F. (2007). Bipolar Disorder-focus on bipolar II disorder and mixed depression. *The Lancet*, *369*(9565), 935–945.
- Beusman, C., & Stoner, R. (2018, March 7). 1 in 10 Women Have Endometriosis. Why Don't Their Doctors Believe Them? Retrieved May 19, 2019, from Vice website: https://www.vice.com/en_us/article/437ejm/1-in-10-women-have-endometriosis-why-dont-their-doctors-believe-them
- Block, N. (1995). How heritability misleads about race. *Cognition*, *56*(2), 99–128. [https://doi.org/10.1016/0010-0277\(95\)00678-R](https://doi.org/10.1016/0010-0277(95)00678-R)
- Block, N. J., & Dworkin, G. (Eds.). (1976). *The I.Q. controversy: Critical readings*. New York: Pantheon Books.

- Cha, B., Kim, J., Ha, T., Chang, J., & Ha, K. (2009). Polarity of the First Episode and Time to Diagnosis of Bipolar I Disorder. *Psychiatry Investigation*, 6(2), 96.
- Dalton, K. (1959). Menstruation and Acute Psychiatric Illnesses. *BMJ (Clinical Research Ed.)*, 1(5115), 148–149.
- Dalton, K. (1960a). Effect of Menstruation on Schoolgirls' Weekly Work. *BMJ (Clinical Research Ed.)*, 1(5169), 326–328.
- Dalton, K. (1960b). Menstruation and Accidents. *BMJ (Clinical Research Ed.)*, 2(5210), 1425–1426.
- Dalton, K. (1961). Menstruation and Crime. *BMJ (Clinical Research Ed.)*, 2(5269), 1752–1753.
- Denny, E. (2004). Women's experience of endometriosis. *Journal of Advanced Nursing*, 46(6), 641–648. <https://doi.org/10.1111/j.1365-2648.2004.03055.x>
- Dotson, K. (2011). Tracking Epistemic Violence, Tracking Practices of Silencing. *Hypatia*, 26(2), 236–257. <https://doi.org/10.1111/j.1527-2001.2011.01177.x>
- Duhem, P. M. M. (1954). *The aim and structure of physical theory*. Princeton: Princeton University Press.
- Endometriosis Foundation of America: Our Commitment to Research. (2019, April 2). Retrieved June 23, 2019, from Endometriosis : Causes - Symptoms - Diagnosis - and Treatment website: <https://www.endofound.org/research>
- Fiala, S. J. (2004). A piece of my mind. Normal is a place I visit. *JAMA*, 291(24), 2924–2926. <https://doi.org/10.1001/jama.291.24.2924>
- Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford ; New York: Oxford University Press.

- Fricker, M. (2016, December). Epistemic Injustice and the Preservation of Ignorance. <https://doi.org/10.1017/9780511820076.010>
- Ghaemi, S., Boiman, E., & Goodwin, F. (2000). Diagnosing bipolar disorder and the effect of antidepressants: a naturalistic study. *Journal of Clinical Psychiatry*, *61*(10), 804–808.
- Goodman, N. (1983). *Fact, Fiction, and Forecast*. Cambridge: Harvard University Press.
- Gould, S. J., & Gould, T. A. A. P. Z. S. J. (1996). *The Mismeasure of Man*. Retrieved from <https://books.google.com/books?id=WTtTiG4eda0C>
- Highsmith, P. (2008). *The Talented Mr. Ripley*. New York: W. W. Norton and Company.
- Hirschfeld, R., Lewis, L., & Vornik, L. (2003). Perceptions and impact of bipolar disorder: how far have we really come? Results of the national depressive and manic-depressive association 2000 survey of individuals with bipolar disorder. *Journal of Clinical Psychiatry*, *64*(2), 161–174.
- Hoffman, D., & Tarzian, A. (2001). The Girl Who Cried Pain: A Bias against Women in the Treatment of Pain. *The Journal of Law, Medicine & Ethics*, *28*(4_suppl), 13–27.
- How Invalidating My Bipolar Disorder Invalidates Me. (2019). Retrieved May 17, 2019, from National Alliance on Mental Illness website: <https://www.nami.org/Personal-Stories/How-Invalidating-My-Bipolar-Disorder-Invalidates-M>
- Hume, D. (2007). *An Enquiry concerning Human Understanding: And Other Writings* (S. Buckle, Ed.). New York: Cambridge University Press.
- KallemWhitman, R. (2017, April 13). I am a good bipolar. Retrieved May 20, 2019, from Medium: Invisible Illness website: <https://medium.com/invisible-illness/i-am-a-good-bipolar-d056289c3e82>

- Kuhn, T. S. (1996). *The Structure of Scientific Revolutions* (3rd ed.). Chicago: University of Chicago Press.
- Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In S. G. Harding (Ed.), *Can Theories Be Refuted?: essays on the Duhem-Quine Thesis* (pp. 205–260). Dordrecht: D. Reidel Publishing Company.
- Laudan, L. (1980). Progress and Its Problems: Toward a Theory of Scientific Growth. *Erkenntnis*, 15(1), 91–103.
- Lohano, K., Loganathan, M., Roberts, R. J., & Gao, Y. (2010). When to suspect bipolar disorder. *The Journal of Family Practice*, 59(12), 682–688.
- Maas, A. h. e. m., & Appleman, Y. e. a. (2010). Gender differences in coronary heart disease. *Netherlands Heart Journal*, 18(12), 598–603.
- Mason, R. (2011). Two Kinds of Unknowing. *Hypatia*, 26(2), 294–307.
- Medina, J. (2012). *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and Resistant Imaginations*. Oxford: Oxford University Press.
- Mikhail, G. (2005). Coronary heart disease in women. *BMJ (Clinical Research Ed.)*, 331, 467–468.
- Minghella, A. (2000). *The Talented Mr. Ripley - Based on Patricia Highsmith's Novel*. London: Methuen.
- Misdiagnosis of heart attacks in women. (2019, January 5). Retrieved May 15, 2019, from British Heart Foundation website: <https://www.bhf.org.uk/information-support/heart-matters-magazine/medical/women/misdiagnosis-of-heart-attacks-in-women>
- Norman, A. (2018). *Ask Me About My Uterus: a quest to make doctors believe in women's pain*. New York: Nation Books.

- Pasha-Robinson, L. (2017, March 12). Endometriosis: Millions of women suffering due to chronic lack of research. Retrieved May 19, 2019, from The Independent website: <http://www.independent.co.uk/news/uk/home-news/endometriosis-women-suffer-chronic-underfunding-research-uterus-womb-lining-pain-endometriosis-a7623731.html>
- Pohlhaus, G. (2012). Relational Knowing and Epistemic Injustice: Toward a Theory of *Willful Hermeneutical Ignorance*. *Hypatia*, 27(4), 715–735. <https://doi.org/10.1111/j.1527-2001.2011.01222.x>
- Pohlhaus, G. (2014). Discerning the Primary Epistemic Harm in Cases of Testimonial Injustice. *Social Epistemology*, 28(2), 99–114. <https://doi.org/10.1080/02691728.2013.782581>
- Publishing, H. H. (n.d.). The heart attack gender gap. Retrieved May 15, 2019, from Harvard Health website: <https://www.health.harvard.edu/heart-health/the-heart-attack-gender-gap>
- Quine, W. van O., & Ullian, J. S. (1978). *The Web of Belief* (2nd ed.). USA: McGraw Hill Inc.
- Roland, J. (2016, April 7). Bipolar 1 vs. Bipolar 2: Know the Difference. Retrieved May 17, 2019, from Healthline website: <https://www.healthline.com/health/bipolar-disorder/bipolar-1-vs-bipolar-2>
- Scott, C. (2015, June 18). The Personal, Painful Ordeal of Women with Endometriosis. Retrieved May 16, 2019, from Healthline website: <https://www.healthline.com/health-news/personal-painful-ordeal-of-women-with-endometriosis-061815>

- Sear, K. (2009). The etiquette of endometriosis: Stigmatisation, menstrual concealment and the diagnostic delay. *Social Science & Medicine*, 69(8), 1220–1227.
- Werkheiser, I. (2014). Asking for Reasons as a Weapon: Epistemic Justification and the Loss of Knowledge. *Journal of Cognition and Neuroethics*, 2(1), 173–190.
- Yu, C. (2017, January 26). The secret battle of high-functioning depression. Retrieved May 16, 2019, from The Orange Dot website: <https://www.headspace.com/blog/2017/01/26/high-functioning-depression/>

CHAPTER 3

MENTAL ILLNESS AND NATURALISM

Introduction

Participants in the mental illness literature seek to give an account of what mental illness is and how it should be studied. Though the terminology of the debate has shifted somewhat over the past 60 years, the central contention remains the same: is mental illness appropriately ‘scientific’? Some - call them the ‘naturalists’ - think that it is, whereas others - call them the ‘constructivists’ - think that it isn’t. In this chapter, I will argue that the distinction between these two views is tangled up in a *problematic inference web*. On my conception, an inference web is a structured collection of claims and principles in which inferential relationships (of outright inference but also of association and common use) dictate the covert meaning of the individual parts of the structure (Think: Quinean web of belief but for inferential practices (Quine & Ullian, 1978)). Though many inference webs are benign, the one that I will draw attention to is problematic because:

- a. they are especially invisible or hard to track.
- b. they are inherited from previous problematic assumptions or views.
- c. they inhibit research programs by foreclosing discussion in certain areas.
- d. they perpetuate real social and epistemological harms.

In section 1, I spend some time overviewing the mental illness literature as a whole. This will serve as important grounding for the concepts and notions that will be under discussion in the rest of the chapter. The rest of the chapter is divided into four further

sections, each of which focusses on one of the following notions; reduction, realism, objectivity, and naturalism. Within each section, I demonstrate the ways in which the target notion is tied up in a problematic inference web as part of the mental illness literature.

1: The Literature

There are deep and prevalent concerns about the legitimacy of mental illness diagnoses, both in the abstract and in individual cases. There are two good reasons for this. The first is that mental illness diagnoses are extremely high-stakes. Being diagnosed as mentally ill has an enormous effect on one's life; it influences not only the way one is regarded by others, but also the way one conceives of one's self. If all acts are 'acts under a description' (Anscombe, 1957), then the acts of the mentally ill are 'acts under a diagnosis.' What may be considered 'laziness' or 'malaise' in the healthy individual, is understood as the symptom of a depressive episode in the person with Bipolar Disorder. What is taken as dieting in many, is seen as relapse in those diagnosed with anorexia. Generally, my behaviour is bound to be interpreted differently (both by myself and by others) if I carry with me a diagnosis of mental illness. This may not be a bad thing; my diagnosis may help me to understand my actions more accurately and to identify how and when to try to control those actions; it may give me access to medical treatment and therapy that are life-saving; and it may accord me important social goods such as additional time to meet deadlines, and empathy when I am struggling with my illness. But it is a significant thing: it means that mental illness diagnoses are extremely high-stakes and worthy of careful philosophical attention.

A second reason to be concerned about the legitimacy of mental illness diagnoses is that they have a history of reflecting the stigmas and prejudices of the day; in the 1850s it was commonly believed that a mental disorder - 'drapetomania' - caused slaves to run away from their owners (Cartwright, 1851; Hunt, 1855), and homosexuality was listed in the Diagnostic and Statistical Manual of Mental Disorders (DSM) as recently as 1973 (American Psychiatric Association, 2013). A tendency for 'othering' may cause the dominant and powerful majority to regard the actions of the minority as pathological, simply due to deep-seated prejudices and ignorance. Labelling such conditions as illnesses not only allows them to be marginalized further, but licenses the use (and even enforcement) of medical treatment to try and eradicate their behaviors.

The mental illness literature attempts to address these deep-seated concerns. Participants in that literature seek to give an account of what mental illness is and how it should be studied. Though the terminology of the debate has shifted somewhat over the past 60 years, the central contention remains the same: is mental illness appropriately 'scientific'? Some think that it is, whereas others think that it isn't. I'll refer to the former as 'naturalists' and the latter as 'constructivists.'

Naturalism and constructivism are familiar views within philosophy today, and there exist naturalistic accounts all sorts of things; species, gender, race, ethical properties, illness... But just what unifies these accounts is surprisingly unclear and different accounts make different assumptions about what it takes for some analysis to count as a *naturalistic* analysis. Here is just a sample of the offerings:

- If something is natural, it's more 'real.'
- If something is natural, it's more fundamental.
- If something is natural, it's non-normative.
- If something is natural, it's human independent.
- If something is natural, it's mind independent.
- If something is natural, it's objective.
- If something is natural, it's non-social.

Some of the figures I will be discussing are self-professed naturalists; others - though they argue that mental illness is 'natural' - do not call themselves by that name; and others still, don't even use the terms 'natural' or 'naturalist.' That makes it extremely difficult to talk about the literature as a whole, since one must do considerable work to figure out how the many views within that debate relate to one another. I take the most accurate and useful way of interpreting naturalism to be as follows:

Naturalism

Naturalist views take mental illness to be...

- 'Scientific' (Wakefield, 1992a, pp. 381–375).
- 'Biological' (Boorse, 1976, pp. 71–72).
- 'Factual' (Boorse, 1976; Wakefield, 1992b, p. 381).
- 'Empirical' (Boorse, 1976, pp. 69, 80).
- 'Natural' (Boorse, 1976, p. 63; Hempel, 1965, p. 147; Kandel, 2005, p. 39).

- 'Real' (Andreasen, 1997, p. 1586).
- 'Objective' (Wakefield, 1992a, p. 375).
- 'Value-free' (Boorse, 1976, p. 63).

Naturalism is characterized by the views of Christopher Boorse (Boorse, 1976, 1977), Jerome Wakefield (Horwitz & Wakefield, 2007; Wakefield, 1992a, 1995, 2000, 2009, 2009), R. Walter Heinrichs (Heinrichs, 2001), and Nancy C. Andreasen (Andreasen, 1997, 2001), according to which mental illness is:

I take the most accurate and useful way of interpreting constructivism to be as follows:

Constructivism

Constructivist views take mental illness to be...

- 'Social' (Lilienfeld & Marino, 1995).
- 'Non-scientific' (Lilienfeld & Marino, 1995).
- 'Constructed' (McHugh & Slavney, 1999, p. 302; Sedgwick, 1973, p. 34).
- 'Subjective' (Lilienfeld & Marino, 1995, p. 418).
- Value-laden (Lilienfeld & Marino, 1995, p. 418; Sedgwick, 1973, p. 32; Szasz, 1960, p. 116)

Constructivism is characterized by the views of Paul McHugh and Phillip Slavney (McHugh & Slavney, 1999), Thomas Szasz (Szasz, 1960, 1974), and Peter Sedgwick (Sedgwick, 1982, 1982).

This division is not perfect nor is it the only interesting distinction that is up for grabs. Still, I think that it is both meaningful and useful. As such, I will adopt the terms ‘naturalist’ and ‘constructivist’ to refer, respectively, to the views characterized above. This will allow me to say what I need to say about these views without getting too bogged down in their individual intricacies which, while interesting and important, do not bear heavily on my discussion.

All that having been said, participants in this debate do have two things in common. First, the common *aim* of this literature is to avoid and expose the abhorrent practices of the past (presented on pages 96-97). Second, it is virtually unanimous in the debate that the following are necessary criteria for a naturalist account of mental illness:

1. **Reduction:** the account must be able to give a *reductive* analysis of mental illness.
2. **Realism:** the account must have a *realist* ontology.
3. **Objectivity:** the account must demonstrate that the study of mental illness is or can be *objective*.
4. **Naturalism:** the account must employ and advocate a naturalist *empirical* methodology.

There are lots of ways of spelling out 1.-4. I submit that, in the current literature, 1.-4. are interpreted as 1N.-4N:

- 1N. **Reduction:** the account must be able to give a *reductive* analysis of mental illness, where ‘reduced’ is taken to imply *free from values*.
- 2N. **Realism:** the account must have a *realist* ontology, where ‘real’ is taken to imply *human independence*.
- 3N. **Objectivity:** the account must demonstrate that the study of mental illness is or can be *objective*, where ‘objective’ is taken to imply *free from human judgments*.
- 4N. **Naturalism:** the account must employ and advocate a naturalist *empirical* methodology, where ‘empirical’ is taken to imply *freedom from presuppositions*.

1N.-4N. are central to the inference web that I wish to expose and criticize. In addition to the idea that N1.-N4. are requirements of a naturalist theory of mental illness, the following claims are also central to the inference web I am targeting:

- (Division)** This is the idea that the social contrasts with the scientific.
- (Aim)** This is the common aim which is social and political.
- (Legitimate)** This is the thought that psychiatry and the study/treatment of mental illness is only legitimate if naturalism about mental illness can be established.

I aim to problematize this inference web by showing that:

- i) 1N.-4N. are taken to be connected by inferential relationships which do not survive philosophical criticism if made explicit.
- ii) The combination of 1N.-4N. with (Division) is covertly influenced by empiricist conceptions of knowledge and of science which would likely be outright rejected if seen in the cold light of day.
- iii) 1N.-4N. are at odds with (Aim), which is the underlying motivation for the debate in the first place.
- iv) The combination of 1N.-4N. with (Aim) is at odds with (Legitimate).

However, I maintain that:

- v) Each of 1-4 originates from a genuine and reasonable concern about social justice and can be given a scientifically and philosophically plausible reading (which is distinct from N1.-N4.). The latter readings should be adopted and used to reframe the debate on mental illness.

The chapter will proceed by dedicating one full section to each of 1N.-4N. Within those sections I will aim to defend i)-v). Through the paper as a whole, I also aim to establish two more general claims:

- vi) The mental illness literature operates on inference webs which are problematic in a variety of ways.

- vii) We have good reason to interrogate *the inference web itself* in addition to focussing on individual claims and accounts.

Due to the nature of the things I'm calling inference webs, my arguments should not be thought of as having any individual or distinct target, at least not in the traditional sense. It is important to my project that I target *the inference web as a whole*. As stated in vii), I think this is an important thing to do independent of the importance of addressing individual accounts of mental illness. Inference webs influence whole literatures, they dictate the language and environment of a debate. I strongly believe that there are aspects of the literature - important ones - that cannot be accounted for by addressing views individually. On my picture, concepts within inference webs code for each other so subtly and so strongly that we tend to use them interchangeably without realizing. And yet, when we do interchange them in this way, it changes the flavor of the dialectic, sometimes radically. This is the purpose for which I introduce the notion of an inference web.

A Note about Terminology:

Two short notes. First; it is standard in the literature to use the terms 'illness' and 'disease' broadly and interchangeably, so that both refer to any unhealthful condition. I will follow that convention. Second; the literatures on illness and mental illness overlap somewhat so that certain parts of what I say here will apply not only to mental illness but to illness more generally. I personally believe that many of my critical arguments will cross over to that literature quite nicely, though this is not the place for an extended discussion of that

point. The reader is invited to think more about how what I say here might apply to other literatures.

2: Reduction

Most participants in the mental illness debate seem to demand that a naturalistic account of mental illness offer some kind of *reductive analysis*. That is, they seem to take the following as a requirement of any naturalistic account of mental illness:

Reduction: the account must be able to give a *reductive* analysis of mental illness.

At first blush, this criterion might seem rather reasonable. It can be motivated by a pretty common intuition about the nature of science and reality which goes something like this:

“Science is the attempt to get at the structure of the natural world. Partly for that reason, some things should be considered the proper subject matter of science and others should not. For example, the question of whether all gold has atomic number 79 is a matter for science, whereas the question of whether all of God’s children are loved is not. This is because whether something has atomic number 79 is a fact that is borne out of the material way that the world is and whether God loves his children is not. Since the real world is, at bottom, made up of material components like atoms, electrons, quarks, etcetera, the ‘natural’ stuff - the stuff about which we conduct scientific investigation - should always boil down to these components. And since science is the study of the natural world, it should be limited to that stuff.”

That said, there are a great many ways of spelling out exactly what the reduction criterion entails. In this section, I will show that the prevalent interpretation in literature embraces 1N:

- 1N. **Reduction:** the account must be able to give a *reductive* analysis of mental illness, where ‘reduced’ is taken to imply *free from values*.

1N. inheres in the backdrop of the literature, but it is problematic. It is influenced by empiricist principles that participants in the mental illness literature and likely to want to reject. Furthermore, it is in tension with both (Aim) and (Legitimate).³⁶

2.1 Reductive Physicalism

I’ll begin my discussion with Szasz. Thomas Szasz (Szasz, 1960, 1974) is rather infamous these days due to his polemical book (and paper of the same name); *The Myth of Mental Illness*. There he makes the controversial claim that mental illness is a mere ‘myth’ or ‘metaphor.’³⁷ According to Szasz, ‘illness’ is a medical term with a ‘materialist-scientific’

³⁶ As outlined on page 99 above, (Aim) is the common social and political aim of those involved in the mental illness literature. (Legitimate) is the thought that psychiatry and the study/treatment of mental illness is only legitimate if naturalism about mental illness can be established.

³⁷ It’s not always clear whether Szasz is making a claim primarily about the ontological nature of individual mental illnesses, or a more linguistic claim about the extension of the term ‘mental illness.’ He often couches his thesis in terms of existence, claiming either that mental illness ‘does not exist’ (Szasz, 1960, pp. 113, 117) or that mental illness is a ‘myth’ (Szasz, 1960, pp. 113, 118) or sometimes that mental illness is a ‘metaphor’ (Szasz, 1974, p. xii). One can cash out this kind of claim in the following two ways:

definition (Szasz, 1974, p. xii); it refers to conditions which involve some breakdown in the tissues of the body (Szasz, 1960, p. 114). True illnesses, claims Szasz, consist in bodily malfunctions which are realized in observable lesions or abnormalities; for example, a fracture in my tibia can be identified as an abnormality in the structure of my leg which has the result that my leg is not able to function as it normally should. Szasz's contention is that mental illnesses fail to be 'real' because (unlike physical illnesses) they cannot be identified with any lesion or abnormality in the body nor do they reflect the malfunction of any 'biological norm' (Szasz, 1960, p. 114). Here's Szasz:

The concept of illness, whether bodily or mental, implies deviation from some clearly defined norm. In the case of physical illness, the norm is the structural and functional integrity of the human body. Thus, although the desirability of physical health, as such, is an ethical value, what health is can be stated in anatomical and physiological terms. What is the norm deviation from which is regarded as mental illness? This question cannot be easily answered. But whatever this norm might be, we can be certain of only one thing: namely, that it is a norm that must be

-
1. the term - 'mental illness' - when understood literally, can have no extension, and therefore must be understood metaphorically.
 2. the things - mental illnesses - that we refer to using the term 'mental illness' are not 'real' phenomena in the sense of science.

In some places, these claims are made together, in others individually, and it's not clear whether Szasz intended to make both claims or whether he mischaracterized his view in places. For this section, I'll be interested in claim 2. Regardless of whether he in fact intended to make this claim, a very reasonable reading of his view does understand him as making it, and the ensuing literature often criticizes this claim explicitly, so I think it is fair enough to focus on that claim here.

stated in terms of psychosocial, ethical, and legal concepts. For example, notions such as “excessive repression” or “acting out an unconscious impulse” illustrate the use of psychological concepts for judging (so-called) mental health and illness. The idea that chronic hostility, vengefulness, or divorce are indicative of mental illness would be illustrations of the use of ethical norms (that is, the desirability of love, kindness, and a stable marriage relationship). Finally, the widespread psychiatric opinion that only a mentally ill person would commit homicide illustrates the use of a legal concept as a norm of mental health. The norm from which deviation is measured whenever one speaks of a mental illness is a psychosocial and ethical one. Yet, the remedy is sought in terms of medical measures which—it is hoped and assumed—are free from wide differences of ethical value. The definition of the disorder and the terms in which its remedy are sought are therefore at serious odds with one another. The practical significance of this covert conflict between the alleged nature of the defect and the remedy can hardly be exaggerated. (Szasz, 1960, p. 114)

Szasz’s position seems to confuse three separate lines of thought. First, in arguing that mental illness is not identifiable with any lesion in the body, he appears to be making a claim about *individual mental illnesses*; namely that they don’t really exist due to the fact that they do not reduce to discrete physical states. Second, in claiming that mental illness is not real because it can only be explained in terms of psychosocial, ethical, and legal concepts, he appears to be making a claim about *kinds*; that mental illness is not a natural kind because individual cases of mental illness are unified merely by ‘psychosocial, ethical, and legal’ concepts. Third, in claiming that diagnoses of mental illness must make use of

‘psychosocial, ethical, and legal’ concepts he appears to be making a claim about *objectivity*; namely that mental illness diagnoses are not real/objective because whether some patient has a mental illness can only be determined if one makes use of psychosocial, ethical, and legal, concepts. Each of these three claims is importantly different, and yet they are run together in Szasz. Moreover, these claims are problematic individually as well as when combined. I will address the first claim here. The second claim will be addresses in section 2 as part of the realism discussion and the third will be addressed in section 3 as part of the objectivity discussion.

Returning to claim (1), Szasz’s claim appears to be that individual mental illnesses don’t (really) exist because they do not correspond to any ‘physical lesion in the body.’ That claim seems to operate on the following assumptions:

- Sa) The real stuff is the stuff that can be identified with discrete material states.
- Sb) Mental illnesses cannot be identified with any discrete physical state.
- Sc) Science only studies the real stuff.
- Sd) So, mental illnesses are not really real.

Sa) is prompted by a desire for reduction. The motivation for reduction here probably comes from a combination of, on the one hand, a desire to avoid cases of ‘made-up’ mental illnesses such as drapetomania and, on the other, a strong materialism. These motivations are common in the literature, and aren’t necessarily bad. But Szasz makes the mistake of taking materialist reduction to require that the reduced states be identifiable

with discrete physical states. Plenty of (most?) contemporary philosophers do not accept this view for familiar reasons (take for example, Jessica Wilson's work on the proper subset-strategy for non-reductive physicalism (Wilson, 1999, 2011)). Hence, we should probably not endorse Sa). Or, at the very least, we should recognize (as Szasz does not) that Sa) does not fall out from the desire for reduction.

Sb) appears to be motivated by a kind of Dualism about the mind. Szasz seems to think that mental and behavioural states are fundamentally non-psychical and hence, by definition, cannot be identified with any physical state (since that would make them physical states rather than mental ones).

There's nothing wrong with believing Dualism, but given that the common aim of involved in this debate is to deal with actual socio-political issues arising from the practice of psychiatry, Szasz's combination of views begins to look largely irrelevant to the issue at hand. After all, Szasz's reason for thinking that mental illnesses are not physically realizable would seem to give just as much reason to think that mental states are not real. If Szasz thinks that mental states aren't real, then it becomes trivially true that mental illnesses aren't real, at least in this sense. But that goes no way towards achieving the original aim that we started out with; namely to avoid any recurrence of the oppressive, immoral, wrongs done to people in the name of psychiatry (i.e. (Aim)). That concern is not assuaged by endorsing Dualism about the mind.

Sc) brings out a further problematic assumption in Szasz's view. Sc) represents the old empiricist commitment that physical is both i) non-social and ii) the only proper subject

matter of science. This principle becomes especially problematic when we conjoin it with Sa) and Sb). To hold all three together would be to rule out lots of important and



Rattus norvegicus

Fig. 3

uncontroversially scientific entities for there are plenty of things which both i) are not realized in discrete physical states and ii) are perfectly scientific. For example, the characteristic pink eyes and white coat of *Rattus norvegicus* (the domestic rat, fig. 3)

can be determined by a number of different combinations of alleles. Albino rats carry genotype a/a ,³⁸ which means that their cells are unable to produce melanin, resulting in fur and skin which lacks pigment. Leucistic rats also have red-eyes and white fur, but in their case this is caused by a mutations in certain genes (including *c-kit*, *mitf*, and *EDNRB*) which reduces the production of all types of pigment (not just melanin). Genetically, these two types are different (the one can make pigment the other cannot) but the trait they carry is the same; they are visually indistinguishable. Hence the trait being a 'pink-eyed-white' rat cannot be reduced to the level of the gene by identification with some discrete physical state, but surely legitimate scientific questions can be asked about this trait. The important lesson here is that the issue of whether mental illness can be reduced to physical states in this way comes apart from the main issue under discussion, which is whether mental illness is a genuinely medical, scientific notion. Whether mental illness can be reduced *in this way* might interest you, but it doesn't tell us whether mental illness is scientific because lots of things that are clearly scientific cannot

³⁸ Where 'A' designates the allele that codes for the ability to make melanin, and 'a' designates the allele that codes for the inability to make melanin.

be given such an analysis. This remains the case regardless of whether or not one is a physicalist.

Szasz's motives here are good. We need a way to explain what goes wrong in the drapetomania and homosexuality cases and it is exactly right to point to the influence of social values and biases in such diagnoses. The problem is that Szasz's response to that desire is infected by old empiricist ideals that have been shown to be defective. Of course we should reject an analysis that makes criminality and moral repugnancy illnesses by definition. To accept such an analysis would be extremely problematic not to mention harmful. But it is misguided to think that such a restriction (legitimate as it is) is reconcilable with Sa) – Sc).

I should note that it is likely proponents of Szasz's view would deny these principles if asked explicitly about it. However, the influence is there if we look for it. Later views expose these problems with Szasz's position but they remain committed to the reductive criterion in a different form. They are also problematic. I'll show this in the next subsection.

2.2 Reductive Functionalism

Various people object to Szasz by arguing that mental illness can in fact be given an appropriately naturalistic definition *if* we make use of the notion of dysfunction. Christopher Boorse is an example of such a person. He makes explicit use of work in the philosophy of mind which has developed the notion of *functional analysis*. Boorse argues that a materialist reduction need not come in the form of literally observable lesions or

abnormalities (as Szasz seems to require) since, after all, plenty of undisputedly scientific phenomena do not meet this criterion (as we saw above). Rather, he claims, naturalism about mental illness can be saved by offering a successful *functional* analysis (Boorse, 1976, 1977). Boorse offers the following, purportedly naturalistic, definition of mental illness:

An organism is healthy at any moment in proportion as it is not diseased; and a disease is a type of internal state of the organism which:

- i. interferes with the performance of some natural function-i.e., some species-typical contribution to survival and reproduction-characteristic of the organism's age; and
- ii. is not simply in the nature of the species, i.e. is either atypical of the species or, if typical, mainly due to environmental causes. (Boorse, 1976, pp. 62–63)

According to this view, mental illness consists of *mental dysfunction*; an organism has a mental illness when some mental process or part fails to adequately perform its natural function. This is fairly intuitive so far, but in order for the account to be genuinely naturalist, it is thought, it must offer a naturalist account of i) what it is to be a 'natural function,' and ii) what it means to 'adequately perform' a function.

A *natural* function is a distinct type of function. Organisms, says Boorse, are goal-directed insofar as they are 'disposed to adjust their behavior to environmental change' (Boorse, 1977, pp. 555–556). On this picture, though we can't speak of my heart as being *designed* or *intended* to have some effect, we can speak of it as having some goal or purpose; namely

the effect which contributes to the overall goal of the organism. According to Boorse, the ultimate goal (biologically speaking) of any organism is that of survival and reproduction and so it is appropriate to understand natural functions as contributions to survival and reproduction. Hence, the natural function of some part or process of an organism is defined by its contribution to the survival and reproductive success of the organism as a whole (Boorse, 1977, pp. 555–556). So for example, we might say that the natural function of the anxiety response in the species *Rattus norvegicus* is to signal when danger is present, since that is its contribution to survival and reproduction (Boorse, 1976, p. 64). Doug-the-rat's anxiety response will be dysfunctional, then, whenever it fails to adequately perform its natural function of signaling danger - either by triggering in the absence of danger or by failing to trigger in the presence of danger.

Adequate functioning is a statistical notion on Boorse's view. Most if not all of the population experience some instances of a misfiring anxiety response - for example in the period that immediately follows watching a scary film - but errors like this are species-typical and hence not enough to constitute true dysfunction. Rather, an organ, tissue, or process is dysfunctional only when it fails to perform its natural function *with the level of efficiency that is typical for members of that species that are of a similar sex and age* (Boorse, 1977, pp. 557–558). *Normal* functioning, then, is simply functional ability that is statistically typical. Take Doug, a member of the species *Rattus norvegicus* (fig. 3). On Boorse's account, Doug's heart is functioning normally if it is performing its function with at least statistically typical efficiency. What is statistically typical for an organism is determined by its reference class, which is an age group of a sex of a species. So, Doug's anxiety

response is functioning adequately when it is (statistically speaking) functioning at least³⁹ as efficiently as is average for other rats of his gender and age group.

In other words: to determine the function of one of Doug's organs or internal processes (e.g. the heart), first work out how that organ or process contributes to individual survival and reproduction of Doug's species (e.g. by pumping blood); this is its function. To determine whether that organ or process is functioning *normally*, compare its functional efficiency with the average functional efficiency of the same organ/process (e.g. other hearts) in Doug's reference class. If other organisms of the same species, sex, age-range have hearts that, on average, pump blood more efficiently than Doug's, then his heart is not functioning normally. Doug is healthy if all of his organs and internal processes have normal functional ability.⁴⁰ A *disease* is any internal state which impairs health.⁴¹

³⁹ Boorse, in later versions of his view, specifies that only functioning below typical efficiently makes for disease since there are cases where an organism functions above typical efficiency and is not ill (for example, a diminished capacity to produce lactate which results in a marathon runners increased competitive ability to run for long periods of time). We will explore this more later, in subsection 2.3.

⁴⁰ Boorse must be careful in his phrasing here. He cannot require that Doug is healthy only if all his organs are functioning with typical efficiency since some organs or processes are designed only to function some of the time (for example, Doug's anxiety response should not always be performing its function). This issue will arise again in section 2.

⁴¹ Notice that Boorse seems to be making a claim about the *kind* - mental illness - rather than about individual illnesses. We can compare this with Szasz, who seemed to move between the two claims. This is potentially problematic thought we won't have time to talk explicitly about it here.

Boorse's main route to showing that his account is appropriately scientific seems to turn on demonstrating that it is non-normative. A functionalist account of mental illness, it is claimed, can only be naturalist insofar as it gives a naturalist account of function more generally. That sounds pretty good as it is, but just how do we tell whether an account of function is properly naturalist or not? It seems to me that the criteria endorsed in the literature are roughly this: the account must offer some *reductive* definition of these natural functions by which the notion of 'purpose' or 'aim' that is operative in the idea of a function is reductively explained via factual, biological notions such as trait selection. On this view, organs, processes, and systems have natural functions which detach from any intention or use put on them by human beings. For example, we might say that my left foot can function as either a doorstep or a means of making noise, but my left foot also has a *natural* function - one which is dictated by the 'facts of biology' and not by my arbitrary needs and desires. As a result, one big constraint on Boorse's account, which is required by both Boorse and the literature as a whole, is that it is non-normative. The reduction criterion then becomes one that is centrally about normativity: a naturalist account of mental illness must be one that is non-normative.

Boorse's functionalist account has given rise to a healthy literature whose purpose it is to determine whether any successful functionalist account of mental illness can be naturalist (read: non-normative). Critiques of Boorse generally take one of the following two routes:

- a. Argue that, i) as it stands, Boorse's definition gets the extension of the concept wrong and ii) any adequate solution to that problem will introduce normativity into the account.

- b. Argue that, as it stands, Boorse's account is already normative.

Responses from Boorse or his supporters usually take one of the following routes.

- c. Argue that the extensional problems can be solved without recourse to normativity.
- d. Argue that Boorse's account is not normative.

Hence the subsequent debate ultimately becomes a discussion of whether any functionalist account can successfully banish norms. Complaint ai) looks atypical in this respect until we recognize that it is invoked almost always as a way of arguing that Boorse's account is or must be normative. One way of pursuing ai) is to argue that there are some conditions that would count as dysfunctions Boorse's account, but that are not genuine illnesses. For example, Boorse would count both osteoporosis and double-jointedness as dysfunctions, but only the former is really an illness. This strategy looks as though it has nothing to do with norms, but that appearance is misleading since it is usually just an initial step which is then used to argue that functionalist definitions will never be able to (non-normatively) differentiate between those dysfunctions that are harmful (such as the inability to go to work due to severe depression) and those that are not (such as the dysfunction of some ageing function which caused the holder to live longer⁴²) because 'harm' and 'harmfulness' are irreducibly normative (Kingma, 2010, 2014; Wakefield, 1992b).

⁴² Example due to (Wakefield, 1992b, p. 383).

Another path to argue for ai) is to show that some genuine illnesses do not count as dysfunctions on Boorse's account. For example, Boorse's statistical notion of dysfunction seems to make it definitionally impossible for any illness to be had by the majority of the people within a reference class despite the fact that there are genuine illnesses that are had by the majority of a population. For example, it is estimated that 70.1 per cent of adults over the age of 65 have Periodontitis (Eke, Dye, Wei, Thornton-Evans, & Genco, 2012), hence this condition would, on Boorse's view, be considered normal for any adult over the age of 65. But again, this kind of argument is usually used to show that Boorse must employ norms in order to bring conditions like Periodontitis into the fold.

This all goes toward demonstrating that, on both sides of the debate, it is agreed that getting rid of normativity is necessary for a successful naturalist account of mental illness. Of course, there are a great many things one could mean by requiring that an account is 'non-normative' and I believe that there are significant discrepancies within the debate about just how to cash out that notion. This issue will re-arise throughout the paper, as norms are taken to play a big role at many levels of the search for a naturalistic account (sections 2 and 3 will both have subsections that focus on normativity). Here I will discuss norms in relation to reduction; not only how these discrepancies obscure the main questions of the debate, but also how they miss the mark individually.

1.2.1. Reduction and Normativity: Language

Some seem to take it that a non-normative account must offer reductive definitions which don't make use of normative *language*. For example, Jerome Wakefield objects to J. G. Scadding's 'purely scientific biological definition' (Scadding, 1967, 1990; Wakefield,

1992b, p. 376) on the basis that it employs the term ‘disadvantage’ which is a value-term. John Sadler and George Agich and others (Agich, 1983; Fulford, 1993; Sadler & Agich, 1995) respond in similar fashion by pointing to normative language such as the terms ‘beneficial’, ‘benefit’ and ‘failure’ that are employed in the analysis (Sadler & Agich, 1995, p. 224).

Underlying these claims is the intuition that what really exists is just states of affairs. We can describe those states of affairs or we can evaluate them but the evaluations merely *grow out* of the descriptions. According to this view, the descriptive world is *out there* as an independent entity which does not bend to our will. When going about our daily lives, we often make judgments about this descriptive world and in so doing we ‘project’ onto it certain ‘subjective’ qualities such as goodness or badness, but all those projections are merely surface-matter that add nothing real to the descriptive stuff. On this picture, when asking whether mental illness is a respectably scientific notion, a big part of what we want to know is whether mental illness exists in the descriptive part of the world, or in the evaluative part of the world. This seems to be what is communicated in this quote from John Sadler and George Agich:

Moral philosophy has distinguished descriptive linguistic terms from evaluative (value) terms (Hare 1963; Fulford 1989). Descriptive terms are descriptive of states of affairs, as in “this pencil is 5 inches long.” Evaluative terms evaluate rather than describe states of affairs, as in "this pencil is good." Value terms cannot retain their full meaning if the evaluative connotations are removed. For example, to be called manipulative (a value term) is distinct from being said to be

indirect in one's intentions; it means that one is scheming and devious in one's indirect intentions. People do not want to be called manipulative because it is not a praiseworthy trait. For a concept to be value-neutral, it should be amenable to re-description without reference to evaluative terms. (Sadler & Agich, 1995, p. 223)

This kind of response really operates on three assumptions.

- La) The world out there admits of two kinds of representation by humans: the descriptive and the evaluative.
- Lb) The descriptive representation is the truly scientific one, the evaluative is not. Hence our representation of the world, if we want it to be scientific, must consist of only description and no evaluation.
- Lc) A response to Szasz and the anti-psychiatrists requires that mental illness is truly scientific (i.e. (Legitimate) is true).

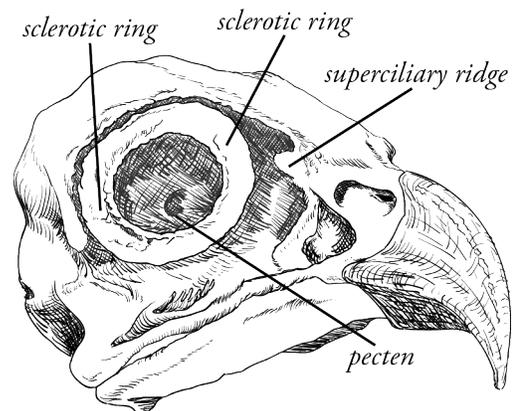
Regardless of whether one takes La) to be true, and that many be a very interesting issue, one really ought not take Lb) to be true. Insofar as terms such as 'benefit' and 'advantage' count as normative (which they surely do on various very reasonable understandings of 'normativity') it cannot be the case that the scientific is non-normative, nor can this be what Wakefield and co. intend. It is straightforwardly false that the subject matter of science is definable without reference *to any value terms whatsoever*. For example, our notion of 'electrical conductor' is perfectly scientific, but in so far as an electrical conductor is a 'substance in which electrical charge carriers, usually electrons, move **easily** from atom to

atom with the application of voltage,’ spelling out that notion is bound to involve some ‘normative’ terms.

1.2.2. *Reduction and Normativity: Teleology*

Jerome Wakefield offers a functionalist account of mental illness which employs a different analysis of function from that adopted by Boorse. According to Wakefield’s view, the natural function of an organ or tissue is the role or task it was ‘selected for.’ This way of understanding function takes advantage of the idea that a thing’s function is the process or effect that it was designed to complete or cause, it is the reason for that thing’s existence (Wakefield, 1992b, p. 383). Hence, according to Wakefield, the account does not suffer the same extensional problems that Boorse’s does because dysfunction is evolutionary rather than statistical.

For example, the skull of *Pandion haliaetus* (the Osprey) contains a ring of bone called the ‘sclerotic ring’ which surrounds the eye (see fig. 4). This ring of bone has the beneficial effect that an osprey’s eyes are harder to damage, but it also has the disadvantageous effect that the eyes cannot move in their sockets. We want to



Osprey (Pandion haliaetus)

Fig. 4

say that its function is to do the *former* rather than the latter, but what allows us to say this? According to Boorse, we can say that the sclerotic ring has the *natural* function of protecting the eye because this is its actual contribution to the survival that organism. According to Wakefield, we can say this because that is the effect that allowed the trait to

survive in the population in the first place. Whereas, for Wakefield, something's natural function is what it evolved to do, for Boorse it is what it actually contributes to survival. In this way, Wakefield's account adopts a Millikonian (Millikan, 1989) view of natural function which is purportedly naturalistically respectable.

But Wakefield's critics remain unconvinced and analogous versions of argumentative strategies a. and b. (see pages 115-116 above) continue to be mobilized. A common response is that normativity sneaks in via the notion of teleology in general (taking route b.). According to this kind of response, a view like Wakefield's fails to give a naturalist account of function because teleology is an inherently and necessarily normative notion. The idea here is that what something was designed to do, what it should do, what its purpose is, is not part of the empirical, non-social world; we can only determine something's teleology by *projecting* onto it some purpose or goal (Agich, 1983; Sadler & Agich, 1995).

Operating in the background of this dialectic is the idea that normative features of the world, such as goal or purpose, though they exist, are in some sense not real or ontologically weighty. The distinction from above - (Disinction) - also rears its head here. That's all fine as far as it goes; this is a legitimate view of ontology (though not one I adhere to). The trouble is that this view is accompanied by the following four (somewhat familiar) assumptions with which it is not compatible:

- Ta) The world out there admits of two kinds of representation by humans: the descriptive and the evaluative.

- Tb) The descriptive representation is the truly scientific one, the evaluative is not. Hence our representation of the world, if we want it to be scientific, must consist of only description and no evaluation.
- Tc) Teleology and any talk of goals or purpose is normative-evaluative.
- Td) A response to Szasz and the anti-psychiatrists requires that mental illness is truly scientific (i.e. (Legitimate) is true).⁴³

Anyone who holds Ta) and Tb) together ought not hold Tc) since to hold all three would rule out talk of, for example, the effects of speed on functionality of car seatbelts, or, at the more meta-level, whether a given experiment was successful or not. Both of these involve making judgments about purpose. In the case of seat belts, we must allocate a function or purpose to the seat belt in order that we may assess its effectiveness. Of course, car seat belts are man-made, and their purpose is heavily norm-laden, but all that will not prevent us from asking whether seat belts effectively perform the function we wish them to perform; *we can conduct scientific investigation on how well an object performs a function that we are contingently interested in*. Elsewhere in science, this is not a contentious issue, but in the case of mental illness it appears to have become one.

2.3 'Hybrid' Accounts

⁴³ Because it's been a while since I introduced these principles, here they are again: (Division) The idea that the social contrasts with the scientific; (Aim) The common social and political aim of those involved in the mental illness literature; (Legitimate) The thought that psychiatry and the study/treatment of mental illness is only legitimate if naturalism about mental illness can be established.

Wakefield (Horwitz & Wakefield, 2007; Wakefield, 1992a, 1992a, 1995, 2000, 2009) offers what he calls a ‘hybrid’ account of mental disorder which combines a ‘scientific’ account of dysfunction with a ‘value-laden’ harm component (his language, not mine). According to this account, mental illness is ‘harmful dysfunction.’ On this account, Doug-the-rat has a mental illness if and only if two conditions obtain; a) the natural function of one of his mental part or process fails to perform properly (i.e. fails to bring about the effect that it was ‘designed’ to have) and b) that dysfunction is harmful to Doug. As Wakefield presents it, part a) represents the ‘factual’ component and part b) represents the ‘value’ component (Wakefield, 1992b, p. 384). Part b) is added in the hopes of avoiding the extensional problems run into by Boorse.⁴⁴

Wakefield seems to take a) and b) to differ in *kind* and not just in content. Within just one paper, he refers to part a) as ‘factual’ (Wakefield, 1992b, pp. 381, 384, 386) ‘scientific’ (Wakefield, 1992b, p. 373), and ‘factual-scientific’ (Wakefield, 1992b, p. 381). He refers to part b) both as the ‘harm’ component (Wakefield, 1992b, p. 1985) and as the ‘value’ component (Wakefield, 1992b, pp. 373, 378, 381).

⁴⁴ Naturally, the ‘scientific-naturalist’ component of Wakefield’s account is subject to objections along routes a. and b. above (see pages 8 and 9). Some take route a. (see page 115 above) and argue that Wakefield’s account still gets the extension of the concept wrong. There is a multitude of indisputably biological processes that fail to count as proper functions on Wakefield’s account. For example, there are plenty of biological functions which appear to have been selected for one purpose but have come to serve another. Feathers are a popular example; our best scientific guess is that feathers were ‘selected for’ on the basis of their insulative properties, only later were they ‘co-opted’ in aid of flight. A Millikan style account of proper function would seem to rule that flight is not one of the proper functions of feathers.

The distinction between a) and b) brings back in that old distinction between the biological and the social that we saw come up in Szasz; (Distinction) Wakefield's account is sometimes touted as one which forgoes that distinction, and as one that is progressive in the way that it 'allows norms back in.' But this isn't really true, and the commitment to old empiricist principles remains. Consider the following quotes from Wakefield:

The concept of disorder must include a factual component so that disorders can be distinguished from a myriad of other disvalued conditions. On the other hand, facts alone are not enough; disorder requires harm, which involves values. Thus both values and facts are involved in the concept of disorder. (Wakefield, 1992b, p. 381)

The mental health theoretician is interested in the functions that people care about and need within the current social environment, not those that are interesting merely on evolutionary theoretical grounds. Thus disorder cannot be simply identified with the scientific concept of the inability of an internal mechanism to perform a naturally selected function. Only dysfunctions that are socially disvalued are disorders. (Wakefield, 1992b, p. 384)

Wakefield claims that, though normative, his account is properly naturalistic because it grounds mental illness in a naturalistically respectable account of dysfunction. The idea is that, while there are many dysfunctions, we only care about some of them. Hence, science can offer us a long list of internal states that would count as dysfunctional, but only the ones we disvalue are *illnesses* in the accepted sense of the word. So goes the thought, this

avoids the extensional problems which plague other functionalist accounts, but maintains the strong naturalistic grounding and hence avoids the worries about ‘pure value accounts’ which take mental illnesses to just be conditions which are ‘disvalued.’

It seems to me that Wakefield’s position takes the following to be true.

- Ha) Harm involves ‘value’ which is fundamentally distinct from ‘fact’
- Hb) Only ‘facts,’ and not ‘values’ are studied by science.

This is problematic because, for one thing it is entirely unclear how to even distinguish between so-called ‘facts’ and so-called ‘values,’ and for another any meaningful/significant way of cashing out Ha) requires that one reject Hb).



Fig. 5 *Ptilonorhynchus violaceus*

When taken at face value, Ha) is familiar both within philosophy and outside of it, and one may endorse this principle, but if one does then one should certainly not endorse Hc). *Ptilonorhynchus violaceus* (the Satin Bowerbird, fig. 5) collects blue materials as an essential part of

its mating ritual. The availability of such materials (for example, blue bottle-caps) is essential to the survival and reproduction of *Ptilonorhynchus violaceus*, and certainly this is an appropriate subject matter for scientific investigation. But such investigation involves

values as its subject matter. As long as we can ask scientific questions about the value of blue bottle-caps for *Ptilonorhynchus violaceus*, Hb) will (on its most face-value interpretation) be false.

Perhaps a more charitable way to understand Wakefield's claim is that *certain* values are not scientific. But which ones? Wakefield frequently distinguishes what he calls 'social norms' from other types of norms such as 'biological' or 'statistical norms.' And there are examples we might use to work out what counts as a social norm. Take Hc) and Hd) below:

Hc) What is evolutionarily disadvantageous is factual.

Hd) The inability of an internal mechanism to perform a naturally selected function is a scientific concept.

But this is confusing since, for both Hc) and Hd), there are cases which seem to involve 'social values' (at least on the most reasonable ways of understand 'social'). Clearly, we can conduct scientific study of the way in



Euplectes progne

which the propensity of a rat's tail to deglove⁴⁵ is conducive to survival (*valuable!*) or of whether the outlandishly long tail feathers of *Euplectes progne* (the long-tailed widowbird, fig. 6) are actually conducive to survival through sexual selection. Some take the equivalent of route b. (see page 116 above) and argue that even the purportedly naturalistic component of Wakefield's account carries normativity; if you think that teleology is inherently normative and you think that science is necessarily non-normative then you are likely to object to Wakefield's account on these grounds.

Wakefield thinks that the inability of an internal mechanism to perform a naturally selected function is a scientific concept, but the inability of an internal mechanism to perform a naturally selected function is surely a normative thing.

Perhaps, then, Wakefield intends to narrow the scope even further and claim that certain social norms are fundamentally distinct from biological norms. But then which ones? Here are some more principles that Wakefield seems to endorse:

- He) The mental health theoretician is interested in the functions that people care about and need within the current social environment, not those that are interesting merely on evolutionary theoretical grounds. Thus disorder cannot be simply identified with the scientific concept of the inability of an internal mechanism to perform a naturally selected function.

⁴⁵ 'Degloving' is a defence mechanism seen in rats and other small mammals whereby the top layer of skin and tissue is caused to be torn away from the bone. This mechanism is thought to be adaptive as it allows the rat (or other small mammal) to escape from immediate danger.

Hf) Because He), the functions that people care about and need within the current social environment, not those that are interesting merely on evolutionary theoretical grounds.

Hg) Because He), what functions people care about and need within the current social environment, is not a scientific concept

Notice that the issue has become less about whether something is ‘normative’ or ‘factual’, and more ‘psychosocial, ethical, and legal’ concepts about which norms are potentially oppressive and which are not. *This* is the kind of question we should be asking as philosophers of science studying mental illness. The question is not ‘how can we banish norms/values from psychiatric practice’ but ‘how can we *identify and banish the* problematic norms/values from psychiatric practice.’ To do this, we must let go of Ha) and Hb). The combination of Ha) and Hb) represents the old empiricist idea that the scientific is separate from the social (Distinction). This distinction has been shown to be deeply problematic, and I think it is also deeply misleading.

The literature on mental illness seems to assume that in order for mental illness to be medically and scientifically respectable as a notion, it must be the case that we can reduce that notion to non-normative components. But it is just not the case that norms can be divorced from or ‘reduced out of’ science in this way. If non-normativity were to be made a general requirement of science, then we would lose much of our scientific practice but we would also lose any justification for (Legitimate). (Legitimate) – which states that psychiatry and the study/treatment of mental illness is only legitimate if naturalism about mental illness can be established - is an important principle insofar as it allows us to

respond to the socio-political issues that have plagued psychiatry and psychiatric practice since the 1800s. Seeking some a-social reduction of mental illness will not help us to answer those concerns.

2.4 The specter of anti-psychiatry

David Cooper introduced the term ‘anti-psychiatry’ in 1967 (Cooper, 2001) to characterize the general movement which criticized psychiatry as it then stood. The introduction of this term went a long way toward making the mental illness literature into a ‘two-party’ affair with the anti-psychiatrists on one side, and the ‘pro-psychiatrists’ on the other (though the latter term never caught on). ‘Anti-psychiatry’ quickly became an unpopular and extreme view, characterized as the antithesis of level-headed, empirically-grounded views of the world. According to their critics, the anti-psychiatrists claimed that psychiatry was merely social and non-scientific; they argued for things like psychoanalysis (Jacques Lacan)⁴⁶ and the importance of narrative accounts of illness (R. D. Laing);⁴⁷ they were dualists who did not recognize that the mind can be grounded in the physical world (Thomas Szasz). Furthermore, the philosophical literature indicates that most people’s conception of anti-psychiatry as a movement was based on claims made by Thomas Szasz and Theodore Sarbin;⁴⁸ that mental illness does not exist; that it is merely social; that mental illnesses can never be detected using scientific tools; that medicine is not a science at all. That this is the case probably reflects the vitriol and fervor that existed in the debate at the time. There was a general sense of this movement ‘against psychiatry’

⁴⁶ See (Lacan, 1934, 2001).

⁴⁷ See (Laing, 1990, 1999).

⁴⁸ See (Sarbin, 1967).

(widely considered to be unacceptably continental) and there were those serious psychiatrists and scientists who stood opposed to this overly political, imprecise position. As a result, the term ‘anti-psychiatrist’ came to represent someone who holds any subset of the following views:

- a. Some mental disorders are not disorders.
- b. Mental disorders should not be treated with ECT etcetera.
- c. Mental disorders should not be treated in the asylum.
- d. Mental disorders should never be treated.
- e. Our methods of diagnosis are problematic.
- f. Our diagnostic criteria are problematic.
- g. Diagnosis is inherently oppressive.
- h. Mental disorder is not a natural kind.
- i. Mental disorders are not scientific categories.
- j. The study of mental illness is not scientific.
- k. Diagnoses can be used in oppressive ways
- l. Diagnostic categories are influenced by prejudices
- m. Diagnoses are influenced by prejudices
- n. Diagnoses can be oppressive
- o. Diagnoses are high stakes (in a particular way)
- p. Diagnostic categories are affected by contingent matters

This version of ‘anti-psychiatry,’ so often objected to, was likely never actually held by anyone (except *perhaps* Szasz, but even then only when interpreted rather ungenerously).

It is likely that the only property actually unifying the ‘anti-psychiatrists’ is a socio-political one; all of these figures had some concern about the welfare and rights of psychiatric patients. It is by no means true that all of these figures were against psychiatry and yet the mental illness literature goes to great pains to try and respond to ‘anti-psychiatry’ of the kind characterized by a-p above.

So why did anti-psychiatry come to be characterized in this way? Two main reasons: first historical, political, and emotional factors make it harder to see nuanced differences between different views that object to psychiatry. All those objecting to psychiatry are lumped together for ease of reference in a period of history where there were a great many people objecting to psychiatry. The high-stakes nature of the issue then obscures nuances within that grouping of views. This by itself helps lead to the fudging of inferential relationships between individual views characterized as being anti psychiatry.

Second, entrenched background views and assumptions form tight inferential connections between the different claims that are counted as ‘anti-psychiatrist.’ In this way, claims a.-j. form problematic inference web of their own; they are tied up in a tangle of dubious inferential practices. Some individual claims are wrongly taken to imply others. For example, h.-k. are invariably taken to be somewhat interchangeable and implied by a.-g. (which are also often taken to be interchangeable). But these associations are only legitimate if various other claims are assumed, such as that science is not affected by socio-political factors or biases; that science can affect the socio-political status of individual people; that scientific results can be oppressive; that the facts of science can be influenced by socio-political factors...

There is a glut of cross-inferences here. One assumption, that we have already mentioned, is that the subject matter of science is exclusively the natural world. A second assumption is that the natural world should be always reducible to the smallest empirical matter. A third assumption is that empirical matter is non-normative (we will consider this more in later sections). What follows from this is the idea that the subject matter of science should include only things that can be reductively defined via reference to smaller empirical (read non-normative) stuff.

Interestingly, all this obscures the fact that the main concern is shared by all. One of the general assumptions (listed at the beginning of this chapter) made in the literature is that any naturalist account of mental illness must accommodate (Legitimate):

(Legitimate) Psychiatry and the study/treatment of mental illness is only legitimate if naturalism about mental illness can be established.

This is encouraged largely by an opposition to anti-psychiatry and (Legitimate) itself is often taken to be the antithesis of anti-psychiatry.

Everyone involved in the debate shares the socio-political worry that mental illness diagnoses might be oppressive and unjust. The ‘anti-psychiatrists’ (though unified by little else) share the concern that diagnoses of mental illness might oppress people, they are spurred on in this worry by the fact that this kind of thing has happened in the past. They wish to legitimize the actual experiences of people who are diagnosed with mental

illnesses. The ‘naturalists’ (such as Boorse, Wakefield and co.) are concerned that people with mental illnesses receive the treatment they need without being subjected to oppressive forces. Their reason for wanting to naturalize mental illness is the same as the anti-psychiatrists’ reason for wanting to problematize it. In many ways, the main aim is shared.

The point of the anti-psychiatry movement was not to demand that mental illness diagnoses be non-normative. This is a common misconception fuelled, in all likelihood, by the fact that the term ‘anti-psychiatry’ is extremely obfuscating. The anti-psychiatrists were characterized as saying that mental illness diagnoses are merely normative judgments made by the powerful to oppress the minority (although it is overwhelmingly likely that no-one actually held this view). The caricature of the main argument of this view was ‘mental illness is not scientific because mental illness just boils down to normative judgments.’ In response, those who wanted to show that psychiatry is legitimate felt the need to show that it is *not* normative.

2.5 Section Summary

In this section, I have demonstrated several ways in which the reduction criterion is brought into play in the mental illness literature. There are generally three assumptions that go along with that criterion.

Ri. The xs reduce to the ys.

Just what this means varies incredibly from one interlocutor to another, as I have shown above. And perhaps one of those interpretations is true. That's not really at issue here. What's important is that i) the big differences between interpretations are not always made transparent and ii) the following assumptions almost always goes along with Ri.

Rii. Science only studies the xs (and ys).

Riii. In order to respond to the worries of Szasz and the anti-psychiatrists we need to establish that mental illness is part of science (i.e. (Legitimate) is true).

One should not think that Ri. and Rii. are true together. As I have shown, for any of the suggested conceptions of Ri. that I have explored in this section, Rii. is flatly false. But say one really did want to take both Ri. and Rii. to be true. In that case, then one should at least reject Riii., for, as I have shown, a combination of Ri. and Rii. is at odds with the actual project of the so-called 'anti-psychiatry' movement.

As I expressed at the beginning of the chapter, my intent here is not only to identify bad inferences individually, but to problematize the conceptual landscape as a whole. For while bad individual inferences certainly are present, they do not tell the whole story. The larger problem is the *general web* of largely unexamined, empiricist, inferential relations, beliefs and concepts of which those individual inferences form a part. My aim is to expose a number of these problematic inferences and assumptions and bring them to light. The nature of my project means that I do not have a 'target' or 'interlocutor' in the typical sense of the term; my target is the debate as a whole.

To make this claim is not to say that participants in the literature are making obvious mistakes. Scientific research and investigation operates against a backdrop of assumptions, conceptual resources, and inferential norms that guide practice within the field. This is not particularly controversial; it is common to agree that science operates against a research paradigm (Kuhn, 1996; Lakatos, 1976; Laudan, 1980). This in itself is not a bad thing; indeed it is an indispensable thing. The presuppositions made within a research paradigm allow us to clear some of the theoretical possibilities from the logical space.

What I wish to highlight is that because of this, those general webs of inference can hide problematic presuppositions (either explicit or implicit) that are essentially invisible to those operating within the field. Different views in the field operate on similar-but-different conceptions of things like reduction, realism, objectivity, and naturalism but, these views being in communication with each other, this makes for a kind of ‘cross-pollination’ which further intensifies and obscures the problem. I think that this type of mismatch is responsible for the erroneous but common move from claiming that gender is socially constructed to claiming that it is not an objectively real phenomenon. The picture that emerges offers a severely restricted understanding of science.

Contrary to being obvious, the errors I intend to expose here are especially insidious and invisible. What I want to bring attention to is the fact that mistakes like these ones can be so very difficult to see when they are reinforced by problematic webs of inference. This is especially the case when many of the individual claims within a web are independently

legitimate, as is the case here. For example, we should be concerned about whether mental illness diagnoses are inherently oppressive, or merely contingently oppressive; psychiatry has a spotted history in this regard and so it is understandable that one might worry about the kinds of norms that govern psychiatrists' diagnoses. We should also be concerned about whether mental illness diagnoses are accurate and reflect any underlying condition; again psychiatry has a spotted history in this regard. Since it is true that empirical testing forms a definitive part of scientific investigation and it is true that scientific investigation is (when done correctly) a generally reliable method of investigation, it's understandable that one might, then, have a desire make the study of mental illness more 'scientific.' The temptation to ask for reduction is motivated by the combination of on the one hand, the legitimate concern about power and control and, on the other, a deep-seated empiricist assumptions about the nature of reality.

Notice, also the looping quality of this phenomenon. Our inferential practices, frameworks, and projectability judgments are not only influenced by abstract and detached theories; they respond to the feedback we get when we use them. As such, an inference that appears to be confirmed by the world will become more entrenched. This is akin to Quine's web of belief (Quine & Ullian, 1978) but applied to inferential practices, and to Hume's notion of custom and habit (Hume, 2007, pp. 43–46). As inferential patterns become more widely adopted and more reliable, they become harder to shift; like tracks in the mud, the more you traverse them the harder they become to escape. In what follows I will refer to 'inference webs' or 'webs of inference' as a way to talk about this kind of phenomenon. My aim in this *paper* is to expose this kind of a problem in the mental illness literature. It is not just the case that bad or under-argued inferences are

being made, but that inferential pathways and presuppositions are being supported by the general inferential web. In this *section*, the point I wish to make is this: while the question ‘can mental illness be reduced to non-normative components?’ may be an interesting one, we ought not expect the answer to that question to tell us anything about:

- whether mental illness is real.
- whether mental illness is scientific.
- whether the treatment of mental illnesses is ethical.

It is important to note that my point is not that the reduction criterion is *false*, but that i) 1N’s interpretation of the reduction criterion is false and ii) 1N is employed in destructive inference webs that distort the literature as a whole.⁴⁹ In the next section, I will make an analogous argument for the realism criterion.

3: Realism

The previous section investigated the ways in which the reduction criterion is introduced in the mental illness literature at large. In this section, we will investigate an additional criterion which participants in this debate appear to endorse; the realism criterion. According to that criterion, in order for an account to be properly naturalist...

Realism: the account must have a *realist* ontology.

⁴⁹ For more on inference webs, see chapter two of this dissertation.

This criterion seems to be supported by the (familiar) intuition that science deals exclusively with a subject matter that is 'real.' If we combine that intuition with a desire to offer a scientifically respectable account of mental illness we get the realist requirement above.

That said, there are a great many ways of spelling out exactly what it means for some account to be 'realist'. In this section, I will show that the prevalent interpretation in literature embraces 2N:

2N. **Realism:** taken to imply *human independence*.

In this section, I will demonstrate this by i) presenting several different places where 2N comes out in different ways and ii) showing that each interpretation is problematic and should be discarded. To reiterate, my aim is not just to argue that the realism criterion fails, but to show that it is employed in problematic inference webs that span the whole literature.

3.1 Anti-Realism

Szasz is an anti-realist about mental illness. According to him, it is part of the definition of illness that it involve deviation from some clearly defined norm. In the case of physical illness, he claims, the norm in question is 'the structural and functional integrity of the human body' (Szasz, 1960, p. 114). What is the norm deviation from which constitutes mental illness? One might suggest that, whereas a physical illness such as a broken arm involves the disruption or distortion of the physical structures in the body at large (in this

case, in the arm), a mental illness involves a disruption or distortion of the physical structures in the brain. Szasz means ‘physical structure’ in a very reductionist sense (something along the lines of type-type identity theory). To take this option is to say that any and all mental illnesses will ultimately be explained by way of some neurological defect; something in the brain that we can point to in an MRI, inject with contrast dye, remove from a cadaver or what have you.

As we have seen, Szasz thinks that it is not possible to give such an account for mental illnesses. The best we can do, claims Szasz, is take mental illness to refer to deviation from some *social* (as opposed to physical or scientific) norm which is spelled out in terms of ‘psychosocial, ethical, and legal concepts’ (Szasz, 1960, p. 114). According to Szasz, this implies that mental illness it is not *real* in the way that physical illness is.

This argument supports Szasz’s infamous claim that mental illness ‘does not exist,’ or is not ‘real.’ It is this claim that prompted many of the vehement responses to his work, and there are very good reasons for abhorring the letter of the remark; to say that mental illness does not exist (if we take that remark at face value) minimizes the experiences of those who suffer daily from mental illness as well as those whose lives are touched by mental illness in other ways and it risks harming the welfare of those individuals by affecting their access to medical treatment and rendering them more likely to be held to blame for the effects of their illnesses.⁵⁰

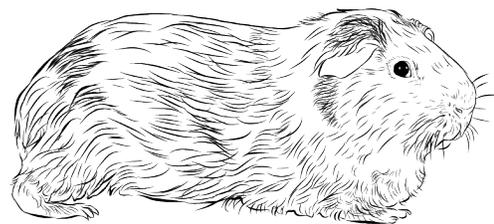
This picture seems to make the following assumptions:

⁵⁰ See the introduction of Peter Sedgwick’s ‘Psychopolitics’ for an especially impassioned presentation of this worry (Sedgwick, 1982).

- URa) The world admits of (at least) two ontological categories; the stuff that is independent of socio-political norms and the stuff that is dependent on socio-political norms
- URb) The independent stuff is the proper realm of *science*. The dependent stuff is not.
- URc) In order to respond to Szasz and the anti-psychiatrists, we must demonstrate that mental illness is part of the domain of science (i.e. (Adequate) is true).

Szasz's position seems to imply that whatever depends for its existence on socio-political norms is socially constructed and hence not really real. This is a legitimate view, but once paired with URb) it starts to run into trouble. The combination of URa) and URb) imply that the subject matter of science includes nothing which involves social or political norms.

But that simply cannot be the case since social norms can *themselves* be scientifically studied. For example, I may want to know the social significance of rumblestrutting⁵¹ in *Cavia*



Cavia porcellus

Fig. 7

⁵¹ Rumblestrutting is a guinea pig behaviour that involves moving side to side on stiff legs, sometimes accompanied by teeth chattering. It can be used as a way of establishing dominance but otherwise it forms part of the mating ritual.

porcellus (the domestic guinea pig, fig. 7) and conduct a scientific investigation into the social behavior of *Cavia porcellus*. Such an endeavor would necessarily involve some irreducibly social notions. In order to determine what significance Maggie's rumblestrutting has on the rest of the herd, I must determine whether and when she is engaging in rumblestrutting which is a behavior that can only be defined in terms of the social dynamics of the herd as a whole. This means that rumblestrutting fails to be independent of socio-political behaviour (the socio-political behaviour of *Cavia porcellus*, anyway) but (contrary to URb)) it does *not* mean that it is impossible to conduct scientific investigation into guinea pig behavior. This ought to lead us to think that it is perfectly possible to study norms in humans.

Szasz's anti-realism is at least partly motivated by the concern that mental illness diagnoses might be used as a tool for social control. This is something we *should* be worried about, the problem is in how Szasz responds to this concern. The fact that mental illnesses do not exist independently of social facts and practices does not mean that they are non-scientific or somehow less real. We *should* worry about whether mental illness diagnoses are taken to be completely 'objective' (in the sense of being detached from judgments of any kind). But that worry arises not because those diagnoses are 'merely judgments of preference,' but because mental illness diagnoses are extremely high-stakes.

That the categorization and diagnosis of mental illnesses can be oppressive is a legitimate concern, and it is relatively easy to motivate without any of the conceptual work that Szasz does. Mental illnesses are notoriously hard to define and diagnose, they are heavily

stigmatized, and most importantly they have a history of oppression (recall the examples of drapetomania and homosexuality above). For this reason, we should be extremely careful with our ascription of mental illnesses. What's problematic, however, is the way in which Szasz responds to this concern; his solution is infected by two things, on the one hand, it is infected by outdated empiricist tendencies, and on the other it is infected by deep confusions about the concerns of the anti-psychiatrist who went before him.

3.2 Social Construction: Dependence

We have just seen that Szasz distinguishes social norms from biological norms and that part of the basis for this claim seems to be an underlying intuition about dependence. According to that intuition, there is a deep difference between the things in the world that exist independently of humans, and the things in the world that exist dependent on humans. An extremely prevalent way of arguing that mental illness is 'social,' 'constructed,' or non-'scientific' is by arguing that it is somehow dependent on human beings. I think that at the core of this idea is a common intuition: the task of the natural sciences is to objectively investigate the natural world. If something depends upon humans, then it is part of the social, as opposed to the robustly natural world; it is projected onto the world, rather than existing inherently within the world.

Here's another example of this type argument being mobilized in the literature: Peter Sedgwick argues that mental illnesses are socially constructed, and hence not natural, because they are *human-dependent*. They are human-dependent in the sense that they exist *because* of humans; if there had been no humans then there would be no diseases (Sedgwick, 1973). The thought here is that while the natural world is external and

independent of us, the social world is merely contingent and ‘socially constructed’⁵² (Sedgwick, 1973, 1982). Medical categories, argues Sedgwick, are projected onto the world by humans, and not to be found in the fabric of the real world. He maintains that medical and psychiatric treatment should still go forward, but that we should be aware that the notions we employ in so doing are socially constructed and not natural.⁵³

Boorse seems to make a similar requirement. On his view, any naturalist theory of health must make it the case that whether some organism is ill or diseased comes apart from our social judgments. The underlying thought here seems to be that illnesses, like rocks, are things that exist out in nature independently of us; they are biological and factual and don’t merely depend on our social and normative preferences (Boorse, 1976, pp. 68–69).

Of course, these theorists come to very different conclusions about the prospective naturalness of mental illness. Szasz thinks that it is dependent and hence not natural; Sedgwick thinks that mental illness is dependent on humans, but is no more so than ‘physical illness’; and Boorse thinks that mental illness is independent of humans and hence natural. But on any of these conceptions, the same general principles are adopted;

⁵² Interestingly, Sedgwick employs the term ‘social construction’ liberally in his work even though Szasz hardly uses it.

⁵³ Sedgwick’s claim makes most sense when it is taken to be a claim about kinds rather than individuals. For example; someone who thinks that only humans (and not non-human animals) can have mental illnesses is trivially committed to the claim that individual mental illnesses wouldn’t exist without humans, because if there were no humans, then there would be nothing to ‘have’ the mental illness. Sedgwick does not seem to me to be making this simple claim. He means to say something stronger; that the very grouping is contingent or constructed.

- Da) The world admits of (at least) two ontological categories; the natural (which is independent of humans) and the socially constructed (which is dependent on humans).
- Db) The independent stuff is factual, empirical, and value-free. The dependent stuff is not.
- Dc) The proper realm of *science* is factual, empirical, and value-free.
- Dd) In order to respond to Szasz and the anti-psychiatrists, we must demonstrate that mental illness is part of the domain of science (i.e. (Legitimate) is true).

Da)-Dd) as a group offer a particularly good example of what I am calling an ‘inferential web.’ Da)-Dd) constitute an inference web because they form a collection of principles or claims that are taken to...

1. be shared by all participants in the debate;
2. have a common justification;

As we will see, each of Da)-Dd) is influenced by certain empiricist assumptions.

3. be inferentially connected to each other;
Some of Da)-Dd) are taken to imply and/or be implied by others of Da)-Dd).
4. be inferentially connected to common external principles or claims.

Some individual assumptions are taken to imply and/or be implied by a common external justification which lies outside of Da)-Dd).

As I've noted, inferential webs of this sort are not always problematic; indeed they are indispensable to epistemic practice. This web in particular, however, is problematic in at least four ways. Here I will discuss those problems in detail. I'll start with a general overview of the four problems, just to act as grounding for the reader, and then I'll move on to elaborate on each problem.

- a) Some of the Da)-Dd) are confused.

Da) and Db) create confusion. There are several things these theorists might mean when they say that something 'depends on' humans and different interlocutors seems to employ different understandings of dependence.

- b) Their common justification is faulty.

Da)-Dd) are influenced by outmoded empiricist assumptions that are likely false if not misused; we have good reason to reject both Db) and Dc) given how most people understand those assumptions.

- c) Some of the internal inferences are faulty.

If one holds Da) and Db) then one is obliged to reject Dc). If one holds Dc), one should reject Dd). These incompatible assumptions are often taken to be connected by inference patters that are defective.

d) Some of the external inferences are faulty.

Further, Db) reflects the combination of Tb), Tc), and Lb) which, as we have seen, are themselves faulty.

To reiterate, it's true that faulty inferences are being made here, but that is only part of the problem. The larger problem is that Da)-Dd) form an inference web which is obscuring and misleading; that there is a knot of problematic inferences that are not only made, but used as part of the fabric of the mental illness debate. This paper tasks itself with exposing the problematic inference webs that operate in the mental illness literature, and this section serves that goal by drawing out problematic inferences that surround the notion of realism.

3.2.1 Counterfactual Dependence

I said that Da) and Db) create confusion (see a) above). Why did I say that? A first issue is that it is not very clear what 'human dependence' consists in; what does Sedgwick mean when he says that mental illness 'depends on humans' for its existence? And what does Boorse mean when he says that mental illness exists independent of humans? Presumably, they don't mean to contest whether mental illness does or does not *counterfactually* depend on humans. It's true that the argument from dependence is often put in terms of a counterfactual claim: x is socially constructed, non-natural, etcetera because if humans had never existed, then mental illness would not have existed. Sedgwick repeatedly emphasizes that the existence of illness is contingent on the existence of humans; that 'there are no illnesses or diseases in nature' (Sedgwick, 1982, p. 30); that if there were no human classifications, there would be no illnesses (Sedgwick,

1973). But the simple truth of such a counterfactual can't be all that these theorists mean when they discuss dependence. This kind of counterfactual dependence doesn't give us any reason to think that mental illnesses are of a demoted status in comparison with things that are not counterfactually dependent on humans; there are a great many things that would fail to exist had humans never existed. Take for example *Biston*



Biston betularia f. typica



Biston betularia f. carbonaria

Biston betularia
Fig. 8

betularia (the peppered moth, fig. 8). *Biston betularia* has two morphs: *Biston betularia f. typica* – which is peppered white and black - and *Biston betularia f. carbonaria* – which is all black. The latter is an evolutionary off-shoot which evolved as a result of having to blend in with the sooty buildings in Manchester during a time when cotton was the main industry in that area. The existence of *Biston betularia f. carbonaria*' is counterfactually dependent on humans; if there had been no humans, it would not have existed. However, this fact doesn't make the study of these creatures any less *scientific*; we can still ask meaningful questions about their genetic makeup, behavior, and evolutionary history despite the fact that their existence is counterfactually dependent on humans.

3.2.2 Projected Dependence

Here's a better attempt at understanding what interlocutors in this debate are getting at when they make arguments like Sedgwick's: the social world is dependent on humans in the sense that it consists of highly contingent, theoretical divisions that we project onto the world. The unadulterated, unobserved world comes apart from our divisions and

definitions of it. As far as the world is concerned, there are just various constellations of physical matter changing over time. It is we who assign to that matter certain groupings or definitions. Call this kind of dependence ‘projected dependence.’

‘Projected dependence’ seems to have a Kuhnian spirit; the kind ‘mental illness’ depends on humans because, as we have learned from Kuhn (Kuhn, 1996), all kinds are paradigm-dependent. This may be a good representation of the following quote:

Animals do not have diseases either, prior to the presence of man in a meaningful relation with them. A tiger may experience pain or feebleness from a variety of causes (we do not intend to build our case on the supposition that animals, especially higher animals, cannot have experiences or feelings). It may be infected by a germ, trodden by an elephant, scratched by another tiger, or subjected to the aging processes of its own cells. It does not present itself as being ill (though it may present itself as being highly distressed or uncomfortable) except in the eyes of a human observer who can discriminate illness from other sources of pain or enfeeblement. (Sedgwick, 1982, p. 30)

The underlying claim here seems to be: ‘mental illness is a social construction; it really isn’t part of the natural world, it’s just ‘projected’ onto the world by us. If we were to go out of existence tomorrow, there would be no money in the full sense of the term.’ This kind of claim is present throughout the mental illness literature as well as in the naturalism literature more broadly. Sometimes, as in Sedgwick, the thought is put forward in explicit argument form.

This is all very well as far as it goes, but we must remember that these theorists are not simply making a claim about what it means to be human-dependent, rather they intend to make use of that claim in order to discern something about the scientific (or unscientific) nature of mental illness. But projected dependence tells us very little (if anything) about whether something is scientific, factual, or empirical. According to the Kuhnian conception, *all* kinds are dependent on humans in this way because all classification is projected. One may wish to engage in a discussion about whether the Kuhnian conception of kinds is accurate, but such a debate would be radically divergent from the issue at discussion here. Endorsing Kuhnianism will go no way towards separating the scientific from the social, or from the empirical. This supports my claim b) (see page 146 above) because this approach misuses Kuhn, and also supports my claim c) because, if one holds Da) and Db) then one is obliged to reject Dc) if one wishes to claim that Kuhn is right and kinds are paradigm-dependent, then one had better not claim that science only studies 'independent' kinds.

3.2.3 Social Paradigm Dependence

Change terminology: don't use 'social paradigms' Perhaps my notion of 'projected dependence' doesn't do justice to Sedgwick's intended conception of dependence.⁵⁴ Perhaps, rather, he intended to say that, though all kinds are paradigm-dependent, some are *badly* dependent; so the notion of atomic number is paradigm-dependent in Kuhn's sense, and so is the notion of IQ, but the latter is distinctive in the sense that it depends

⁵⁴ I'd suggest that the unclarity about the subject matter of dependence claims further supports my claim a) above.

on (something like) the ways we assess and value intelligence. Sedgwick claims that different things would be called illnesses if humans happened to value different things (Sedgwick, 1982, p. 30) and as such the class of things we call illnesses is highly arbitrary and highly context-dependent. So perhaps he means to say that the category ‘mental illness’ is not a ‘natural’ grouping because mental illnesses form a coherent class or category only if we take for granted the social and ethical preferences and norms of human beings. Call this type of dependence ‘Social Paradigm Dependence.’

But this kind of account would re-invoke the distinction between the scientific and the social (Distinction), and we have already seen that that distinction is extremely problematic. Indeed, nowadays the distinction between a natural and a social kind is often contested in the philosophy of science (and for good reason), but in any case even if one wished to hold onto that problematic distinction one would still run into troubles. This is because if one invokes the notion of ‘Social Paradigm Dependence,’ one will be forced to give up either Db) or Dc). Lots of clearly scientific kinds are ‘social paradigm dependent,’



Poecile atricapillus

Fig. 9

including kinds that we both want and need to do scientific research on. For example; the alarm call of *Poecile atricapillus* (the black-capped chickadee, fig. 9) is ‘social paradigm dependent’; it wouldn’t exist or have the function it does have had it not been for the (highly contingent) social norms of *Poecile atricapillus*. Chickadees

learn calls from their parents, and the different functions of those calls are determined socially; it takes more than one chickadee to develop an alarm call. Interestingly, the alarm

call in chickadees – *chick-a-dee-dee-dee* – has been shown to carry information about the size of the threat; a larger threat results in more ‘*dee*’s’ being added to the end of the call (Templeton, Greene, & Davis, 2005). What could be more dependent? Clearly, whether something counts as an alarm call depends on the existence and practices of chickadees. And yet, it would be extremely odd to say that chickadee calls are ‘non-scientific’ or ‘non-naturalistic.’⁵⁵

3.2.4 Restricted Social Paradigm Dependence

But perhaps Sedgwick meant to restrict his notion of human-dependence even further. Perhaps he meant to allow that not all, but *some* social-paradigm-dependent kinds are less real. Call this kind of dependence ‘restricted social paradigm dependence.’ Which social-paradigm-dependent kinds are less real on this account? The naturalistic picture we have seen so far would likely demand that the answer to that question be reductive, norm-free, and committed to realism in the ways discussed so far. But we have shown pretty well that such an account is extremely hard, if not impossible to come by. As another example, consider floods. Flooding (of a lake or river) sounds like a perfectly scientific notion, one which we have very good reason to research and gain knowledge about. Yet what it means for something to be flooded is both fuzzy and socially mediated. Rivers and lakes rise and fall very naturally, whether one is flooded depends on our judgments about where the boundary of that lake or river is; on where its banks are. Flooding, then, is human-dependent in all possible senses of that word. Yet it would be extremely problematic to claim that scientific study of flooding is impossible.

⁵⁵ This supports my claim c) above, p. 21.

Notice that, as we have worked through different possible ways of understanding dependence, this issue has become less and less about whether something is ‘real’ or how it ‘exists,’ and more and more about which norms are potentially oppressive and which are not. A similar thing happened when we were walking through the responses to Wakefield’s view. That this feeling recurs is telling. The *important* question in all of this is not one about dependence or about ‘reducing away’ normativity, it is about how to differentiate between those norms and values that should influence scientific practice, and those that shouldn’t. Our intuition is to distinguish the flooding case from, for example, the case of homosexuality as a mental illness because the latter feels ‘made up’ in a way that the former doesn’t. But this need not reflect anything like the stark ontological divide that is described in the previous passages. The flooding case can be distinguished from the homosexuality case because, the flooding case depends on a social paradigm which is benign or even beneficial whereas the homosexuality case rests on a paradigm which is oppressive and prejudiced. The ‘disease of homosexuality’ reflects the prejudices of western society and in this way exists ‘merely dependent’ on those paradigms. In contrast, flooding – while it does depend on social paradigms – depends on paradigms which we want to keep.

Human-dependence – of the sort generally described in the literature - may well be interesting, and perhaps we should do serious work to answer the question ‘is mental illness human-dependent?’ But the answer to that question does not and will not tell us

anything either about the socio-political questions that are in the background of this debate or about the scientific status of mental illness categories and diagnoses.⁵⁶

3.3 *Social Construction: Normativity*

In section 1, we saw that a large part of the literature consists in a disagreement about whether any naturalist account is sufficiently non-normative. There we looked at the normativity arguments that reflect the reduction criterion, but there are also normativity arguments that reflect the realism criterion. As we saw above, many assume that the reductive base of the material universe is non-normative and that if something is material (read: natural and scientific), then it must either be in that base or else be reducible to that base. A similar assumption is made in connection with realism. It is assumed that the *real* world is made up of non-normative matters of fact; if something is really real (read:

⁵⁶ Analogous with my remarks in previous footnotes, there is also unclarity in the intended subject matter of dependence claims like these. In talking through the idea that the social is counterfactually dependent on the natural, it is most fitting to see the claim as one about individual things rather than about kinds. Contrastingly, it feels odd to say that some *kind* counterfactually depends on humans (unless what we mean is that all the individual members of the kind counterfactually depend on humans). Analogously, the idea of projected dependence makes much more sense when taken as a claim about kinds, rather than individuals; humans project categories onto the world, not individual items. However, Sedgwick and Boorse both talk fairly often about individual mental illnesses and Szasz sometimes talks about the concept of mental illness. I mention this only to make a small point about the ways in which those engaged in this debate can slip between talk of individuals and talk of kinds. A similar problem is present in the naturalism literature at large, though I won't have time to investigate this claim further here.

natural and scientific) then it must be non-normative. This assumption comes out in a variety of ways. In this sub-section, we will consider those arguments.

3.3.1 Realism and Normativity: Judgments

Some seem to assume that facts about reality are entirely detachable from the judgments of humans. Elselijn Kingma argues, for example, that Boorse's definition of illness fails because it operates on the assignment of comparison classes which can be selected only using evaluative judgments. On Boorse's account, we assess whether Jeremy has a demyelinating disease we check his reflexes and balance and compare their functioning to that of members of his comparison class; if his reflexes function below typical efficiency then his spinal cord is dysfunctional. But whether his reflexes are functioning normally will vary greatly depending on which comparison class we use. Since Jeremy is a rat, his reflexes and coordination may be statistically normal when compared with a human child, or when compared with geriatric rats, but abnormal when compared with other rats of his age. In order for Boorse's definition to get the extension of the term 'mental illness' right, we must employ the *correct* comparison classes, but Kingma argues that there is no 'empirical fact' that determines which are the right comparison classes (Kingma, 2007, pp. 129–131). Goes the objection, this judgment is socially mediated and does not reduce to any empirical fact: 'since he [Boorse] claims to offer an account of health that is grounded in empirical fact, not evaluative judgement, he must show that empirical facts underlie the distinction between appropriate and inappropriate reference classes' (Kingma, 2007, p. 129).

Others (for example(Lilienfeld & Marino, 1995; Murphy, 2006)) make a similar

assumption that science, if it is done well, should tend towards agreement and if there is persistent disagreement between qualified scientists, then that is an indication that either the science is bad or the underlying phenomenon is not objective.

For example Scott Lilienfeld and Lori Marino (Lilienfeld & Marino, 1995) argue that naturalist accounts of mental illness will necessarily always fail because, while the proper subject matter of science necessary includes ‘clear demarcations,’ the notion of mental disorder is an inherently ‘fuzzy’ concept (Lilienfeld & Marino, 1995, p. 417). The thought here is that demarcations in nature are clean and well-defined. Therefore, if our concept ‘mental disorder’ corresponds to any category in nature, it must be possible to define it in such a way that there are no genuinely fuzzy cases. But since this is not possible, Lilienfeld and Merino propose an account of mental illness which does not aspire to be natural or scientific but rather identifies mental illness as a ‘mental construction.’

3.3.2 Realism and Normativity: Purpose

A related, but narrower, set of concerns puts an embargo only on judgments about *purpose*. These responses tend to be levelled against Wakefield’s account, arguing that, since there is no objective way of identifying what the evolutionary purpose of some organ or process is, we are forced to develop hypotheses about what the natural functions of organs are on the basis of what we can observe. Such hypotheses will, so the objection goes, necessarily involve value judgements as we make decisions about which conditions to call normal and which to call harmful. As such, they argue that Wakefield’s definition cannot be properly value-free (Murphy & Woolfolk, 2000, p. 250).

This brings us back, once again, to the empiricist intuition that the world naturally divides itself into the descriptive and the normative such that the descriptive world, and not the normative world, is the world of science. According to this view, the descriptive world is *out there* as an independent entity which does not bend to our will. When going about our daily lives, we often make judgments about this *real* external world and in so doing we ‘project’ onto it certain ‘subjective’ qualities such as goodness or badness. Take the following quote from R. E. Kendell for example:

The most fundamental issue, and also the most contentious one, is whether disease and illness are normative concepts based on value judgments, or whether they are valuefree scientific terms; in other words, whether they are biomedical terms or sociopolitical ones. (Kendell, 1975, p. 25)

This kind of picture is motivated by the (very reasonable) desire not to see certain prejudiced and wrongful preferences (for example, the preference for others not to engage in homosexual sex) influence medical judgments. I have this desire too. But it does not imply that we are barred from studying things on the basis of the fact that we think it important or that we care about them. In order to



Fig. 10 *Malurus cyaneus*

conduct a study on the capacity of *Malurus cyaneus* (the fairy wren, fig. 10) to learn the alarm calls of other species, we must ask how well different songs perform their function. In order to do that we must make judgments about what the fairy wren is trying to do

when listening to the call of another species. *This is still good science.* We can perfectly well ask ‘how successful are fairy wrens at avoiding predation by eavesdropping on the calls of other species?’ without losing scientific credibility. Asking how well some x performs a function f is a perfectly legitimate way to do science. What makes the homosexuality case different is *which* function we are asking about.

3.3 Section Summary

In this section, I have demonstrated several ways in which the realism criterion is brought into play in the mental illness literature. There are generally three assumptions that go along with that criterion.

Ei. The *real* stuff is the stuff that is x.

Two things are important to notice here. The first is that interlocutors do not share the same conception of how to fill in ‘x,’ the second is that Eii. And Eiii. almost always go along with Ei.

Eii. Science only studies the *real* stuff

Eiii. In order to respond to the worries of Szasz and the anti-psychiatrists we need to establish that mental illness is part of science

Above, I said that if one holds Dc), one should reject Dd) (see option c. on page 146). This is the case for the same reasons that accepting Ta)-Tc) obliged one to reject Td). Anti-realism is often seen as a defining feature of the ‘anti-psychiatrist’ position, but it’s

not really accurate to call many of the so-called anti-psychiatrists ‘anti-realists.’ Erving Goffman (Goffman, 1961, 2009), for example, argued that institutionalization is both immoral and ineffective as it forces patients merely to ‘learn how to be a good patient.’ This claim is critical of psychiatric practice, but not inconsistent with a strong realism about mental illness. One might think (something close to Lacan’s view) that mental illness is a real kind which can be studied and whose sufferers can be treated if we employ the correct methods of treatment. Even those who think that psychiatric treatment is inherently oppressive (in the spirit of Laing for example) might think that mental illness is a *real* phenomenon. Michel Foucault, in ‘The History of Madness’ (Foucault, 1961) - also associated with the anti-psychiatrists - discusses the historical development of the asylum and problematizes its increasingly institutionalized nature, but this really doesn’t address the ontological question of whether mental illness is really *real, normative*, dependent or what have you. If one wishes to respond to the concerns of the anti-psychiatrists (and those concerns are diffuse), one must attend to the socio-political aspects of that notion; one needs to consider how treatment can be given in a way that is ethical, how power structures can be exposed and their influence minimized, and how we can protect the rights of the mentally ill.

To reiterate, the issue I wish to raise here is not so much with the bare question of whether mental illness is ‘projecting dependent’ or whatever. Rather the issue is in the fact that the answer to that question is taken to tell us something about the answers to other important questions, such as whether mental illness is scientific, empirical, factual, objective, social... etcetera. My general claim is that, while there questions about dependence may well be interesting, and their answers may well be informative, those

answers do not tell us whether mental illness is a ‘scientific’ phenomenon. To do justice to the underlying *legitimate* concern (which is that the mentally ill be treated in a way that is humane, non-oppressive, and helpful) we should focus instead on whether the assumptions and paradigms of thought we make use of are the ones we *want* to make use of. Are those paradigms just? Do we know what they are? Have we reflected on them? These are the important, nuanced questions which require serious analysis by philosophers of science.

4: Objectivity

Our third criterion is the criterion of objectivity. Most participants in this literature seem to demand that a naturalistic account of mental illness show that mental illness can be studied objectively. That is, they seem to take the following as a requirement of any naturalistic account of mental illness:

Objectivity: the account must demonstrate that the study of mental illness is or can be *objective*.

Our previous two criteria made demands about the *ontology* of naturalist analyses of mental illness; that it be reductive and that it be realist. This criterion focuses more on *methodology*; the idea is that a genuinely naturalist account of mental illness will allow for and/or involve properly scientific (read ‘objective’) methods of investigation.

This is pretty intuitive. Scientific investigation is taken to be special in the sense that it is particularly rigorous, very reliable, and especially resistant to (distorting) subjective

opinions. The thought seems to be that if, as the naturalist claims, mental illness is truly scientific, then we should be able (at least in principle) to investigate it without making subjective judgments. Moreover (continues the thought), if we can achieve this then the anti-psychiatrists' worry about power and control will be dealt with. According to this picture, objective investigation plus realism (about the ontology of mental illness gives us truth, and truth *is just truth*; it is detached from any political or prejudicial slant.

Once again there are a great many ways of spelling out exactly what the objectivity criterion entails. I believe, however, that most conceptions of the criterion share the common assumption that *objective investigation is not normative*. We have talked in both previous sections about the influence that 'normativity' has on naturalists. In section 1, we saw that normativity is often thought of as being irreducible to the material and for that reason is taken to be forbidden in naturalist analysis of mental illness. In section 2, we saw how normativity is thought to indicate a reduced ontological status, of dependence on humans or contingency or something like that. But in this section the normativity concern will really take center stage. Here I will focus on the different ways in which 'free from norms' has been understood in connection with objectivity in the mental illness literature. I do this with a motive to arguing that i) normativity is understood very differently across the literature and ii) given the prevalent ways of understanding normativity, participants in the literature ought not accept the objectivity criterion. Hence, my point once again is not only that the stated criterion is false, but that it is employed in destructive inference webs that do an extremely effective job of mutating the literature as a whole.

4.1 Objectivity and Normativity: Judgments

As previously noted (page 162 above), some argue that Boorse's account fails to be properly naturalist because there is no non-normative way of determining the appropriate reference class for any prospective case of illness. According to Boorse, Jeremy has a demyelinating disease in virtue of the fact that his spinal cord is functioning below typical efficiency; his reflexes and balance are poor in comparison with members of his comparison class. But this is only the case as long as we compare Jeremy with the right class of people; since Jeremy is a rat, his reflexes and coordination may be statistically normal when compared with human children, or when compared with geriatric rats, but abnormal when compared with other rats of his age. Goes the objection, the only way for us to determine the correct reference class is to make normative judgments about how efficiently Jeremy *should* be functioning (Kingma, 2007, pp. 129–131).

Above we interpreted that concern as being influenced by ontological commitments; the fact that judgments were required to settle the matter meant that there was *no external fact of the matter* that determined which reference class was the right one. However, one might also take this concern to be motivated by *epistemological* commitments. This reading would have it that judgments are not permitted on naturalist analyses because judgments are merely subjective, and subjectivity means a lack of objectivity hence Boorse's account cannot be naturalist since objectivity is a requirement of naturalist accounts of mental illness.

This line of thought seems to make the following assumptions:

- Ja) Objectivity is a requirement of science.
- Jb) Judgments are not objective.
- Jc) In order to respond to Szasz and the anti-psychiatrists, we must demonstrate that mental illness is part of the domain of science (i.e. (Legitimate) is true).

Ja) and Jb) require that any naturalistic account of mental illness be completely free from the ‘judgments’ of scientists. On the most straight-forward reading of Jb), this will include any kind of judgment made by scientists. But this is not a viable position, for familiar reasons. As the philosophy of science has taught us, the judgments of scientists are indispensable at any level of scientific investigation: we only manage to learn anything from our scientific experimentation because we introduce what is sometimes called ‘extra-experiential criteria’ to allow us to tell between empirically equivalent theories. (Duhem, 1954, pp. 185–187; Quine, 1963, p. 42). The extra-experimental criteria allow us to narrow down the field of infinitely many empirically equivalent theories to a manageable number of theories that can be tested against one another. In essence, they let us kick out all of the outlandish theories and focus on the theories that are more likely to be true. How do we decide which theories to kick out and which to keep? On the basis of ‘projectability judgments’; judgments guided by what we think a good theory looks like in the context under scrutiny. Though the term is technical, projectability judgments themselves are extremely familiar to us all. When assessing a situation for example making a diagnosis, we are presented with information. We must process that information and differentiate the relevant from the irrelevant, the more useful from the less useful. We do this somewhat holistically. For example, a relatively uncontroversial ornithological fact is

that *Troglodytes troglodytes* (the Eurasian Wren) makes ‘dummy nests.’ Male wrens have been seen to build up to twelve separate nesting sites as a way of encouraging females to mate with them. We know this on the basis of observation (male wrens have been observed actually building the nests and later observed using only one of those nests) but not *just* on the basis of observation. In order to establish our hypothesis (that male wrens build dummy nests to attract a mate) on the basis of our observations, we are obliged to make a whole host of additional assumptions such as; that the bird I see really is a wren and not (for example) a sparrow; that I am managing to track the same nests day-by-day; that the nests I am tracking were genuinely made by the wren in question... etcetera. Most of these assumptions likely seem self-evident, but they are assumptions nonetheless, and they are essential to any kind of empirical testing.

Hence, scientific investigation necessarily requires us to make judgments in the form of projectability assessments. Even when studying, for example, the comparative lubricating properties of different fluids one must identify which tests to conduct under which conditions and using which tools. It simply cannot be the case that scientific investigation is conducted in the absence of *any judgments whatsoever*.

When read this way, Ja) and Jb) are incompatible with our common-sense notion of science. Perhaps this isn’t a charitable enough reading of Jb), though. One might argue that that worry is not about judgments wholesale, but about judgments *about value*; the concern about reference classes demonstrates that scientists must judge whether a condition is *desirable* in order to be able to determine which reference class is appropriate. For example, when determining how to test for lubrication, scientists make judgments

about the chemical properties of fluids, the rate of change, the movement of atoms and molecules, etcetera. This is permitted since these judgments concern empirical facts of the matter. But were they to instead base their determination on judgments about which fluid is the nicest color, for example, or on which was the cheapest to produce, this would not be scientific. According to this line of thought, judgments may be objective if they reflect empirical facts, but are subjective if they reflect values and norms. The latter kind of judgment is not scientifically respectable though the former is.

But still, judgments about value are ten-a-penny in scientific discourse. Our ability to differentiate genes and study gene expression requires that we make judgments about what functions genes perform, and these assumptions must be guided by intuitions about what would be adaptive in the relevant contexts.

If one is happy to dispense with the folk intuitions about what counts as science, then one might get away with holding both of Ja) and Jb) together, but in that case there is absolutely no reason to hold Jc). The spirit behind Jc) is respectable; it reflects a desire to avoid situations in which people are determined to be mentally ill merely on the basis that their behavior is not valued by those in power. One might suggest that this is part of what resulted in the pathologizing of homosexuality; more?. But in this context adopting Jb) only makes sense if Jc) is false. This legitimate concern about oppressive normative judgments seems to have been infected by the thought that ‘judgments’ or ‘norms’ are the enemy of objectivity and that facts about the natural, non-social world can be discovered without the use of judgments. This is not true, as we have seen.

One might propose that we further restrict the notion of ‘judgement’ that is operative in Jb). Perhaps we should take ‘judgment’ in this context to refer only to certain judgments about value. For example, we might say that judgments about how well some liquid lubricates a knee joint is permitted, but judgments about whether gay sex is abnormal is not. This looks like progress to me, but we must acknowledge that the question is now no longer one about whether psychiatry can be ‘judgment-free.’ In order to ask this (better, more relevant) question, it is imperative that we move away from the model of science as an a-social, a-political, endeavour. We must break out of the restrictive inference web that forces an overly restrictive, empiricist, picture of science onto a debate that is inherently both scientific and political. As it currently stands, the inferential web in which the literature is tied up forecloses discussion of the type of view I outline here. This is both philosophically problematic and deeply harmful.

4.2 The Medical Model

Participants in the contemporary mental illness debate almost universally assent to what they call ‘the medical model’ for mental illness. According to the medical model, psychiatry and the study of mental illness should be conducted in a way directly analogous to the practice of medicine more generally; it should employ the methods and background of the modern medical sciences. On at least an initial viewing, adherence to the medical model seems to imply naturalism about mental illness; no one claims, as Szasz (Szasz, 1974) and Sarbin (Sarbin, 1967) did, that the study of mental illness can necessarily not be conducted as a legitimate science. Yet despite this seeming agreement, many of the same kinds of controversies endure. This is due to the fact that there is significant disagreement over just what the medical model entails.

Some advocate for what tends to be referred to as a ‘minimal’ or ‘weak’ interpretation of the medical model. On this approach, the correct way to understand our diagnostic criteria is as classifications or groupings of cases based on observable symptoms and signs. The DSM is usually taken to present mental illnesses in this way. For example, the diagnostic criteria for a manic episode are as follows:

- A. A distinct period of abnormally and persistently elevated, expansive, or irritable mood and abnormally and persistently increased goal-directed activity or energy, lasting at least 1 week and present most of the day, nearly every day (or any duration if hospitalization is necessary).
- B. During the period of mood disturbance and increased energy or activity, three (or more) of the following symptoms (four if the mood is only irritable) are present to a significant degree and represent a noticeable change from usual behavior:
 - 1. Inflated self-esteem or grandiosity.
 - 2. Decreased need for sleep (e.g., feels rested after only 3 hours of sleep).
 - 3. More talkative than usual or pressure to keep talking.
 - 4. Flight of ideas or subjective experience that thoughts are racing.
 - 5. Distractibility (i.e., attention too easily drawn to unimportant or irrelevant external stimuli), as reported or observed.
 - 6. Increase in goal-directed activity (either socially, at work or school, or sexually) or psychomotor agitation (i.e., purposeless non-goal-directed activity).

7. Excessive involvement in activities that have a high potential for painful consequences (e.g., engaging in unrestrained buying sprees, sexual indiscretions, or foolish business investments).
- C. The mood disturbance is sufficiently severe to cause marked impairment in social or occupational functioning or to necessitate hospitalization to prevent harm to self or others, or there are psychotic features.
- D. The episode is not attributable to the physiological effects of a substance (e.g., a drug of abuse, a medication, other treatment) or to another medical condition. (American Psychiatric Association, 2013, p. 124)

The DSM dictates; ‘Criteria A-D constitute a manic episode. At least one lifetime manic episode is required for the diagnosis of bipolar I disorder.’ (American Psychiatric Association, 2013, p. 124). On this approach, one meets the diagnostic criteria for bipolar I disorder in virtue of the fact that they demonstrate symptoms which are enough in line with the symptoms that make up the diagnostic criteria. Note that the claim here is not simply that all we can achieve right now is a rough taxonomy based on observed symptoms, but rather (something like) that such classifications *is all we should or can aim for*. According to the minimal interpretation, individual mental illnesses are united by their characteristic symptoms and typical progression. Taxonomy, on this view, should be carried out via the ‘objective’ and precise analysis of observational data.

In contrast with those who accept the minimal interpretation of the medical model, there are those who accept the strong interpretation of the medical model. According to this view, medicine must aim at giving diagnostic criteria that reflect not just the groupings of

symptoms on the basis of evidence, but the ‘underlying causal mechanisms’ of disorders. Proponents of the strong interpretation suggest that we employ extant scientific resources - such as the resources of cognitive neuroscience (Andreasen, 1997; Heinrichs, 2001) or molecular biology (Kandel, 2005) - as background theories to guide our work within the realm of psychiatry. The back-and-forth between these two schools of thought brings out some additional problematic inference webs. I’ll investigate those in this sub-section.

4.2.1 Objectivity and Normativity: Assumptions

Advocates of the minimal interpretation of the medical model (for example (Guze, 1978; McHugh & Slavney, 1999)) tend to claim the medical model dictates that all we can hope to do is create diagnostic criteria on the basis of observation and similarity judgments. For example, Paul McHugh and Phillip Slavney (McHugh & Slavney, 1999) claim that the scientific method requires that observation is the fundamental method of investigation. According to them, we cannot observe underlying causes, but can only observe symptoms and construct diagnostic categories on the basis of those observations. For this reason, we should understand diagnostic categories as clusters of symptoms that occur together and we should define mental illness itself as ‘a construct that conceptualizes a constellation of signs and symptoms as due to an underlying biological pathology, mechanism and cause’ (McHugh & Slavney, 1999, p. 302). Hence, on this account, disease is not a ‘physical process’; ‘its essence is conceptual and inferential’ (McHugh & Slavney, 1999, p. 48).

The underlying thought here seems to be one about objectivity; all we can aim to do is create diagnostic criteria on the basis of observation and similarity judgments because an

account that seeks to create diagnostic criteria on the basis of underlying causes would have to make assumptions about, e.g., the correct theory of mind, and to make such assumptions would be contra to objectivity and hence not scientific. This is a kind of picture that trades on the following assumptions:

- Ua) Objectivity is a requirement of science.
- Ub) Objectivity requires that we do not make any assumptions.
- Uc) In order to respond to Szasz and the anti-psychiatrists, we must demonstrate that mental illness is part of the domain of science (i.e. (Legitimate) is true).

In keeping with what we have seen elsewhere, assenting to both Ua) and Ub) will result in huge losses for science. Indeed, if one wishes to maintain Ua), Ub) becomes close to impossible to maintain since it would imply that science must be conducted in the absence of all assumptions. This is impossible because individual hypotheses do not support testable predictions; in order to generate a testable prediction one must have both individual hypotheses and theoretical principles, and assumptions. In the absence of assumptions, science becomes impossible because assumptions are needed in order to make predictions on the basis of theories. This follows simply from the inverse of Quine and Duhem's classic lesson that only 'whole sciences'⁵⁷ - and not lone hypotheses - can furnish us with testable predictions (Duhem, 1954; Quine, 1963).

⁵⁷ A 'whole science' in this sense is a huge collection of posits consisting of the hypothesis under consideration *plus* all the background assumptions that support that hypothesis (Quine, 1963).

4.2.2. *Objectivity and Normativity: Background Theories*

Proponents of the strong interpretation do not make the same requirement that assumptions be banished, though they do enforce strict restrictions on *which* assumptions are allowed in a properly ‘medical,’ ‘scientific’ approach. Generally, the accepted background theories tend to be ones that would meet the reduction and realism requirements of the previous sections. Nancy C. Andreasen (Andreasen, 1997) argues that we should utilize the notions of cognitive neuroscience MORE. R. Walter Heinrichs (Heinrichs, 2001) claims that we need a developed theory of mind/brain that lets us identify psychological abnormalities. In general there is a preference for ‘algorithmic’ or explicitly formulated background theories, which are seen as being objective, over ‘the (tacit) judgments of scientists,’ which are being subjective. For example, Boorse is happy to utilize statistical and evolutionary theory/assumptions in his naturalist account and Wakefield contrasts ‘statistical norms’ with ‘social norms.’

In general there is a preference for ‘algorithmic’ or mechanical background theories - which are seen as objective - over ‘the judgments of scientists’ - which are seen as subjective. For example, Boorse is happy to utilize statistical and evolutionary theory/assumptions in his naturalist account. This is problematic. The strong interpretation is right to reject the idea that science can be conducted in the complete absence of any background theories, but it makes the same mistake at a higher level when it requires that the accepted background theories be neuroscience, statistical facts, evolutionary facts etcetera on the basis that these theories are more objective. They assume that algorithmic methods etcetera remove judgments. Others (for example (Lilienfeld & Marino, 1995; Murphy, 2006)) make a similar assumption that science, if it is

done well, should tend towards agreement and if there is persistent disagreement between qualified scientists, then that is an indication that either the science is bad or the underlying phenomenon is not objective.

For example according to Scott Lilienfeld and Lori Marino (Lilienfeld & Marino, 1995) argue naturalist accounts of mental illness will necessarily always fail because, while the proper subject matter of science necessary includes ‘clear demarcations,’ the notion of mental disorder is an inherently ‘fuzzy’ concept (Lilienfeld & Marino, 1995, p. 417). The thought here is that demarcations in nature are clean and well-defined. Therefore, if our concept ‘mental disorder’ corresponds to any category in nature, it must be possible to define it in such a way that there are no genuinely fuzzy cases. But since this is not possible, Lilienfeld and Merino propose an account of mental illness which does not aspire to be natural or scientific but rather identifies mental illness as a ‘mental construction.’

But it’s just not true that science is ‘objective’ in the way that these figures understand. Algorithmic methods do not remove judgment from calculations of similarity but rather write them in to the mechanism itself. The judgments of scientists and the use of background theories are indispensable at any level of scientific investigation. It may well be right that we should make use of the results of cognitive neuroscience in developing diagnostic criteria, but if it is the reason for that will not be because cognitive neuroscience is more objective than similarity judgments made on the basis of observed symptoms. This is to say, one really ought not hold the following three claims together:

- Ba) Naturalistic accounts of mental illness must make the study of mental illness scientifically respectable (i.e. (Legitimate) is true).
- Bb) Objectivity is a requirement of science.
- Bc) Objectivity requires freedom from background assumptions.

Apart from Ba)-Bc) being inconsistent with actual science, they may also be inconsistent with ideal science, since background assumptions may actively aid in the pursuit of good science. Standpoint theories, for example, suggest that an agent's particular standpoint – their background and social status - may enable them to better assess evidence (Harding, 1992). This makes sense given the role that projectability judgments play in our epistemic practices. If sexist beliefs form part of the background assumptions for one's theorizing, then those assumptions are likely to guide one away from the more accurate interpretations of the data. Racist background beliefs are often posited as a contributing factor in the racist research on IQ conducted in the 1960's (Block & Dworkin, 1976). Standpoint theory gives us reason to believe that, not only is Bc) practically impossible, it is epistemically non-ideal.

Psychiatry has a spotted history, marred by prejudiced beliefs and oppressive acts. That history leads many to be concerned about the role that assumptions and background beliefs play in psychiatric research and practice. Rightly so. However, it is a mistake to express this worry in the form of calls for 'assumption free' science, as the current literature frequently does. The wrongs perpetrated in the name of psychiatry are not wrongful because they involve assumptions, they are wrongful because they involve prejudiced assumptions. It is common in the literature to tacitly assume that the remedy to

this worry is to i) demand that psychiatry be more scientific and ii) understand ‘scientific-ness’ to require freedom from assumptions. This kind of reaction is misguided; assumption-free science is neither possible nor desirable. Moreover, to aim for such a thing renders important, tenable views invisible by effectively ‘framing them out’ of the debate.

4.3 Section summary

The concern about objectivity is motivated by the thought that ‘judgments’ are the enemy of objectivity and that facts about the natural, non-social world can be discovered without the use of judgments. This is not true. It may well be right that we should make use of the results of cognitive neuroscience in developing diagnostic criteria but, if it is, the reason for that will not be because cognitive neuroscience is more objective than similarity judgments made on the basis of observed symptoms.

In this section, I have demonstrated several ways in which the objectivity criterion is brought into play in the mental illness literature. As we saw at the beginning of the section, there are generally three assumptions that go along with that criterion.

Oi. Objective investigation is not normative.

Just what is meant by ‘objective’ varies incredibly from one interlocutor to another, as I have shown above. And perhaps one of those interpretations is true. That’s not really at issue here. Two things are important to notice here. The first is that the different

conceptions are wildly different and this is not always made transparent. Second is that the following assumption almost always goes along with Oi.

Oii. Scientific study is always objective.

Moreover, Oiii. invariably goes along with Oi. and Oii.:

Oiii. In order to respond to the worries of Szasz and the anti-psychiatrists we need to establish that mental illness is part of science.

We are right to be concerned about the legitimacy of mental illness diagnoses. A tendency for ‘othering’ may cause the dominant and powerful majority to regard the actions the minority as pathological, simply due to deep-seated prejudices and ignorance. Labelling such conditions as illnesses not only allows those who have those conditioned to be marginalized further, but licenses the use (and even enforcement) of medical treatment to try and eradicate their behaviors. A natural response to this is to look to restrict the influence of socio-political judgments on medical diagnoses and to make the study of mental illness more reliable and rigorous; to ask that the study of mental illness be more objective and scientific. But it is misguided to think that such a restriction (legitimate as it is) is reconcilable with the idea that objectivity requires that scientific practice be conducted without recourse to norms of any kind.

5: Naturalism

Our fourth and final criterion is the criterion of naturalism. According to this criterion, the following is a requirement for any naturalist account of mental illness:

Naturalism: the account must employ and advocate a naturalist *empirical* methodology.

This criterion focuses on methods of investigation that are empirically grounded. Most participants in this literature seem to demand that a naturalistic account of mental illness show that mental illness can be studied *empirically* in this way.

But what does it mean for a method of investigation to be ‘empirical’? Once again, there are a number of ways of spelling this out and this section will look primarily at the different conceptions of ‘empirical investigation’ that are conveyed in the literature. My aim will be familiar; to argue that i) theorists participating in the literature are operating under quite different understandings of the naturalism criterion and ii) given their other assumptions, these theorists ought not accept that criterion. As a reminder; my aim here is not only show that the stated criterion is false, but that it is employed in destructive inference webs that distort the literature in harmful ways.

5.1 Observation

Common among advocates of the minimal interpretation of the medical model is the idea that properly empirical investigation goes no further than what is directly observable. As we saw above, McHugh and Slavney claim that the scientific method requires that observation is the fundamental method of investigation (McHugh & Slavney, 1999). This

sentiment is also seen in the DSM and in Samuel B. Guze's work (American Psychiatric Association, 2013; Guze, 1978). Above, we interpreted this statement as being motivated by certain views about objectivity, but it can also be motivated by a certain picture of scientific investigation. On that picture, observation, being our only connection with the outer world, must be the principle method of investigation and hence, any naturalist account of mental illness must have it that diagnostic criteria reflect only groupings of observable symptoms (McHugh & Slavney, 1999). According to this approach, while these symptoms are likely due to a common cause or mechanism, the job of science is not to speculate about that underlying cause but to operationalize diagnostic categories enough that we can predict and treat disorders as they are presented to us.

This kind of view appears to invoke the following assumptions:

- Sa) The scientific method requires that observation is the fundamental method of investigation.
- Sb) We cannot observe underlying causes.

And on the basis of Sa) and Sc), it is concluded that:

- Sc) We cannot, in good scientific conscience, claim to discover underlying mechanisms; we can only observe symptoms and construct diagnostic categories of those observations.

Added to Sa)-Sc) is the following familiar claim:

Sd) In order to respond to Szasz and the anti-psychiatrists, we must demonstrate that mental illness is part of the domain of science (i.e. (Legitimate) is true).

Sa)-Sd) as a group have the implication that we can never know the underlying causes of mental illnesses.⁵⁸ For this reason, these theorists argue, we should understand diagnostic categories clusters of symptoms that occur together.

This kind of argument rests on a problematic picture that is based on old empiricist conceptions of science. This position fails to recognize the fact that it is impossible to conduct scientific investigation on the basis of observation alone. As discussed in section 3, only whole sciences (and not individual hypotheses) are confirmed by data. As such, we must assume the accuracy of a whole host of background theories in order to establish a single hypothesis, even if that hypothesis is just ‘these symptoms tend to co-occur.’ If

⁵⁸ It is worth highlighting that this does not necessarily entail anything about existence. Proponents of the minimal medical model do not necessarily claim that mental illnesses *are identified with* or *reduce to* groupings of symptoms (though some of them may think this). Rather, the issue is over i) what methods of investigation we are allowed to use and ii) what we are allowed to conclude on the basis of those methods. According to the minimal model, our method must be observation and, because we can only use observation, our conclusions can only be that certain observable symptoms can be grouped together; somewhat like the Humean picture of cause and effect. The strong interpretation responds that the methodology of science allows us to make observations but also to make inferences on the basis of those observations including using background theories to guide our investigation.

proponents of the minimal medical model were to truly and whole-heartedly assent to Sa)-Sc), they would be ruling out all of science as unscientific.

5.2 Predictive Validity

Both strong and minimal interpretations agree that predictive validity is important for the medical model but they disagree about what predictive validity consists in. The minimal interpretation claims that predictive validity requires only that we have a certain level of success in predicting events. The strong interpretation claims that predictive validity requires that we also have an understanding of the underlying cause of some symptom or disease. The underlying cause gets are the 'reality' of the disease and should be identified as some biological process that can be identified and observed.

Proponents of both the strong and minimal interpretations agree that predictive validity is important for the medical model. They disagree about what predictive validity consists in, though. The minimal interpretation claims that predictive validity requires only that we have a certain level of success in predicting events. The strong interpretation claims that predictive validity requires that we also have an understanding of the underlying cause of some symptom or disease. The underlying cause gets are the 'reality' of the disease and should be identified as some biological process that can be identified and observed.

5.3 Ontology and Methodology

Above I flagged the fact that the objectivity criterion makes a methodological requirement while the realism and reduction criteria make ontological requirements. This distinction reflects a more general distinction between two types of naturalist position.

Some take naturalism to be an ontological thesis according to which certain things (the material, empirical, real things) have a weightier ontological status than others. Others take it to be more of a *methodological* thesis; they see naturalism as a view which countenances certain methods of scientific investigation (empirical testing, observation etcetera) over others. Very roughly, the former approach will take naturalism about mental illness to be a thesis about the ontology of mental illness and responses to this view are ones that argue mental illness is not a 'real' existent or that it is ontologically dependent on humans. The latter approach represents naturalism about mental illness as the (related but) distinct claim that the study of psychiatry is (or can be) scientifically respectable. Responses to this view argue that the study of mental illness is somehow importantly distinct from the study of medicine or of science more generally.

It is important to note that these two views do not entail one another. However, they have a tendency to be conflated in the literature. For example, the idea that science objective is sometimes taken to imply that its subject matter be realist in some sense.

According to naturalism, the concepts of health and disorder are predominantly driven by objective natural categories, that is, categories that exist independent from our values and interests: biological function and dysfunction. Whilst these categories may interact with values, social considerations, and/or social norms to generate more complex judgments about what conditions qualify for particular social and medical treatment, and whilst values may also play a role in our identification of these categories (Murphy 2009), naturalists maintain that these

values or social norms do not determine what disorder/dysfunction is; nature does. (Wakefield, 1992a, pp. 364–365)

The typical response to the skeptics is to argue that mental disorder is an objective scientific concept, like physical disorder. (Wakefield, 1992a, p. 375)

Here the notion of objectivity is being used to refer to things in the world; Wakefield speaks of the categories and concepts as being objective. But it is quite unclear how to make sense of this notion of objectivity. Objectivity is a property of investigation, not of objects. The best interpretation of claims like this is that to say that some category is objective is to say that it was defined objectively. That's fine as far as it goes but it does not imply realism, which it is taken to imply by many in the debate. This provides further evidence in support of my claim about the inference webs that are routinely employed within the mental illness literature.

5.4 Section Summary

In this section, I have demonstrated several ways in which the naturalism criterion is brought into play in the mental illness literature. There are generally three assumptions that go along with that criterion.

- Ni. Naturalistic investigation is x.
- Nii. Scientific study is always naturalistic.

Moreover, Ni. and Nii. invariably go along with Niii.:

Niii. In order to respond to the worries of Szasz and the anti-psychiatrists we need to establish that mental illness is part of science (i.e. (Legitimate) is true).

As in other cases, what x stands in for varies incredibly from one interlocutor to another, as I have shown above. But our task here is not to ‘solve for x.’ Rather, I wish to draw attention to the problematic ways in which Ni.-Niii. Have been influenced by old empiricist ideals that would likely be rejected in brought into the cold light of day. The inferential webs that the notion of ‘naturalism’ gets tangled up in are both harmful and obscuring. While it’s right to be concerned about just which assumptions we are assuming in the background of psychiatric practice, the remedy to this is not to require that background assumptions themselves be outlawed, or even that non-algorithmic background assumptions be outlawed. Such a thing is neither possible nor desirable. Rather, we should ask which background theories are present, and which are problematic.

5: Conclusion

The mental illness debate centralizes around a single project: that of exposing and avoiding wrongful practice in the psychiatric field. In trying to achieve this aim, a good thing to ask is ‘how can we make the study and treatment of mental illness more scientific?’ Items 1.-4. *can* help with that and do reflect important insights relevant to our main socio-political motives. *However*, 1.-4. Are currently understood as 1N.-4N. which are problematic, potentially false, and misleading due to their inclusion on a problematic inference web. Optimistically, we can save 1-4 if we understand them properly. That is, if

we drop the problematic dichotomy between the natural and the social. Hence, for progress we should:

1. Reject 1N-4N
2. Re-evaluate 1-4
3. Retain (Aim) and (Legitimate) but reject (Division) (at least on the vast majority of understandings)
4. Emphasise and clarify (Aim), putting particular stress on its political and social character.
5. Endorse a version of 1-4 that does not rely on (Distinction); the distinction between the social and the natural.

By their nature, inferential webs of the sort I am targeting are opaque. This makes it especially hard to see the faults within those webs both within disciplines and between them. When a literature operates against a web that is also faulty it can mess up work in the adjacent literatures because the combined usage of these disparate inference webs brings in new relationships under the table. We need a conception of mental illness that moves away from the project of divorcing that notion from ‘the social.’ In this paper, I have suggested a way of reframing the debate that does a better job of getting at what we care about in asking what mental illness is. That the categorization and diagnosis of mental illnesses can be oppressive is legitimate concern, and it is relatively easy to motivate without any of the conceptual work that Szasz does. Mental illnesses are notoriously hard to define and diagnose, they are heavily stigmatized, and most importantly they have a history of oppression. For this reason, we should be extremely

careful with our ascription of mental illnesses. Subsequent writers, however, have (quite correctly) questioned the relevance of the mind-body distinction in motivating this concern.

References

- Agich, G. J. (1983). Disease and value: A rejection of the value-neutrality thesis. *Theoretical Medicine and Bioethics*, 4(1).
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA.
- Andreasen, N. (1997). Linking Mind and Brain in the Study of Mental Illnesses: A Project for a Scientific Psychopathology. *Science*, 275(5306), 1586–1593. <https://doi.org/10.1126/science.275.5306.1586>
- Anscombe, G. E. M. (1957). *Intention* (Vol. 57). Harvard University Press.
- Boorse, C. (1976a). What a Theory of Mental Health should be. *Journal for the Theory of Social Behaviour*, 6(1), 61–84. <https://doi.org/10.1111/j.1468-5914.1976.tb00359.x>
- Boorse, C. (1976b). What a Theory of Mental Health should be. *Journal for the Theory of Social Behaviour*, 6(1), 61–84. <https://doi.org/10.1111/j.1468-5914.1976.tb00359.x>
- Boorse, C. (1977). Health as a Theoretical Concept. *Philosophy of Science*, 44(4), 542–573. <https://doi.org/10.1086/288768>
- Cartwright, S. A. (1851). Diseases and Peculiarities of the Negro Race. *DeBow's Review*, XI.
- Cooper, D. (2001). *Psychiatry and anti-psychiatry* (Repr. of the ed. London 1967). London: Routledge.
- Foucault, M. (1961). *History of Madness*. Routledge.
- Fulford, K. W. (1999). Nine variations and a coda on the theme of an evolutionary definition of dysfunction. *Journal of Abnormal Psychology*, 108(3), 412–420.
- Fulford, K. W. M. (1993). Moral Theory and Medical Practice. *Noûs*, 27(3), 401–403.

- Goffman, E. (1961). *Asylums: Essays on the Social Situation of Mental Patients and Other Inmates*. USA: Anchor Books.
- Goffman, E. (2009). *Stigma: Notes on the Management of Spoiled Identity*. Simon and Schuster.
- Heinrichs, R. W. (2001). *In search of madness: Schizophrenia and neuroscience*. Oxford: Oxford Univ. Press.
- Hempel, C. G. (1965). Fundamentals of Taxonomy. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, 137–154.
- Horwitz, A. V., & Wakefield, J. (2007). *The loss of sadness: How psychiatry transformed normal sorrow into depressive disorder*. Oxford: Oxford University Press.
- Hunt, S. B. (1855). Dr. Cartwright on “Drapetomania.” *Buffalo Medical Journal*, 10, 438–442.
- Kandel, E. R. (2005). *Psychiatry, psychoanalysis, and the new biology of mind* (1st ed). Washington: American Psychiatric Publishing.
- Kendell, R. E. (1986). What are Mental Disorders? In A. M. Freedman, R. Brotman, I. Silverman, & D. Hutson, *Issues in psychiatric classification: Science, practice and social policy* (pp. 23–45). New York: Human Sciences Press.
- Kingma, E. (2007). What is it to be healthy? *Analysis*, 67(2), 128–133.
- Kuhn, T. S. (1970). *The structure of scientific revolutions*. Retrieved from <https://search.library.wisc.edu/catalog/999466601902121>
- Lacan, J. (1934). *De la psychose paranoïaque dans ses rapports avec la personnalité*. Paris: Éditions du Seuil, 1975.

- Lacan, J. (2001). *Ecrits: A selection* (A. Sheridan, Trans.). Retrieved from <http://www.tandfebooks.com/isbn/9780203995839>
- Laing, R. D. (1999). *The Politics of the Family and Other Essays*. <https://doi.org/10.4324/9781351054089>
- Laing, Ronald David. (1990). *The Divided Self (An Existential Study in Sanity and Madness)*. London: Penguin Books.
- Lakatos, I. (1978). *The Methodology of Scientific Research Programmes*. Cambridge University Press.
- Laudan, L. (1980). Progress and Its Problems: Toward a Theory of Scientific Growth. *Erkenntnis*, 15(1), 91–103.
- Lilienfeld, S. O., & Marino, L. (1995). Mental Disorder as a Roschian Concept: A Critique of Wakefield's "Harmful Dysfunction" Analysis. *Journal of Abnormal Psychology*, 104(3), 411–420.
- McHugh, P. R., & Slavney, P. R. M. D. (1999). *The Perspectives of Psychiatry*. Baltimore: Johns Hopkins University Press.
- Millikan, R. G. (1989). In defense of proper functions. *Philosophy of Science*, 56(June), 288–302.
- Murphy, D., & Woolfolk, R. L. (2000). The Harmful Dysfunction Analysis of Mental Disorder. *Philosophy, Psychiatry, & Psychology*, 7(4), 241–252.
- Sadler, J. Z., & Agich, G. J. (1995). Diseases, functions, values, and psychiatric classification. *Philosophy, Psychiatry, and Psychology*, 2(3), 219–231.

- Sarbin, T. R. (1967). On the futility of the proposition that some people be labeled “mentally ill.” *Journal of Consulting Psychology*, 31(5), 447–453.
<https://doi.org/10.1037/h0025018>
- Scadding, J. G. (1967). Diagnosis: The Clinician and the Computer. *The Lancet*, 290(7521), 877–882.
- Scadding, J. G. (1990). The semantic problems of psychiatry. *Psychological Medicine*, 20(2), 243–248. <https://doi.org/10.1017/S0033291700017566>
- Sedgwick, P. (1973). Illness: Mental and Otherwise. *The Hastings Center Studies*, 1(3), 19–40.
<https://doi.org/10.2307/3527464>
- Sedgwick, P. (1982). *Psycho Politics: Laing, Foucault, Goffman, Szasz, and the Future of Mass Psychiatry*. Harper & Row.
- Szasz, T. (1960). The Myth of Mental Illness. *American Psychologist*, (15), 113–118.
- Szasz, T. (1974). *The myth of mental illness : foundations of a theory of personal conduct*. New York: Harper & Row.
- Wakefield, J. (1992a). The concept of mental disorder: diagnostic implications of the harmful dysfunction analysis. *American Psychologist*, 47(3), 373–388.
- Wakefield, J. (1992b). The Concept of Mental Disorder: On the Boundary Between Biological Facts and Social Values. *American Psychologist*, 47(3), 373–388.
- Wakefield, J. (1995). Dysfunction as a value-free concept: A reply to Sadler and Agich. *Philosophy, Psychiatry, and Psychology*, 2(3), 233–246.

Wakefield, J. (2000). Spandrels, Vestigial Organs, and Such: Reply to Murphy and Woolfolk's "The Harmful Dysfunction Analysis of Mental Disorder." *Philosophy, Psychiatry, and Psychology*, 7(4), 253–269.

Wakefield, J. (2009). Mental Disorder and Moral Responsibility: Disorders of Personhood as Harmful Dysfunctions, With Special Reference to Alcoholism. *Philosophy, Psychiatry, & Psychology*, 16(1), 91–99.

Wakefield, J. (2011). Darwin, functional explanation, and the philosophy of psychiatry. In P. R. Adriaens & A. de Block (Eds.), *Maladapting Minds: Philosophy, Psychiatry, and Evolutionary Theory* (pp. 43–172). Oxford: Oxford University Press.