Reflective Equilibrium: A Wittgensteinian Approach

A Doctoral Thesis
by
Jamie Potter
University of East Anglia
August 2010

Acknowledgements

I would like to thank Mr Angus Ross for the patience and support he showed me during initial supervision, as well as members of the Philosophy department at the University of East Anglia, particularly Dr Rupert Read, Prof Catherine Osborne and Dr John Collins for the support and advice they have given me. I would also like to thank Dr Folke Tersman for some very helpful suggestions and criticism he gave via E-mail correspondence, and Dr Max Kölbel for his patient and rigorous examination and criticism, from which I believe I have benefitted enormously. Special thanks must go, however, to Dr Oskari Kuusela, without whom I simply would not have been able to complete this resubmitted version of my thesis.

Abstract

In contemporary moral epistemology, recommending the method of reflective equilibrium in moral reasoning is seldom recognised as a distinct concern from the more typical use of the term 'reflective equilibrium' to refer to the coherence theory of justification in ethics. I argue here that these concerns should be clearly distinguished, since the latter is not a viable approach. The argument proceeds by detaching the *epistemological* import of Wittgenstein's rule-following considerations from Kripkean *constitutive* concerns. In doing so, we recognise an inherent limit to our ability to justify regarding a particular pattern of response to some concept as correct. Nonetheless, as Wittgenstein reminds us, it remains essential to our linguistic practices that, in those cases where we have reached this limit, we accept, without explanation, that some pattern *just is* correct. In certain, problematic, cases of moral disagreement, however, we have likewise exhausted our justificatory powers, and yet our responses to a situation diverge in terms of whether we regard some moral concept (and any principle invoking such a concept) as correctly applicable. Faced with identical situations and given identical rational commitments, agents may nonetheless disagree in what moral judgements they regard as correct, with no method available to settle this problematic disagreement. This, I argue, gives rise to problematic epistemological disagreements in terms of estimations of the coherence of an agent's moral conception, which creates an insuperable difficulty for reflective equilibrium *qua* coherence theory of justification in ethics. Even if we grant that justification in ethics turns on matters of coherence, the coherence theory in ethics is not a viable option. I propose that, if we wish to recommend the *method* of reflective equilibrium, as I believe we should, it pays to do so independently of reflective equilibrium *qua* coherence theory of justification in ethics.

Table of Contents

Introduction 1

Chapter 1 – Reflective Equilibrium 8

1.1 – Overview 8

1.2 – Rawlsian Reflective Equilibrium 10

1.3 – The Standard and Alternative Interpretations 20

1.4 – Brandt's Garbage-in/Garbage-out Objection 25

1.5 – Daniels' Wide Reflective Equilibrium 32

1.6 – The Usefulness Objection 41

Chapter 2 – Explication and Objectivity 47

2.1 – Overview 47

2.2 – Stich on Explication 49

2.3 – Carnap on Explication 54

2.4 – Exactness and Objectivity 62

2.5 – The Explicative Project 72

Chapter 3 – Kripke's Sceptical Paradox 78

3.1 – Overview 78

3.2 – Meaning Factualism 80

3.3 – The Sceptical Paradox 87

3.4 – Arguments Against Dispositionalism 94

3.5 – The Sceptical Solution 102

Chapter 4 – Rule-Following and Explication 109

4.1 – Overview 109

4.2 – Rational Demonstrability 112

4.3 – Wright on Basic/Non-Basic Rule-Following 118

4.4 – Wright's Wittgenstein and *Prima Facie* Legitimacy 124

4.5 – Rational Demonstrability and Explication 133

4.6 – *Prima Facie* Legitimate Explication 140

Chapter 5 – Explicating 'Being Justified' in Ethics 148

5.1 – Overview 148

5.2 – Sensitivity 150

5.3 – Example 1: Eating Meat 159

5.4 – Example 2: Pornography 166

5.5 – Against the 'Theoretical Presumption' 173

5.6 – Reflective Equilibrium: Explicative Failure 180

Chapter 6 – The Alternative Approach 187

6.1 – The Distinctiveness of the Argument 187

6.2 – The Anti-Theory Worry 195

6.3 – The Objectivity Worry 204

6.4 – The Vacuity Worry 212

Glossary 219

Bibliography 221

Introduction

The notion of 'reflective equilibrium' derives from John Rawls' 1971 work, *A Theory of Justice*. (Rawls 1999) The idea has attracted considerable scrutiny and interest, and remains of significant interest to philosophers, enjoying especial prominence, orthodoxy even, in applied ethics, bioethics in particular.[1] Whilst most of this interest no doubt derives from the widespread influence of Rawls' classic work, the work of Norman Daniels (himself an active contributor in bioethics) on *wide* reflective equilibrium[2] in the late 1970s and 1980s was instrumental in sustaining interest in the idea. A little more recently, book-length treatments by Michael R. DePaul (DePaul 1993) and Folke Tersman (Tersman 1993) point to its enduring appeal, and current debates concerning the use of 'intuitions' in ethics (Singer 2005; Tersman 2008) have taken the form of concentrating on the merits of reflective equilibrium methodology in particular.

'Reflective equilibrium' may be held to refer to a certain *methodology* which, to use Daniels' most recent and authoritative description, "Consists in working back and forth among our considered judgments (some say our "intuitions") about particular instances or cases, the principles or rules that we believe govern them, and the theoretical considerations that we believe bear on accepting these considered judgments, principles, or rules, revising any of these elements wherever necessary in order to achieve an acceptable coherence among them." (Daniels 2009) In addition, 'reflective equilibrium' may be held to refer to a *coherence theory of justification* whereby "a moral principle or moral judgment about a particular case (or, alternatively, a rule of inductive or deductive inference or a particular inference) would be justified if it cohered with the rest of our beliefs about right action (or correct inferences) on due reflection and after appropriate revisions throughout our system of beliefs." (Daniels 2009) One may subscribe to either commitment without necessarily subscribing to the other.[3]

---

[1] The title of John Arras' introductory piece 'The Way We Reason Now: Reflective Equilibrium in Bioethics' (Arras 2007) is in itself instructive. Arras maintains that "In the world of bioethics, the air is abuzz with reflective equilibrium." (Arras 2007, p.46)

[2] See §1.5 below.

[3] As we shall see in §1.3, one can regard being in a state of reflective equilibrium in holding some moral conception (or set of inferential rules), i.e. being in a state of coherence, as constitutive (in some way) of justification, without thereby thinking that adopting the method of reflective equilibrium is the

The interest in reflective equilibrium is itself bifurcated as a result of deriving from these two distinct concerns. On the applied, normative side of things, there those who are interested in reflective equilibrium as a method of moral or ethical[4] theory construction.[5] For such philosophical ethicists, reflective equilibrium is introduced as a way of retaining traditional, principle-based approaches, whilst incorporating lessons from alternative approaches emphasising the importance of the particular, or the centrality of narratives in ethical reasoning, or lessons garnered from broader scientific and social scientific enquiry. (Arras 2007, pp.48-9) The method of reflective equilibrium represents a broad methodological consensus from which an ethicist can go about constructing a moral theory in a particular area. It is, in short, a methodological recommendation.

On the more abstract, theoretical side of things, philosophers are interested in reflective equilibrium as a coherence theory of justification (in ethics).[6] The interest here is (moral) epistemological in nature: philosophers are interested in the notion of 'reflective equilibrium' inasmuch as it is intended to tell us something[7] about the nature of justification in ethics.[8] Reflective equilibrium is of interest because it may

---

best way to achieve such a state of coherence. It may be that adopting some other methodology is better for arriving at the state of reflective equilibrium. Consequently, adopting reflective equilibrium *qua* coherence theory does not entail adopting reflective equilibrium *qua* methodological recommendation.

One can also regard adopting the method of reflective equilibrium as worthwhile without thereby believing that the end-products of following such a methodology are constitutive (in some way) of being justified in one's views. One might, for instance, simply take that adopting the method of reflective equilibrium *improves* one's views. Consequently, adopting reflective equilibrium *qua* methodological recommendation does not entail adopting reflective equilibrium *qua* coherence theory.

[4] I shall use the terms 'ethical' and 'moral' interchangeably.

[5] See, for instance, Martin Benjamin's recent *Philosophy and This Actual World* (Benjamin 2003), or Stephen Cohen's *The Nature of Moral Reasoning* (Cohen 2006). Daniels, as is implied by the title of his collected papers, *Justice and Justification, Reflective Equilibrium in Theory and Practice,* (Daniels 1996) likewise regards his own work in applied ethics as that of putting the method of (wide) reflective equilibrium into practice. (c.f. Daniels 1996a)

[6] Opinion is somewhat divided on whether 'reflective equilibrium' is intended as a general coherence theory of justification, or whether its domain is restricted to ethics. In what follows, I follow the narrower interpretation and restrict reflective equilibrium to ethical concerns. See also ft.23.

[7] In this introduction and in chapter 1, I shall of necessity employ the somewhat awkward and vague locution 'tells us something about'. The reason I do this is because it is a delicate question, one I delay until chapter 2, as to what thinking of reflective equilibrium as a coherence theory of justification in ethics actually commits a person. For instance, a natural and appealing way of understanding reflective equilibrium *qua* coherence theory of justification (in ethics) is to regard 'reflective equilibrium' is an *analysis*, in the traditional mould, of what it is for an agent to be justified in holding some moral conception. Here, to hold a moral conception in reflective equilibrium would *constitute* being justified in holding that moral conception.

[8] Representative examples here are Tersman's *Reflective Equilibrium: An Essay in Moral Epistemology* (Tersman 1993) and DePaul's *Balance and Refinement.* (DePaul 1993) See the beginning of §1.4 for a

well serve to clarify and expand upon our understanding of the nature of justification in ethics. Potentially, this is useful as a tool for settling questions of moral theory acceptance, e.g. for discounting moral theories that may not be justifiably held.

Although these theoretical and normative concerns are distinct, they are typically run together. Reflective equilibrium is referred to as a 'coherence method' of justification in ethics,[9] for instance. This locution nicely captures what I take to be the 'standard' view,[10] well-represented by Daniels' own work, that reflective equilibrium is primarily of interest as a coherence theory. That is, reflective equilibrium describes an epistemic state of coherence in one's overall moral conception,[11] which, on the coherence theory of justification, tells us something about whether one is justified in holding that moral conception. On the standard view, one engages in the method of reflective equilibrium only because one seeks the epistemic state of reflective equilibrium. The 'alternative' view, by contrast, is to regard the method of reflective equilibrium as of value in its own right, regardless of whether it leads to such an epistemic state. One would advocate reflective equilibrium as a distinctive method independently of whether the products of following such a method represent a state of coherence in one's overall moral view.

My principle contention in this thesis is that it is a good thing that reflective equilibrium *qua* methodological recommendation for moral reasoning and reflective equilibrium *qua* coherence theory of justification in ethics reflect distinct concerns. This is because the latter involves commitments that cannot, given the nature and prevalence of certain problematic kinds of moral disagreement, be upheld. This leaves us free, however, to pursue reflective equilibrium as a methodological

---

comprehensive list of philosophers who have taken reflective equilibrium to be a coherence theory of justification (in ethics and elsewhere).

[9] See, for instance, (Sayre-McCord 1996, p.141), or (DePaul 1993, p.13).

[10] See §1.3 below for the distinction between the 'standard' and 'alternative' understandings of reflective equilibrium, distinguishable by whether there is an assumed priority, as there is under the 'standard' view, of reflective equilibrium as primarily picking out a distinctive epistemic state of coherence, at which point one recommends the *method* of reflective equilibrium insofar as it leads the moral enquirer to attain such a state.

[11] In what follows, I use the terms 'moral conception' and 'moral theory' interchangeably. Thus, 'moral theory' is not intended to pick out some technical notion, but is rather to be understood loosely as a more or less interrelated set of moral principles and judgements at varying levels of generality. On occasion, it seems more natural to utilise the term 'moral conception' instead, but I do not mean to thereby move the argument in any direction. 'Moral conception' is likewise intended to pick out this idea. The central requirement here is that we understand by these terms the sort of thing required to form the 'principles' part of the coherent triad of principles, judgements and background theory in reflective equilibrium.

recommendation in its own right, and to regard the method of reflective equilibrium as nonetheless important and relevant to contemporary moral epistemology. My claim, then, is that, if one is at all interested in the method of reflective equilibrium, as I believe one ought to be, one would be well-advised to avoid recommending that method on the basis of its role in attaining an epistemic state of reflective equilibrium, i.e. coherence in one's moral conception. One should reject the standard approach to reflective equilibrium, and pursue an alternative understanding in which one recommends the method of reflective equilibrium for reasons unconnected to the role of reflective equilibrium *qua* coherence theory of justification in ethics.

I shall now give some broad indications as to what to expect in what follows by describing the distinctive nature of my argument and outlining the general structure of the thesis. Concerning the distinctive nature of the argument, there are three considerations I shall briefly explain: firstly, that my argument is directed against the viability of reflective equilibrium understood as a coherence theory in *ethics*, as opposed to a general coherence theory; secondly, that my argument against reflective equilibrium *qua* coherence theory (in ethics) takes a different tack to arguments normally employed in that I already grant the approach a certain degree of latitude it does not usually receive; thirdly, that the argument traces connections from Wittgenstein's 'rule-following considerations'[12] to within the moral realm along a path that is not entirely untrodden, but goes further in noting *moral epistemological* implications, i.e. for reflective equilibrium *qua* coherence theory of justification in ethics.

Coherence theories of justification in general epistemology have, arguably, been about since the advent of logical positivism, but have attracted considerable attention (and criticism) since the 1970s following the work of Laurence Bonjour, Gilbert Harman, and Keith Lehrer. (Bonjour 1976; Harman 1974; Lehrer 1974) Evidently, what follows cannot be a mere rehearsal of traditional complaints about general coherence theories of justification, such as the 'input problem', the complaint that one might have a coherent set of beliefs "In spite of being utterly out of contact with the world that it purports to describe", (Bonjour 1985, p.108) rendering the thought

[12] Typically, this refers to §§138-242 of Wittgenstein's *Philosophical Investigations*. (Wittgenstein 2001, §§138-242)

that coherence is a theory of *justification* somewhat suspect.[13] Instead, the argument I put forward targets the idea of having a coherence theory of justification *in ethics*. I make no claims about the prospects of maintaining such a theory in epistemology more generally. The claim is that a standard reflective equilibrium approach, in which reflective equilibrium functions primarily *qua* coherence theory of justification in ethics, is constrained by certain commitments that we cannot hope to satisfy. It will remain an open question as to whether analogous commitments could be satisfied in other areas.

This relates to the second consideration to bear in mind considering the distinctive nature of my argument. As my argument is directed specifically against reflective equilibrium as a coherence theory of justification in ethics, I shall argue that there is something about the moral sphere, at least in its *de facto* state, that renders it uncongenial to coherence considerations, and this means that I'm already granting the standard reflective equilibrium approach a certain degree of latitude. Even if one grants that justification in ethics is usefully understood by thinking about the coherence of an agent's moral conception, I shall demonstrate that the coherence theory is nonetheless not viable. As such, I shall not question what is generally regarded as contentious: the underlying claim that coherence considerations help us to understand the nature of justification, in ethics or elsewhere.

Another way to put essentially the same point is that my argument takes place further down the road than existing objections to reflective equilibrium *qua* coherence theory of justification in ethics. In what follows, I shall assume that there is no difficulty as such with the idea that reflective equilibrium *qua* coherence theory of justification tells us something about the nature of justification in ethics, and that, for the sake of my argument, one may take this road with impunity. The problem occurs, however, when it comes to understanding what being in an epistemic state of coherence amounts to when it comes to ethical cases. The problem is that the moral sphere, at least in its current state, is uncongenial for a coherence approach.

[13] I shall, however, consider an analogue of this objection for *moral* cases in §1.4: the 'garbage-in/garbage-out objection', as it is an established line of criticism for reflective equilibrium as a coherence theory.

Evidently, much turns on establishing that there is something about moral cases that poses a problem for coherence considerations. It is here that the third consideration concerning the distinctive nature of my approach is relevant. Existing commentators, notably John McDowell and, more recently, Alice Crary,[14] have argued that Wittgenstein's rule-following considerations establish the need for us to take seriously the idea of a 'sensitivity', a notion I can only vaguely describe in advance of substantive investigation as referring to language-users' sets of responses to concepts that are not inferred from applying further rules, but are taken as 'basic' in some way, and that the existence of such 'sensitivities' has a serious impact on how we ought to view the moral realm in general. What is distinctive about my argument here is that I shall demonstrate that it also has a serious impact on how we ought to think about reflective equilibrium.

These remarks shall, of course, become clearer over the course of the thesis. I conclude this introduction with an indication of the broad structure of what follows. The thesis comprises of three parts. The task of the first part, undertaken in chapters 1-2, is to clarify the *target* of my argument. More specifically, its function is to, having clarified the distinction between the standard and alternative approaches to reflective equilibrium, characterise the standard approach to reflective equilibrium in a manner that models, in a sufficiently charitable way, the necessary commitments that a would-be standard reflective equilibrium theorist would need to uphold if they are to continue to regard their approach as viable.

The task of the second part, which comprises the bulk of the thesis, chapters 3-5, is to make the case that the standard reflective equilibrium approach is misguided, since a coherence theory is not viable for ethical cases. Within this broad division, a sub-division may be helpful: Chapters 3 and 4 are concerned with drawing out what I take to be the Wittgensteinian lesson, under Wright's recent interpretation, (Wright 2007) to draw from the Kripke's sceptical paradox/the rule-following considerations. Chapter 5 is concerned with illustrating how this bears on the moral realm in particular, and then the implications for reflective equilibrium, i.e. that a coherence theory is not viable in ethics.

[14] e.g. (McDowell 1979; Crary 2007a). See §5.2 below.

The task of the third part of the thesis is to make it clear that adopting the *method* of reflective equilibrium remains a live option in spite of the insuperable difficulty that besets the standard approach, regarding reflective equilibrium primarily *qua* coherence theory of justification in ethics. I develop, largely in response to potential worries one might have about my argumentative position, my own alternative interpretation of the merits of the method of reflective equilibrium. In giving an outline of a potentially fruitful approach to the method of reflective equilibrium, the aim is to reinforce my principle contention: if one is interested in the method of reflective equilibrium, as I maintain one ought to be, it is well to avoid recommending the method in virtue of its reliably leading to an epistemic state of coherence in one's moral conception.

Chapter 1 – Reflective Equilibrium

1.1 – Overview

This chapter is a selective survey of the philosophical literature concerning reflective equilibrium. Aside from the obvious consideration that such a survey is necessary before one discusses any topic in philosophy, it also fulfils two more specific aims. Firstly, in terms of the 'standard'/'alternative' distinction, I shall illustrate that commentators have overwhelmingly assumed that the method of reflective equilibrium is to be recommended only insofar as it would reliably lead to an epistemic state of reflective equilibrium (i.e. coherence in one's moral view), which in turn tells us something about whether an agent is justified in holding their moral conception. In demonstrating that commentators have overwhelmingly assumed that reflective equilibrium functions primarily as a coherence theory of justification in ethics, I shall show that what I'm referring to as the 'standard interpretation' of reflective equilibrium really is the standard view. Secondly, following the shape of the philosophical discussion, in particular two salient lines of criticism that have emerged regarding reflective equilibrium, enables us to see the constraints it is reasonable to impose on reflective equilibrium *qua* coherence theory of justification in ethics.

To these ends, I begin with Rawls' exposition of the notion of 'reflective equilibrium', removing four relevant points of unclarity concerning the notion so that they do not interfere with subsequent discussion. These unclarities are: 1. concerning the role of the 'original position' within reflective equilibrium, 2. whether reflective equilibrium is intended as restricted to ethical domains or applies more generally, 3. whether the method of reflective equilibrium is a matter of mere description or whether it is a genuinely justificatory method, and, most crucially, 4. how we should understand Rawlsian reflective equilibrium in terms of the standard/alternative distinction. To emphasise the importance of this fourth consideration, I address it in a separate section (§1.3).

I then examine, in turn, two criticisms of reflective equilibrium that illustrate the centrality of reflective equilibrium *qua* coherence theory of justification in ethics. The first criticism, the garbage-in/garbage-out objection, is urged upon us by Richard

Brandt and Peter Singer. This line of criticism bears upon the arguably egregious reliance on ethical 'intuitions' within reflective equilibrium, and runs to the effect that a set of ethical 'intuitions' held in a state of coherence is no more justified than a set of ethical 'intuitions' not held in a state of coherence. I then consider Daniels' response, where we shall see how a *wide* reflective equilibrium account can counter such criticism.

This form of response, however, would appear to leave Daniels' wide reflective equilibrium vulnerable to another line of criticism, 'the usefulness objection'. This line of criticism centres upon Daniels' desire for the notion of 'wide reflective equilibrium' to render questions of moral theory acceptance *more tractable*, and yet, at least on his formulation, it is highly unlikely, due to immense complexity of the considerations involved, that one could ever *ascertain* that an agent was in a state of wide reflective equilibrium. Here I shall argue that, whilst the objection helps us to clarify the commitments Daniels' approach needs to satisfy, the criticism itself is not as damaging as it first appears.

What shall emerge is an preliminary sketch of the commitments that a standard approach to reflective equilibrium needs to satisfy. My intention over this chapter, and the next, is to develop a charitable understanding of to what one is committed in holding the standard view of reflective equilibrium as functioning primarily *qua* coherence theory of justification in ethics. By the end of chapter 2, we shall be in a position to see that the standard view of reflective equilibrium involves a commitment to the viability of a certain 'explicative project' for the would-be standard reflective equilibrium theorist. This 'explicative project', as I shall argue in chapters 3-5, is, unfortunately, not viable. However, this argument is not usefully surveyed in advance. For the time being, we need to concentrate solely upon getting clear on our *target*: the 'standard' understanding of reflective equilibrium.

1.2 – Rawlsian Reflective Equilibrium

We start then, with Rawls' presentation of reflective equilibrium found within *A Theory of Justice*.[15] (Rawls 1999) As Thomas Scanlon points out, (Scanlon 2003, p.139) Rawls' use of reflective equilibrium is actually one of three important contributions he makes to the nature of justification in the area of moral and political reasoning throughout his philosophical career, the other two being the 'original position',[16] and 'public reason'.[17] Indeed, Rawls introduces the notion of reflective equilibrium within the context of a discussion of the original position, a hypothetical situation designed to model *fair* conditions for the selection of regulative principles of justice for a 'well-ordered society'.[18] The idea is to come up with a description of the original position such that any regulative principles of justice that would be chosen in such a situation

---

[15] It is worth noting here that Rawls' presentation of reflective equilibrium in that work is anticipated to an extent by his 1951 paper, 'Outline of a Decision Procedure for Ethics', (Rawls 1951) originally part of Rawls' PhD dissertation at Princeton. This paper concerns the design of a decision-procedure for the selection of moral rules in a specific area, whereby one selects a class of *competent* (intelligent, knowledgeable, reasonable, and sympathetic) moral judges who make *considered moral judgements* on a range of non-hypothetical moral situations, and then attempt to 'explicate' those considered moral judgements via constructing an explicit set of moral principles, which if followed correctly, would result in those particular considered moral judgements, at least approximately.

[16] The original position is a theoretical device utilised by Rawls in *A Theory of Justice*. Rawls asks us (Rawls 1999, pp.102-30) to imagine ourselves placed behind a 'veil of ignorance' such that we are deprived of knowing individuating facts about ourselves (e..g our position in society, job, ethnic origin, what conception of the good we happen to hold, etc.), and then select a system of public principles of justice (from a list of alternatives) to effectively regulate a 'well-ordered society' (see ft.18). Rawls' claim is that the procedural constraints imposed by the original position thereby "represent equality between human beings as moral persons, as creatures having a conception of the good and capable of a sense of justice." (Rawls 1999, p.17)

[17] The notion of 'public reason' is set out in *Political Liberalism* (Rawls 1996), and in the subsequent essay 'The Idea of Public Reason Revisited' (Rawls 1997). Essentially, the thought is that reasonable citizens in a pluralistic modern democracy, can, through the exercise of public reason achieve a point of agreement on constitutional essentials. (Larmore 2003, pp.380ff.) The notion is connected to a 'political' conception of *public justification*, where the emphasis is on finding a conception of 'justice' such that it can reasonably form the basis of a modern constitutional arrangement due to it being a point of 'overlapping consensus' amongst reasonable conceptions of the good. Through these ideas of public reason and public justification, Rawls' concern is not with justification *simpliciter*, but rather justification in the sense of *best suited* to a particular purpose, that of "finding a basis for public agreement" (Rawls 1996, p.9).

Joseph Raz charges that the underlying notion of 'public justification' amounts to a case of 'epistemic abstinence', since it seems to confuse *justification* with finding a pragmatic *modus vivendi*. (Raz 1990) For an excellent discussion of these issues, see Burton Dreben's 'On Rawls and Political Liberalism'. (Dreben 2003) It is beyond our scope, however, to examine the notion of 'public reason' here.

[18] The notion of a 'well-ordered society', whilst present in *A Theory of Justice* (Rawls 1999, p.397), is set out more clearly in 'Fairness to Goodness' (Rawls 1975a), as one in which:

"(1) Everyone accepts, and knows that others accept, the same principles (the same conception) of justice.
(2) Basic social institutions and their arrangement into one scheme (the basic structure of society) satisfy, and are with reason believed by everyone to satisfy, these principles.
(3) The public conception of justice is founded on reasonable beliefs that have been established by generally accepted methods of inquiry." (Rawls 1975a, p.278)

will be justified. Rawls is explicit, however, that the justification of the original position takes place within reflective equilibrium:

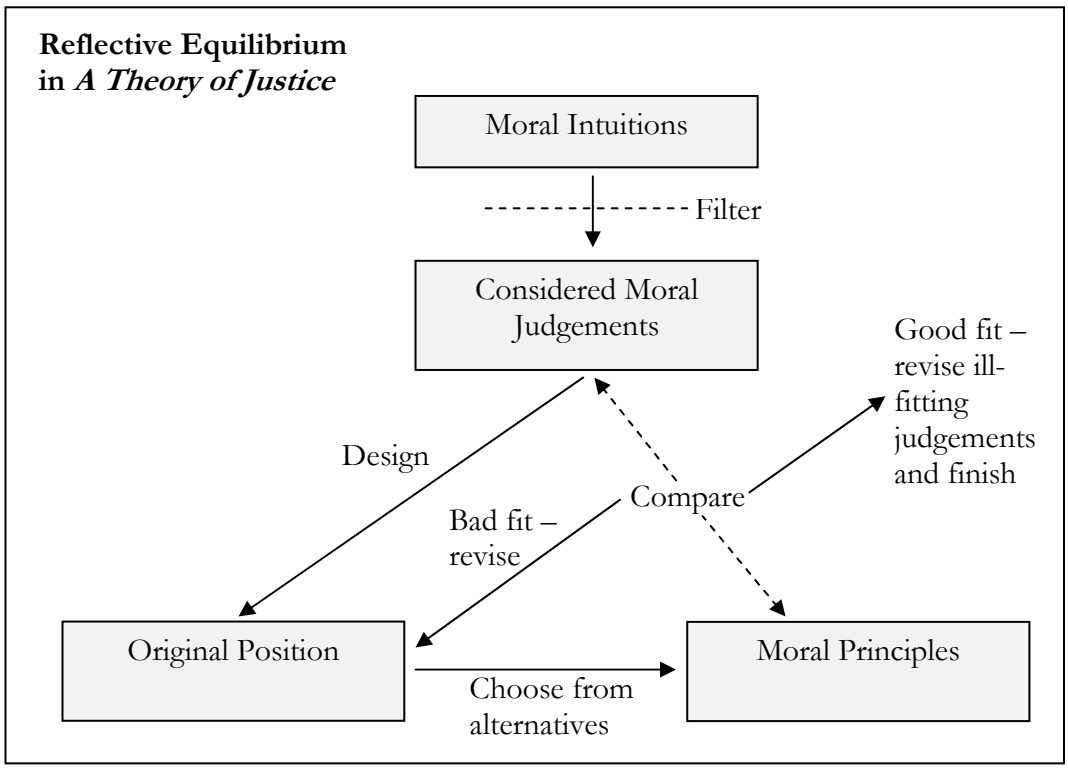
> In searching for the most favoured description of this situation we work from both ends. We begin by describing it so that it represents generally shared and preferably weak conditions. We then see if these conditions are strong enough to yield a significant set of principles. If not, we look for further premises equally reasonable. But if so, and these principles match our considered convictions of justice, then so far well and good. But presumably there will be discrepancies. In this case we have a choice. We can either modify the account of the initial situation or we can revise our existing judgements, for even the judgements we take provisionally as fixed points are liable to revision. By going back and forth, sometimes altering the conditions of the contractual circumstances, at others withdrawing our judgements and conforming them to principle, I assume that eventually we shall find a description of the initial situation that both expresses reasonable conditions and yields principles which match our considered judgements duly pruned and adjusted. This state of affairs I refer to as reflective equilibrium. It is an equilibrium because at last our principles and judgments coincide; and it is reflective since we know to what principles our judgments conform and the premises of their derivation. (Rawls 1999, p.18)

Rather than attempting to design the original position purely from first principles and then derive whatever further principles and judgements follow from this description, what we might call a 'top-down' approach,[19] Rawls allows the design of the original position to be influenced by considerations of whether any resultant principles or judgements match our intuitive understanding of distributive justice. In Rawlsian reflective equilibrium, we begin with 'considered moral judgements', a set of judgements that survive the filtering out of ill-considered or rash judgements that, he argues, poorly reflect our moral sense.[20] We then take these judgements and attempt to formulate a situation, the original position, in which we select principles of justice designed to regulate a well-ordered society. We then compare the outcome of applying such principles in particular cases (i.e. resultant moral judgements) with what we would intuitively think in ordinary cases. Should we achieve a reasonable fit,

[19] Stephen Cohen utilises a helpful distinction between 'top-down' reasoning, where "the first principles of moral reasoning are general or universal moral principles, which can be applied to specific situations" (Cohen 2006, pp.59-60), and 'bottom-up' reasoning, where "the first principles of moral reasoning are… the moral judgements we make." (Cohen 2006, pp.61-2) The method of reflective equilibrium represents a mixture of, and balance between, these two basic modes of reasoning in ethics.

[20] "We can discard those judgements made with hesitation, or in which we have little confidence... Considered judgements are simply those rendered under conditions favourable to the exercise of the sense of justice, and therefore in circumstances where the more common excuses and explanations for making a mistake do not obtain." (Rawls 1999, p.42)

we then accept the theory, revising any ill-fitting judgements to accord with the theory. If a reasonable fit is not achieved in that round, we return to the description of the initial situation, make adjustments, and repeat the process. I represent the matter visually below:



There are four points of unclarity regarding Rawlsian reflective equilibrium that require some attention before we can advance the discussion. The first concerns the role of the original position within reflective equilibrium, the second concerns whether reflective equilibrium is to be utilised solely in *ethical* domains, and the third concerns whether we ought to regard the method as *descriptive* or '*deliberative*'.[21] The fourth, and most important for our purposes, concerns the connection between reflective equilibrium *qua* methodological recommendation and reflective equilibrium *qua* coherence theory of justification. I shall address the first three points in the rest of this section. Since the last point is of central importance to the thesis, I address it in a separate section below.

[21] This distinction I borrow from Scanlon (Scanlon 2003). A descriptive interpretation of reflective equilibrium would see it as working towards setting out how, as a matter of fact, people are inclined to judge on certain moral matters, i.e. a kind of moral psychology; a deliberative interpretation sees seeking reflective equilibrium as a matter of working out exactly what one *ought* to think about moral matters.

The first point of unclarity may be settled relatively easily, as it is by and large a terminological issue. One can see that the design of an 'initial situation' or 'original position' for the construction of principles of justice is part of the overall reflective equilibrium procedure Rawls utilises in *A Theory of Justice*, but it seems clear from ancillary remarks that this is merely a feature of the justificatory task undertaken within that work. The reason the original position is used is because it is crucial to Rawls' basic idea of *justice as fairness*, the idea that justified principles of justice are those that are selected under fair conditions: "One conception of justice is more reasonable than another, or justifiable with respect to it, if rational persons in the initial situation would choose its principles over those of the other for the role of justice." (Rawls 1999, pp.15-6) This suggests that it is due to the peculiarities of Rawls' particular theory that the original position is included, not because using the original position is integral to the method of reflective equilibrium. The original position is simply a device used "to make vivid to ourselves the restrictions that it seems reasonable to impose on arguments for principles of justice, and therefore on these principles themselves." (Rawls 1999, p.16) It seems appropriate, then, to think of Rawlsian reflective equilibrium as not requiring the use of such a device, though of course any such device may be happily incorporated within reflective equilibrium if it helps in the formulation of acceptable principles.

The second issue, whether reflective equilibrium is restricted to ethical domains or if it has broader scope, is also largely terminological. As Rawls indicates in an important footnote (Rawls 1999, p.18n.7), the notion of 'reflective equilibrium' borrows heavily from Nelson Goodman's *Fact, Fiction and Forecast* (Goodman 1983),[22] indicating that the scope of the method of reflective equilibrium may potentially extend beyond ethical considerations to, in this instance, formulating justifiable rules of inference. Rawls is careful, however, not to actually make such a claim,[23] and thus it seems appropriate to interpret him in this more cautious vein. Henceforth, then, I shall

[22] Goodman's project in this work is the creation of justified inductive rules of inference. He suggests the following procedure: "*A rule is amended if it yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend.* The process of justification is the delicate one of making mutual adjustments between rules and accepted inferences; and in the agreement achieved lies the only justification needed for either." (Goodman 1983, p.64)

[23] Here, Samuel Freeman (Freeman 2007, p.31, p.36, p.41) argues persuasively that, whilst Rawls is generous to acknowledge his indebtedness to Goodman, it is a mistake to assume that 'reflective equilibrium' is intended as a general epistemological theory of justification for all areas of enquiry. This commitment is, by the lights of *A Theory of Justice,* optional: one *may* extend it in that way if one wishes, but Rawls should be understood as only requiring the more restricted claim that the method of reflective equilibrium is appropriate specifically to moral enquiry.

interpret the notion of 'reflective equilibrium' as bearing only on ethical considerations.[24]

The third point of unclarity is less easily settled, however, and it concerns whether reflective equilibrium is merely a matter of *description*. One salient criticism of the methodology of *A Theory of Justice*, courtesy of Richard Hare, runs as follows:

> If (as will certainly be the case) he finds a large number of readers who can share with him a cosy unanimity in their considered judgements, he and they will think that they adequately represent 'people generally', and congratulate themselves on having attained the truth. This is how phrases like 'reasonable and generally acceptable' are often used by philosophers in lieu of argument.
>
> Rawls, in short, is here advocating a kind of subjectivism, in the narrowest and most old-fashioned sense. He is making the answer to the question 'Am I right in what I say about moral questions?' depend on the answer to the question 'Do you, the reader, and I agree in what we say?' (Hare 1973, p.82)

It is one thing, argues Hare, to characterise adequately the 'moral sense' of a particular moral evaluator, but this is entirely a distinct question from supplying a justified moral theory concerning what one *ought* to think about some particular moral matter. According to Hare, Rawls conflates the former, descriptive task, for the latter, normative/deliberative task, and that it is the latter task that ought to be regarded as the preserve of moral philosophy.

Hare's interpretation of Rawls' methodology is not without textual support, both within *A Theory of Justice* and in the subsequent paper 'The Independence of Moral Theory' (Rawls 1975).[25] It may well appear that Rawls' use of reflective equilibrium is

---

[24] I say this with an important proviso: I do not take restricting the notion of 'reflective equilibrium' to ethical domains to imply that it is not worth considering criticism of a reflective equilibrium-like approach in other areas. As we shall see when we examine some of Stich's remarks concerning a 'neo-Goodmanian' project of justifying inferential rules via demonstrating that they are upheld in 'reflective equilibrium', there are certain problems with the notion of 'reflective equilibrium' in the area of inference that can be made with equal force within the area of moral enquiry. Provided we are not broaching some criticism that is obviously inappropriate to moral enquiry, there is no difficulty in canvassing criticism that is directed at a 'neo-Goodmanian' approach for inferential rules.

[25] In this paper, Rawls lays out a distinctive conception of 'moral theory' that he finds in Henry Sidgwick's *The Method of Ethics* (Sidgwick 1981) whereby 'moral theory' is to be distinguished from promoting and/or justifying a particular moral conception:

> "[In moral theory] one tries to find a scheme of principles that match people's considered judgments and general convictions in reflective equilibrium. This scheme of principles represents their moral conception and characterizes their moral sensibility. One thinks of the moral theorist as an observer, so to speak, who seeks to set out the structure of other people's moral conceptions and attitudes... We may also include ourselves, since we are ready to hand for detailed self-examination.

to be understood solely as a kind of moral psychology when one encounters remarks such as the following:

> Now one may think of moral theory at first (and I stress the provisional nature of this view) as the attempt to describe our moral capacity; or, in the present case, one may regard a theory of justice as describing our sense of justice. By such a description is not meant simply a list of the judgments on institutions and actions that we are prepared to render, accompanied with supporting reasons when these are offered. Rather, what is required is a formulation of a set of principles, which, when conjoined to our beliefs and knowledge of the circumstances, would lead us to make these judgements with their supporting reasons were we to apply these principles conscientiously and intelligently. A conception of justice characterizes our moral sensibility when the everyday judgements we do make are in accordance with its principles. (Rawls 1999, p.41)

This aim of 'characterising our moral sensibility' implies a detached, neutral vantage-point from where one observes and systematises people's considered moral judgements. From here, one can propose particular conceptions of justice, say, that could, as a matter of fact, plausibly serve as a basis of agreement on broad constitutional principles. This reading of reflective equilibrium as merely descriptive moral psychology is further encouraged by Rawls' invocation of a comparison between the method of reflective equilibrium in ethics and methods in descriptive linguistics:[26]

> A useful comparison here is with the problem of describing the sense of grammaticalness that we have for the sentences of our native language. In this case the aim is to characterise the ability to recognised well-formed sentences by formulating clearly expressed principles which make the same discriminations as the native speaker. This undertaking is known to require theoretical constructions that far outrun the ad hoc precepts of our explicit grammatical knowledge. A similar situation presumably holds in moral theory. There is no reason to assume that our sense of justice can be

---

But in studying oneself, one must separate one's role as a moral theorist from one's role as someone who has a particular conception. In the former role we are investigating an aspect of human psychology, the structure of our moral sensibility; in the latter we are applying a moral conception, which we may regard (though not necessarily) as a correct theory about what is objectively right and wrong." (Rawls 1975, p.288)

[26] One might think here that Rawls has in mind *grammarians*, who genuinely do issue prescriptions on correct use of grammar, such as the prescriptions not to wantonly split infinitives or end sentences with prepositions that people normally make use of. However, given his explicit reference (Rawls 1999, p.41n.25) to Chomsky's *Aspects of the Theory of Syntax,* in which Chomsky is explicit that his generative approach to linguistics "Is not a model for a speaker or a hearer… [and] attempts to characterise in the most neutral possible terms the knowledge of the language that provides the basis for actual use of language by a speaker-hearer," (Chomsky 1965, p.9) it is clear that Rawls intends to compare reflective equilibrium with descriptive linguistics.

> adequately characterised by familiar common sense precepts, or derived from the more obvious learning principles. A correct account of moral capacities will certainly involve principles and theoretical constructions which go much beyond the norms and standards cited in everyday life; it may eventually require sophisticated mathematics as well. (Rawls 1999, pp.41-2)

The implication of the comparison seems fairly clear: just as the linguist attempts to *describe* our grammatical competence via collecting linguistic judgements (subject to certain filters to remove performance errors), and then creating systematic grammatical rules that account for those judgements, the moral theorist attempts to *describe* our moral competence via collecting moral judgements (again, subject to certain filters), and then creating systematic moral rules that account for those judgements.[27] At this point, Hare's criticism looks pressing: characterising what a person *happens to believe* in terms of their moral conception does not thereby *justify* those moral beliefs, or demonstrate them to be true. It does not do so even where one establishes that lots of people happen to believe the same thing. Rawls appears to conflate consensus with justification.

However, I do not believe that it is, on balance, a fair representation of Rawls' characterisation of reflective equilibrium to regard it solely as a method of moral psychology, or a matter of discovering an underlying 'moral grammar' governing our considered moral judgements. One ought to note the various caveats Rawls utilises in stressing the descriptive aspect of the method of reflective equilibrium: "Now one *may* think of moral theory *at first* (and *I stress the provisional nature of this view*) as the attempt to describe our moral capacity... A *useful* comparison here is with the problem of describing the sense of grammaticalness that we have for the sentences of our native language." (Rawls 1999, p.41, my italics) Even in 'The Independence of Moral Theory', where 'moral theory' is explicitly characterised as "a kind of psychology" (Rawls 1975, p.290), Rawls stresses that "*provisionally* we may bracket the problem of moral truth and turn to moral theory." (Rawls 1975, p.288, my italics) These caveats strongly suggest that Rawls' developed position is to regard 'moral

[27] This is precisely how Hauser characterises his 'Rawlsian creature', representing a distinctive approach to moral thinking, in his recent *Moral Minds*:

"In the same way that grammaticality judgments emerge from a universal grammar of principles and parameters, the Rawlsian creature's ethicality judgments would emerge from a universal moral grammar, replete with shared principles and culturally switchable parameters." (Hauser 2006, p.43) Hauser even interprets Rawls' work as an attempt at *causal explanation*, whereby the aim is to "uncover the set of principles that unconsciously guide our moral judgments of permissible, obligatory, and forbidden actions." (Hauser 2006, p.48)

theory', conceived of solely as a descriptive enterprise, as an important *first step* in the method of reflective equilibrium, rather than the whole method.

How might this work? Well, if one is engaged in a deliberation about what one ought to think about some moral case, it seems reasonable to begin by attempting to elucidate what one already thinks, prior to considering relevant alternatives (accompanied by argument). This is implied by Rawls' distinction between *narrow* and *wide* reflective equilibrium.[28] Shortly after the comparison with methodology in linguistics, Rawls points out that the interpretation of reflective equilibrium "varies depending upon whether one is to be presented with only those descriptions which more or less match one's existing judgements except for minor discrepancies [narrow reflective equilibrium], or whether one is to be presented with all possible descriptions to which one might plausibly conform one's judgements together with all relevant philosophical arguments for them [wide reflective equilibrium]", and that "Clearly, it is the second kind of reflective equilibrium that one is concerned with in moral philosophy." (Rawls, 1999, p.43) In 'wide' reflective equilibrium, then, one is attempting to characterise the moral sense one would have if one considered many alternative moral conceptions, together with relevant supporting arguments, revising one's position accordingly.[29]

This emphasis on the *revisability* of one's existing moral conception (revealed through engaging in Rawlsian narrow reflective equilibrium) once we engage in Rawlsian wide reflective equilibrium takes us a long way away from regarding the method of wide reflective equilibrium as a kind of moral psychology. It can no longer be sensibly regarded as a matter of describing a particular person's moral sense because one is

[28] One should note here that Daniels' distinction between 'wide' and 'narrow' reflective equilibrium differs from that of Rawls'. Rawls' notion of 'wide reflective equilibrium' occurs merely once one considers a maximally inclusive set of alternative moral principles in the reflective equilibrium process; Daniels' notion of 'wide reflective equilibrium', by contrast, also explicitly includes relevant scientific and social scientific theories and relevant philosophical arguments, all of which may reasonably be supposed to significantly influence the reflective equilibrium process as a whole.

In what follows, I adopt the following convention: unless explicitly indicated otherwise, e.g. by using the modifier 'Rawlsian', 'wide reflective equilibrium' refers to Daniels' account of wide reflective equilibrium.

[29] Ideally, Rawls has it that one considers *all* possible alternatives to attain Rawlsian wide reflective equilibrium, though Rawls admits that it is "doubtful whether one can reach this state [of wide reflective equilibrium]." (Rawls 1999, p.43) As such, a state of wide reflective equilibrium represents a 'philosophical ideal' to which we approximate in our actual carrying out of a reflective equilibrium process.

characterising the hypothetical moral sense of a person who had likely made extensive revisions to their existing moral sense.

Furthermore, it is worth stressing that the only person in a position to *revise* one's moral beliefs is the person undertaking the evaluation. The point is well-made by Scanlon:

> The person whose considered judgments are in question has to be involved in this process – since only that person is in a position to revise his or her judgments – and that for this person it is a constant process of making up his or her mind about what to believe. When we are engaged in the enterprise that Rawls calls moral theory, we may have reason to consider the results of other people's search for reflective equilibrium as well as our own. But the process of seeking reflective equilibrium is something we each must carry out for ourselves, and it is a process of deciding what to think, not merely one of describing what we do think. (Scanlon 2003, p.149)

If one thinks of reflective equilibrium merely as a method to delineate what people happen to think, then one can only engage in the 'purely quantitative' matter of selecting moral principles that best approximate to the set of considered moral judgements that person is inclined to make. This is where the comparison with method in linguistics is appropriate. Rawls, however, stresses that the state of Rawlsian wide reflective equilibrium is "reached after a person has weighed various proposed conceptions and he has either revised his judgments... or held fast to his initial convictions." (Rawls 1999, p.43) Scanlon's point here is that the only way to make sense of such remarks is to regard the person whose moral sense is being 'described' as themselves involved in a deliberative process of seeking reflective equilibrium. Rawls' preferred understanding of reflective equilibrium is that of wide reflective equilibrium, and this is something that an evaluator must carry out for themselves, making revisions as they go. 'Moral theory', as Rawls construes it, is only a first step in this broader process.

The stress on revisability and the need for an evaluator to carry this out for themselves, argues Scanlon, means that we must regard reflective equilibrium as primarily a matter of deliberation about what one ought to think, rather than a matter of attempting to describe what someone (who might happen to be oneself) does in fact think, about moral matters. It seems to me Scanlon is quite correct about this.

Hare seems to have confused a characterisation of a *provisional aspect* of the method of reflective equilibrium with a characterisation of *the method itself*. If reflective equilibrium were merely a matter of describing what one happens to think, it would be a bizarre practice to make extensive and systematic revisions to one's moral conception.

1.3 – The Standard and Alternative Interpretations

These remarks bring us to the fourth, and most important, area of unclarity regarding Rawls' understanding of reflective equilibrium: the question of whether we ought to regard reflective equilibrium as primarily a methodological recommendation or as a coherence theory of justification in ethics where 'reflective equilibrium' is intended to pick out a distinctive epistemic state. As I noted above,[30] and as noted by Geoffrey Sayre-McCord,[31] these commitments are distinct. Nonetheless, commentators overwhelmingly favour the standard interpretation of reflective equilibrium, whereby reflective equilibrium is primarily of interest as a coherence theory of justification in ethics. That is, where holding a moral conception in the distinctive epistemic state of reflective equilibrium tells us something about whether one is justified in holding that moral conception.

In order to distinguish between the standard and the alternative interpretations of reflective equilibrium, we need to enquire as to the respective priorities of the *method* of reflective equilibrium as opposed to the *state* of reflective equilibrium. The easiest way to do this is to ask ourselves on what basis we justify adopting the method of reflective equilibrium. Shall we justify the method of reflective equilibrium via arguing that it helps us achieve the state of being in reflective equilibrium (standard interpretation), or for other reasons (alternative interpretation)? When asking as to this philosophical 'meta-justification'[32] for engaging in the method of reflective equilibrium, the standard view is that the method of reflective equilibrium is of value only insofar as it leads to the epistemic state of reflective equilibrium. This state is whereby one holds one's moral principles and moral judgements, including beliefs as

---

[30] See ft.3.

[31] Sayre-McCord makes distinguishes between epistemic, practical, and heuristic uses of the method of reflective equilibrium. On an *epistemic* understanding of the method of reflective equilibrium, the method of reflective equilibrium is to be recommended because it leads to a state of reflective equilibrium, which is understood as a coherence theory of justification (in moral matters, at least). (Sayre-McCord, p.143) As such, it corresponds to my 'standard' view. On a *heuristic* understanding, we might engage in the method of reflective equilibrium because it does lead to *justified* moral beliefs, but for reasons other than having achieved a coherent epistemic state, being in reflective equilibrium. (Sayre-McCord, pp.143-4) On a *practical* understanding, engaging in the method of reflective equilibrium is of practical value in that, say, it better prepares us for engaging in debate in a pluralistic society. (Sayre-McCord, p.144) Both the heuristic and the practical understandings would fall under the alternative view.

[32] The phrase is Klemens Kappel's (Kappel 2006). Since reflective equilibrium is intended to tell us something about the nature of justification, justifying the method itself amounts to a 'meta-justification'.

to why one rejects competing moral conceptions, in a coherent manner. The suggestion then is that one's being in this particular epistemic state tells us something about whether one is justified in one's particular moral conception. The philosophical motivation for engaging in the method of reflective equilibrium is that it is the best way to achieve this state. The method of reflective equilibrium is to be pursued because it reliably leads us to a state that is valuable independently, being in reflective equilibrium, a state which tells us something about whether one is justified in one's moral conception. If it were to turn out that some other distinguishable method were available that led to being in the state of reflective equilibrium more reliably, say, then we would no longer continue to recommend reflective equilibrium *qua* methodology.

A necessary feature of the standard interpretation, then, is that the state of reflective equilibrium is an independent consideration from considerations of how one reached that point. If we just take the state of (Rawlsian) narrow reflective equilibrium, for instance, we can characterise this state as one where an agent holds their moral principles and moral judgements in a coherent manner. That is, the agent is in a particular epistemic state, where their various moral beliefs stand in the particular logical and evidential relationships characteristic of coherence. The obtaining[33] of this state of (Rawlsian) narrow reflective equilibrium is a matter that is independent of the methodology used in achieving the state.

Then, if we incorporate, as (Rawlsian) wide reflective equilibrium requires of us, the consideration of alternative moral conceptions, together with accompanying arguments, we require a state of coherence between moral principles, moral judgements, and various beliefs at all levels of generality regarding alternatives and accompanying arguments (e.g. beliefs about whether particular arguments are sound). Similarly, the obtaining of the state of (Rawlsian) wide reflective equilibrium is an independent matter. It is possible, though incredibly unlikely, that an agent happens to hold their moral conception in a state of (Rawlsian) wide reflective equilibrium, but in fact reached such a point of coherence in their moral conception by believing everything they read on the back of a cereal packet. The standard view, overall, is this: the state of reflective equilibrium, a desirable state to be in because it tells us

[33] I will return to what I mean when I say that a state of reflective equilibrium 'obtains' in §2.5. For the moment, and for the purposes of brevity, I leave it unanalysed.

something about the justificatory status of that agent, is one where an agent happens to hold their moral beliefs (and, depending on how one characterises 'reflective equilibrium', non-moral beliefs) in a coherent manner. One may, in addition, recommend the method of reflective equilibrium as one that is best for achieving such a state, but this is an independent consideration.

On an alternative understanding of reflective equilibrium, however, one justifies engaging in the method of reflective equilibrium for reasons not connected with the coherence theory. There might be many ways in which one does this. One obvious way in which one might do this is by claiming that the method is of value because it represents a method by which one enriches and improves upon one's existing moral conception, without necessarily reaching some distinctive end-product that is coherent. Here, then, the philosophical motivation for engaging in the method of reflective equilibrium stems from regarding the method itself as of value independently of the particular results to which it leads. One simply has a methodological recommendation: what one ought to do is to first of all attempt to work out what one already thinks, attempt to systematise those judgements, and compare the resultant principles with alternatives (together with supporting arguments). This process may even have no distinctive end-points, but could well be an open-ended task.

Having said this, an alternative reflective equilibrium advocate might continue to utilise the notion of a 'state' of reflective equilibrium for reasons of grammatical convenience. For instance, suppose one had undertaken reflective equilibrium methodology for an extended period of time, and found that one's moral conception was exceptionally stable when faced with alternative moral conceptions. Here it seems appropriate to regard oneself as in a 'state' of (Rawlsian) wide reflective equilibrium, where this should be understood as the claim that 'Engaging in the method of reflective equilibrium has led me to a point where my moral conception is stable under challenge. Further consideration of relevant alternatives is unlikely to destabilise my views.' Here, the 'state' of being in reflective equilibrium, if one is happy to call it a 'state',[34] is merely a placeholder for a claim about what further

[34] This is a terminological issue. I would regard it as misleading, given the prevalence of the standard understanding of reflective equilibrium, having adopted an alternative view of reflective equilibrium, to make claims about being in a 'state' of reflective equilibrium. To my mind (and as we shall see),

engagement in the method would bring. It does not describe some epistemic state that obtains independently.

As the presence of alternative understandings of reflective equilibrium make clear, one can recommend the method of reflective equilibrium without thinking of reflective equilibrium as a coherence theory of justification. The standard view is to regard the coherence theory as of primary importance, and to subordinate the recommendation of the method of reflective equilibrium to achieving the goal of attaining a state of reflective equilibrium in one's moral conception. The alternative view, however, is to recommend the method reflective equilibrium without claiming that it leads to such an independently-obtaining state of coherence. One may wish to utilise the notion of a 'state' of reflective equilibrium, but here the notion of a 'state' functions as a placeholder for claims about the outcome of further engagement in the method of reflective equilibrium.

Having clarified the distinction between these two approaches, we can now enquire as to what view Rawls himself took. One can see that Rawls may well have had the standard understanding in mind, whereby reflective equilibrium is primarily denotes an independently-characterised epistemic state, and where one ought to engage in the method of reflective equilibrium only insofar as it reliably leads to this state. Take, for instance, his insistence that "A conception of justice cannot be deduced from self-evident premises or conditions on principles; instead, its justification is a matter of the mutual support of many considerations, of everything fitting together into one coherent view." (Rawls 1999, p.19) Rawls here only mentions the resultant epistemic *state* of 'everything fitting together into one coherent view' as telling us something about justification.[35] As we've already seen, he also refers to reflective equilibrium as a 'state of affairs' when introducing the notion (Rawls 1999, p.18), which again might be taken to suggest that 'reflective equilibrium' functions primarily as a description of an epistemic state, as opposed to delineating a particular methodology.

---

claiming one is in a *state* of reflective equilibrium typically amounts to a claim about the current logical status of one's belief set, i.e. that one's moral conception is coherent. As such, it is akin to claims such as "My argument is valid", or "My position is consistent".

[35] This is reaffirmed at the very end of *A Theory of Justice*, where Rawls reminds the reader that his view is that the theory's "Justification rests upon the entire conception and how it fits in with and organizes our considered judgments in reflective equilibrium." (Rawls 1999, p.507)

However, this is scant evidence for attributing to Rawls the thought that reflective equilibrium characterises an epistemic state that obtains independently of the methodology involved. His favoured locution with regards to reflective equilibrium is to talk of judgements one holds *in* reflective equilibrium (Rawls 1999, p.40; p.43; p.44; p.96; p.104; p.105; p.159; p.379; p.381; p.507), which is ambiguous between the two interpretations I've outlined. One can either regard holding judgements in reflective equilibrium as a matter of holding certain moral beliefs whilst undergoing the reflective equilibrium process, or as being in an independently-characterised epistemic state that one may or may not reach via undergoing the reflective equilibrium process.

Positive evidence for the alternative view that reflective equilibrium is simply a methodological recommendation without any commitment to a coherence theory of justification is perhaps even more sparse. It is a consideration in favour of regarding Rawls as holding the view that reflective equilibrium is important primarily as a methodology that there is not a single instance in *A Theory of Justice* where Rawls discusses achieving a state of reflective equilibrium independently of discussing the method of moving between one's principles and judgements, etc. Furthermore, his description of being in a state of reflective equilibrium as a 'resting point' (Rawls 1999, p.44) does suggest that being in that 'state' is more a matter of a certain point of stability having adopting the method of reflective equilibrium: something is only a *resting* point in relation to a process that has, for the time being at least, ceased. Nonetheless, one must recognise that this is insufficient evidence to clarify Rawls' position on the priority of the method or state of reflective equilibrium. I thus conclude that, having clarified the distinction, either interpretation represents a legitimate reading of Rawls' intent. He may well have taken the standard view that reflective equilibrium is of primarily importance *qua* coherence theory of justification in ethics; or he may simply have taken a view on methodology independently of this claim.

1.4 – Brandt's Garbage-in/Garbage-out Objection

We have, then, a characterisation of the standard view of reflective equilibrium with its commitment to reflective equilibrium *qua* coherence theory as of primary interest. Commentators already note that the coherence theory occupies a central position in the literature on reflective equilibrium.[36] If anything, however, this underplays the significance of the coherence theory, and consequently the overwhelming predominance of the standard view of reflective equilibrium. References to reflective equilibrium *qua* coherence theory of justification in ethics are to be found in (Brink 1989, pp.103-4; Brink 1987, pp.73-4; Brandt 1990, pp.272-3; Brandt 1998, pp.18-9; Daniels 1979, p.22; Daniels 1980a pp.60-1; Daniels 1996, p.1; Daniels 2008; DePaul 1987, p.463; DePaul 1993, p.13; Ebertz 1993, pp.193-8; Haslett 1987, p.306; Lyons 1989, pp.144-5; Millgram 2005, pp.8-9; Petersson 1998, p.127; Sayre-McCord 1996, pp.141-5; Singer 2005, pp.344-5; Sinnott-Armstrong 1996, pp.31-2; Tersman 2006, p.26; Tersman 2008, p.398; Timmons 1999, pp.236-9). Where there is discussion of the *method* of reflective equilibrium, there is either no mention of its relationship to the coherence theory of justification (which leaves it ambiguous as to whether they are adopting the standard or the alternative view), or the method is explicitly linked to the coherence theory of justification. The overwhelming cursory impression is that the standard view is taken for granted.

However, we can bolster this cursory impression by examining the shape of a central line of discussion concerning reflective equilibrium. This will also enable us to garner a clearer view of how the standard view of reflective equilibrium needs to be understood. This central line of discussion concerning reflective equilibrium is known as the 'garbage-in/garbage-out objection'.[37] A clear presentation of the objection(s)[38] is to be found in Brandt's *A Theory of the Good and the Right*. (Brandt 1998) Brandt targets the method of reflective equilibrium inasmuch as it involves,

[36] David Copp, for instance, avers that "The so-called 'method' of WRE [wide reflective equilibrium] is, I think, usually viewed as an account of epistemic justification rather than as a prescription about the conduct of moral inquiry". (Copp 1996, p.960) Karen Jones, in her survey of moral epistemology, tells us that "Coherentism is typically thought to gain support from the philosophical and common-sense use of reflective equilibrium." (Jones 2007, p.74)

[37] The term is Jones' (Jones 2007, p.66).

[38] One may regard the garbage-in/garbage-out objection as two distinct objections, what we might call a 'no-credibility objection', and a 'low-credibility objection'. Alternatively, one may regard it as essentially the same objection, albeit in different guises. I have adopted the latter way of presenting the matter.

through the use of considered moral judgements, reliance upon moral 'intuitions', which he (partially) defines as "Beliefs in, and dispositions on occasion to utter, certain normative statements." (Brandt 1998, p.17) Like beliefs more generally, intuitions can be held at varying levels of generality, but the distinguishing feature of an intuition is that one gives it credence independently of its logical connections to other beliefs. (Brandt 1998, p.18) That is, a moral belief counts as an intuition if one gives it *some* credence on its own terms, independently of considering how it relates to other beliefs one happens to hold.

Brandt then characterises the method of reflective equilibrium as involving testing normative principles against the set of moral intuitions one happens to hold, until one achieves a coherent system of moral beliefs. (Brandt 1998, p.20) To put it in our terms, the method is intended to facilitate the attainment of the epistemic state of being in reflective equilibrium, which is clearly to take the standard view of reflective equilibrium. At this point, however, Brandt asks why we ought to be interested in a *coherent* set of moral beliefs (i.e. in a state of reflective equilibrium) in the first place:

> There is a problem here quite similar to that which faces the traditional coherence theory of justification of belief: that the theory claims that a more coherent system of beliefs is better justified than a less coherent one, but there is no reason to think this claim is true unless some of the beliefs are initially credible – and not merely initially believed – for some reason other than their coherence, say, because they state facts of observation. In the case of normative beliefs, no reason has been offered why we should think that initial credence levels, for a person, correspond to credibilities. The fact that a person has a firm normative conviction gives that belief a status no better than fiction. Is one coherent set of fictions supposed to be better than another? (Brandt 1998, p.20)

One can grant, then, that one would like, for all sorts of reasons, to have a coherent moral conception. It may even be a necessary condition for that moral conception to be justifiably held. However, argues Brandt, it cannot be sufficient: we still need reasons to think that the moral beliefs within this coherent epistemic state are credible to begin with. If they are not credible, it is mysterious why the mere systematisation of moral beliefs that are merely sincerely held (i.e. given credence) will make any difference to the credibility of any of those initial beliefs. Arriving at an epistemic state of being in reflective equilibrium puts one's moral intuitions into a coherent form, but this consideration seems moot in the absence of some reason for

thinking those intuitions credible in the first place. If you put garbage in, you get garbage out.

Brandt makes much of a comparison between the use of moral intuitions and the use of observation in science, a comparison that is encouraged by Rawls' remarks on reflective equilibrium.[39] In the case of observation is science, presumably we do have reasons for taking observations as initially credible, i.e. credible independently of the fact that they cohere with our best scientific theories.[40] However, we have no corresponding reason(s) for assigning moral intuitions any initial credibility. As such, Brandt's objection is that reflective equilibrium utilises considered moral judgements, a form of moral intuitions, and moral intuitions lack *any* initial credibility at all. In thinking of intuitions as normative beliefs which derive some credence independently of their logical relationships to other beliefs, Brandt issues an implicit demand for the 'coherentist-intuitionist' philosophical account of justification: we need to give some reasons, independently of whether those intuitions cohere with our moral theory, why we ought to regard moral intuitions as initially credible. These sorts of considerations are arguably available when it comes to justifying the use of observation in science; a similar case needs to be made for the use of moral intuitions in ethics. Unless such considerations are provided, we can only presume that moral intuitions lack any initial credibility.

However, various further considerations suggest that Brandt's objection may well come in a different guise, where the argument is that, regardless of whether moral intuitions have any initial credibility, they have, even after considering their relationships to other beliefs, *very low* credibility:

---

[39] We've already seen Rawls' comparison of the method of reflective equilibrium with methodology in linguistics, with the thought that moral principles are constrained by the "Definite if limited class of facts against which conjectured principles can be checked, namely, our considered judgements in reflective equilibrium." (Rawls 1999, p.44) The thought seems to be that, much like the range of acceptable (justified) scientific theories is constrained by the available evidence provided via scientific observation, the range of acceptable (justified) moral theories is constrained by the class of considered moral judgements available upon reflection.

[40] This, of course, is a matter of considerable contention. Presumably the thought is that scientific observations demonstrate their *reliability* and thus initial credibility independently of their role in confirming scientific theories and hypotheses. Daniels demurs, however, largely on grounds of Quinean holism. I will briefly canvass Daniels' reply in the next section; a full examination of the issue is well beyond my scope, however.

> Various facts about the genesis of our moral beliefs militate against mere appeal to intuitions in ethics. Our normative beliefs are strongly affected by the particular cultural tradition which nurtured us, and would be different if we had been in a learning situation with different parents, teachers, or peers. Moreover, the moral convictions of some people derive, to use words of Peter Singer, 'from discarded religious systems, from warped views of sex and bodily functions, or from customs necessary for the survival of the group in social and economic circumstances that now lie in our distant past'. [(Singer 1974, p.516)] What we should aim to do is to step outside our own tradition somehow, see it from the outside, and evaluate it, separating what is only the vestige of a possibly once useful moral tradition from what is justifiable at present. (Brandt 1998, pp.21-2)

Brandt notes the inter- and intra-cultural variation of moral beliefs,[41] the tendency for our moral beliefs to be determined to some extent by the essentially arbitrary fact as to where one grew up, and the tendency for our moral beliefs to survive beyond the point where they may have at least initially had empirical support. In doing so, Brandt gives us reasons precisely not to think that justifying some moral theory would partially involve accounting for the intuitions (in the form of considered moral judgements) one is initially inclined to hold. Even if moral intuitions hold any initial credibility, we have plenty of reasons to treat them as having, in the final analysis, *very low* credibility.

These suspicions appear to have received some confirmation by recent work by Joshua Greene and Jonathan Haidt, (Green & Haidt 2002) which suggests that moral intuitions are, as Brandt and Singer claim, rather susceptible to various biases (in this case, biases that look to be explainable in evolutionary psychological terms). For instance, people tend to avoid making decisions which would reliably cause some adverse emotional reaction (e.g. pushing an obese man off a bridge to stop a runaway trolley, i.e. tram, that's about to kill 5 bystanders), but seem to have less qualms about decisions where the emotional reaction is less likely to occur (e.g. throwing a switch that directs the trolley down a track that kills one obese gentleman rather than 5 bystanders).[42] As such, differences in one's moral intuitions seem to track arguably morally non-relevant aspects of a situation, which casts serious doubt on the credibility of such intuitions. As Singer puts it:

---

[41] "Moral disagreement does not exist only between our own reflective equilibria and those of some primitive tribes, or on relatively superficial matters. It exists among sophisticated civilized persons and in core areas." (Brandt 1998, p.22)

[42] The situations derive from Judith Jarvis Thomson's 'Killing, Letting Die, and the Trolley Problem' (Thomson 1976).

> Very probably, there is no morally relevant distinction between the cases. At the more general level of method in ethics, this same understanding of how we make moral judgments casts serious doubt on the method of reflective equilibrium. There is little point in constructing a moral theory designed to match considered moral judgments that themselves stem from our evolved responses to the situations in which we and our ancestors lived during the period of our evolution as social mammals, primates, and finally, human beings. We should, with our current powers of reasoning and our rapidly changing circumstances, be able to do better than that. (Singer 2005, p.348)

The suspicion is that ethical intuitions in the form of considered moral judgements have low credibility because they often track morally non-relevant aspects of the situation. Often, our moral intuitions are simply the arbitrary products of our evolutionary and cultural history. Achieving a state of reflective equilibrium, however, requires that we take ethical intuitions, in the form of considered moral judgements, as playing a key role in constraining the range of justifiable moral conceptions. This raises the possibility that ethical intuitions that track morally non-relevant aspects of a situation would be accounted for in a state of reflective equilibrium.

At this point, the garbage-in/garbage-out objection takes the form of illustrating how intuitions with very low credibility can nonetheless survive into a state of reflective equilibrium. We can design relatively plausible[43] counter-examples whereby one starts off with a set of transparently unjustified ethical intuitions that can nonetheless survive into a state of reflective equilibrium because they can be rendered internally coherent. The above trolley problems seem to require of us to maintain, as a general moral principle, that there is some morally relevant difference between the case of pushing an obese man off a bridge, as opposed to flicking a switch, where the consequences of the actions are identical. One can easily imagine that "It is morally better to perform some action though a technological intermediary as opposed to having to do it oneself" might well be upheld in a state of reflective equilibrium. However, such a 'moral principle' looks to be transparently unjustifiable, and this casts doubt on the idea that reflective equilibrium has got anything to tell us about the nature of justification in ethics.

[43] 'Plausible' because, as we shall see in §2.2, one need not claim that a reflective equilibrium account is immune to *any* counter-examples as an *a priori* matter.

Since we'll be relying on the work of Stephen Stich in the next chapter, it is worth mentioning here that Steven Stich uses exactly this sort of argument via invoking plausible counter-examples, albeit in a different context. In arguing against 'neo-Goodmanian'[44] attempts to justify inferential rules via maintaining they are upheld in reflective equilibrium, (Stich 1990) Stich uses the gambler's fallacy[45] as an instance of a patently unjustified inferential principle (i.e. with low credibility) that could nonetheless plausibly survive into reflective equilibrium. People utilising this inferential principle will insist that the non-occurrence of a 6 on many throws of a die increases the likelihood that it will subsequently occur. One will even find that such errant inferences have been enshrined in an everyday principle known as 'the law of averages'. This 'law' maintains that there is active pressure on independent events to conform to our long-run expectations, so three 6's in a row makes it less likely that another will result on the next throw. If anything, however, three 6's in a row makes it *more* likely than a one in six chance that a 6 will turn up, since it makes it more likely that the die is biased. If such a principle could be held in a state of reflective equilibrium, then this strongly suggests that something is amiss with regarding being in a state of reflective equilibrium as telling us something about justification. The underlying problem here is that 'daffy' (i.e. garbage) judgements, whether moral or inferential, could plausibly survive into a state of reflective equilibrium, and one would have 'daffy' principles as a result. One could not conceivably claim here that being in a state of reflective equilibrium told us a great deal about whether one was justified in holding such principles.

In such cases, the proponent of reflective equilibrium is committed to claiming that the resultant theory is in reflective equilibrium, and that this tells us something about whether it may be justifiably held. At this point, the critic holds, the reflective equilibrium defender has egregiously parted company with common-sense. Whatever 'justification' amounts to, it had better not admit of such plausible counter-examples. It is far from clear, however, how reflective equilibrium theorists could weed out such cases, committed as they seem to be to taking considered moral (or inferential) judgements as initially credible. It seems hard to credit reflective equilibrium with telling us anything about what it is for a person to be justified in holding a certain

[44] See ft.63 below.

[45] "These people infer that the likelihood of throwing a seven in a game of craps increases each time a nonseven is thrown. What is more, there is every reason to think that the principle underlying their inference is in reflective equilibrium for them." (Stich 1990, p.83)

moral conception if such intuitions can survive into a state of reflective equilibrium. As such, the claim is that moral intuitions are 'garbage' in that they have, as a general group, very low credibility. Even if such moral intuitions are held in an epistemic state of coherence, they remain garbage, and are thus unlikely to tell us much about the justificatory status of one's moral conception.

1.5 – Daniels' Wide Reflective Equilibrium

In this section, I shall examine some of Daniels' remarks on reflective equilibrium, and demonstrate how he addresses the garbage-in/garbage-out objection in its two guises as either the claim that moral intuitions lack any credibility, or the claim that intuitions with low credibility can be upheld in reflective equilibrium. From this point on, it will be Daniels' account of wide reflective equilibrium that will serve as our focus in terms of reflective equilibrium, since his exposition is notably clearer and more explicit than Rawls', and has itself received more concentrated attention than Rawls'. In following the contours of the discussion, we shall see, not only that Daniels' wide reflective equilibrium approach follows the standard view of reflective equilibrium, but also get a clearer picture of the particular commitments that the standard view needs to maintain.

At a cursory level, it is appropriate to regard Daniels as taking the standard view of reflective equilibrium in which it functions primarily as a coherence theory of justification in ethics. Daniels is content to regard his account of reflective equilibrium as a coherence theory of epistemic justification for moral considerations.[46] Daniels is also by-and-large content with Brandt's understanding of reflective equilibrium as a coherence theory,[47] but with an important qualification: Daniels takes it that Brandt's view of reflective equilibrium is too simplistic.

[46] Daniels proposes a contrast between reflective equilibrium construed in a 'modest' and a 'daring' way. In its 'modest' role, the method of reflective equilibrium is primarily descriptive, allowing us to discern the structure of a person's moral conception. This corresponds to Rawls' understanding of 'moral theory', which we surveyed above. In its 'daring' role, however, wide reflective equilibrium "serves as the basis for a coherence account of moral justification." (Daniels 1980a, p.60) As subsequent remarks in the article make clear (Daniels 1980a, pp.61-2), Daniels clearly prefers the 'daring' view. This is confirmed in other relevant articles: "Wide reflective equilibrium is... a theoretical account of justification in ethics." (Daniels 1996, p.2); "If we construe wide reflective equilibrium as providing us with the basis for a full-blown coherence theory of moral justification, then my argument suggests that it faces the same difficulties and advantages as coherence theories of non-moral justification. I cannot here defend my view that a coherence theory of justification can be made compatible with a non-coherence account of truth." (Daniels 1979, p.45n.29); "Justification in ethics rests, I have long thought, on a broad coherentist approach involving beliefs at many levels." (Daniels 1996a, p.338)

[47] Daniels is careful to address what he describes as Brandt's criticisms of reflective equilibrium as Brandt understands it, as opposed to simply addressing Brandt's criticisms of reflective equilibrium: "Brandt characterises the method of reflective equilibrium as follows." (Daniels 1996, p.29); "Brandt elaborates his own view in response to the "intuitionism" he thinks undermines the method of wide reflective equilibrium" (Daniels 1996, p.81), This, to my mind, suggests that Daniels isn't entirely happy with Brandt's characterisation of reflective equilibrium, but not sufficiently displeased to regard Brandt as having *mis*characterised reflective equilibrium. The natural way to understand this is that Daniels regards Brandt's understanding as too simplistic, but is nonetheless broadly correct in regarding reflective equilibrium as a coherence theory of justification in ethics.

Contrasting the 'traditional' understanding of reflective equilibrium, narrow reflective equilibrium, with his version of wide reflective equilibrium,[48] he regards Brandt as addressing only the former. He regards narrow reflective equilibrium as a "simple coherence view of justification" whereby all one need do is make one's moral judgements and moral principles cohere, whereas he regards his own account of 'wide reflective equilibrium' as allowing for "a greater complexity in the structure of moral theories than the traditional view" (Daniels 1996, p.21), since it introduces the possibility that one's reflective equilibrium may well be disturbed by background theoretical considerations.

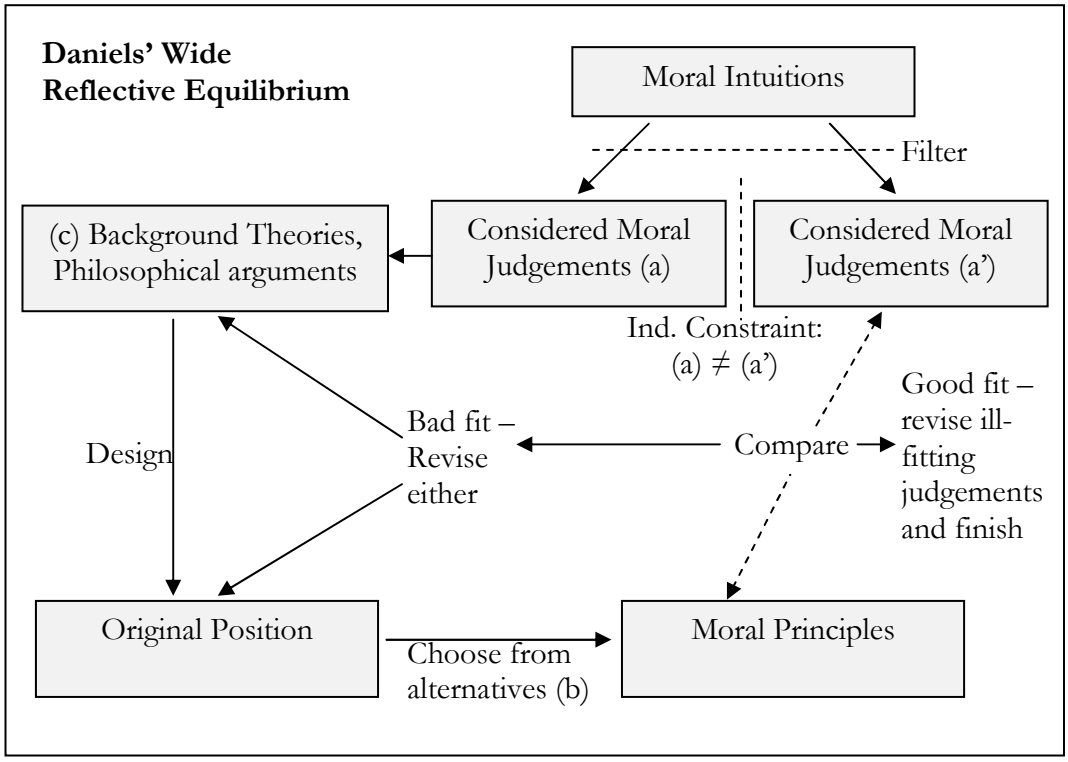
Daniels defines a state of wide reflective equilibrium as follows:

> A wide reflective equilibrium is a coherent triple of sets of beliefs held by a particular person; namely (a) a set of considered moral judgements; (b) a set of moral principles; and (c) a set of relevant background theories, which may include both moral and nonmoral theories. We collect the person's initial moral judgements, which may be particular or general, and filter them to include only those of which he is relatively confident and which have been made under conditions generally conducive to avoiding errors of judgement. We propose alternative sets of moral principles which have varying degrees of "fit" with the moral judgements, Rather than settling immediately for the "best fit" of principles with judgements, which would give us only a narrow equilibrium, we advance philosophical arguments that reveal the strengths and weaknesses of the competing sets of principles (that is, competing moral conceptions). I construe these arguments as inferences from relevant background theories (I use the term loosely). Assume that some particular set of arguments wins and the moral agent is thus persuaded that one set of principles is more acceptable than the others (and perhaps than the conception that might have emerged in narrow equilibrium). The agent may work back and forth, revising his initial considered judgements, moral principles, and background theories, to arrive at an equilibrium point that consists of the triple – (a), (b), and (c). (Daniels 1996, p.82)

A "fit" between considered moral judgements and moral principles for a particular moral agent results merely in a state of *narrow* reflective equilibrium; if one then incorporates all relevant philosophical arguments and background theories in the process, the agent arrives at a state of *wide* reflective equilibrium. Here, the sort of considerations that might be regarded as relevant background theories are theories

[48] Again, as noted in ft.28, the 'wide' and the 'narrow' of Daniels' reflective equilibrium do not correspond to Rawls' use of 'wide' and 'narrow' reflective equilibrium. The principle difference concerns the explicit inclusion of relevant *background theories* in Daniels' wide reflective equilibrium: as such, Rawlsian wide reflective equilibrium would still be a case of Daniels' narrow reflective equilibrium.

from psychology, sociology, biology, cognitive science, anthropology etc.[49] For instance, the fact that Rawls makes certain assumptions about how human beings would behave in a well-ordered society is arguably informed by reasonable sociological and psychological assumptions: people, in a well-ordered society regulated by public norms, tend to have a sense of justice and fairness, but also have a certain conception of what constitutes good and bad actions. Rawls' practice here would be an example of wide reflective equilibrium in action since he relies on contemporary theories from the human and social sciences. I represent Daniels' version of wide reflective equilibrium below:



It is important for Daniels' purposes that the background theories included in a state of wide reflective equilibrium are not merely introduced to bolster the set of considered moral judgements one is already inclined to make, or one is simply making 'accidental generalisations' purely for the purposes of one's own theory

[49] "It is important that we see how diverse the types of beliefs included in wide reflective equilibrium are, as well as the kinds of arguments that may be based on them. They include our beliefs about particular cases; about rules, principles, and virtues and how to apply or act on them; about the right-making properties of actions, policies, and institutions; about the conflict between consequentialist and deontological views; about partiality and impartiality and the moral point of view; about motivation, moral development, strains of moral commitment, and the limits of ethics; about the nature of persons; about the role or function of ethics in our lives; about the implications of game theory, decision theory, and accounts of rationality for morality; about human psychology, sociology, and political and economic behaviour; about the ways we should reply to moral scepticism and moral disagreement; and about moral justification itself. As is evident from this broad and encompassing list, the elements of moral theory are diverse. Moral theory is not simply a set of principles." (Daniels 1996, p.6)

rather than appealing to theories that have considerable scope beyond one's own theory. Daniels therefore requires that there the considered moral judgements (a'), that are used in comparing the moral principles selected from the list of alternatives (b), should not be the same set of considered moral judgements (a) used to constrain the set of relevant background theories (c). Provided a significant portion of the intuitions that govern the choice of background theories are independent of the intuitions that govern the adoption of certain moral principles, then Daniels' 'independence constraint' is satisfied (that (a) should be sufficiently independent of (a')), and we can say that the resulting "fit" represents a state of wide reflective equilibrium.

Before we proceed to consider how Daniels responds to Brandt's garbage-in/garbage-out objection, it is worth bringing out how Daniels' wide reflective equilibrium approach is a paradigm example of the standard interpretation of reflective equilibrium in that it characterises wide reflective equilibrium as a state that obtains (or not) in virtue of the particular epistemic state of agent at the time of asking. We note that a state of wide reflective equilibrium obtains where an agent holds a coherent triple of sets of beliefs involving their moral principles, moral judgements, and background theories (including relevant philosophical considerations). If the agent happens to hold a set of principles, judgements, and relevant background theories in a manner characteristic of coherence, they are in a state of wide reflective equilibrium. As such, it is natural to think of the agent as a repository for various moral and non-moral beliefs, and whether or not a state of wide reflective equilibrium obtains as instantiated by some function operating over those beliefs. The function in question here is whether the various beliefs 'fit' together. If they are in the various logical and evidential relationships characteristic of such 'fitting', then a state of wide reflective equilibrium obtains. In all of this, it is strictly speaking irrelevant *how* the agent arrived at that point.

Evidently, we'd need to say much more about what such 'fitting' involved: clearly it needs to be a more substantial function than mere absence of internal contradictions.[50] There would need to be, for instance, positive epistemic support from relevant background theories, and the various considered moral judgements one initially was inclined to hold would need to be transparently derivable from the

[50] See Tersman's *Reflective Equilibrium: an Essay in Moral Epistemology* (Tersman 1993) for a lengthy discussion of exactly how one might characterise the various logical and evidential relationships characteristic of being in reflective equilibrium. Bo Petersson also discusses this problem in some depth in his 'Wide Reflective Equilibrium and the Justification of Moral Theory'. (Peterson 1998, pp.130ff.)

more generalised moral principles the agent holds. However such 'fitting' is characterised, it is nonetheless some function involving the logical and evidential relationships between the various beliefs the agent happens to hold. Given that an agent has some set of moral and non-moral beliefs, and given a full, precise description of what such 'fitting' involved, their status in terms of whether they hold those beliefs in wide reflective equilibrium is already set.

How, then, might one utilise a conception of *wide* reflective equilibrium to counter the garbage-in/garbage-out objection in its two guises? On the first point, the criticism that some account of why moral intuitions have *any* credibility is needed yet lacking, Daniels first of all targets the idea that the initial credibility of observation reports in science is established independently of its logical relationships to other scientific beliefs:

> Observation reports are neither self-warranting nor unrevisable, and our willingness to grant them initial credibility depends on our acceptance of various other relevant theories and beliefs. Similarly, in rejecting the view that wide equilibrium merely systematizes a determinate set of moral judgements, and arguing instead for the revisability of these inputs, I suggest that wide equilibrium closely resembles scientific practice. Neither in science nor in ethics do we merely "test" our theories against a predetermined, relatively fixed body of data. Rather, we continually reassess and reevaluate both the plausibility and the relevance of these data against theories we are inclined to accept. (Daniels 1996, p.33)

The claim, then, is that, if one really wishes to demand that the credibility of the relevant inputs to a process of theorising, whether scientific or moral, needs to be established *independently* of that process of theorising, then one seems committed to the claim that scientific observations lack credibility. To put it in more recognisable terms, Daniels emphasises the *theory-ladenness* of scientific observation.[51] The implication is that Brandt's implicit demand that the credibility of moral intuitions be established independently of coherence consideration is too onerous. One may here acknowledge that there is a large *difference of degree* with regards to the theory-ladenness of scientific observation and moral intuitions, but that is not to say that there is a *difference of kind* here.[52] Nonetheless, he maintains, the demand that one establish the

[51] "The credibility assignment [for scientific observations]... draws implicitly on a broadly accepted body of theory which explains why those judgments are credible." (Daniels 1979, pp.31-2)

[52] Daniels elsewhere acknowledges that it is possible to have, using Ronald Dworkin's phrase, a 'false sophistication' about science, whereby one takes the analogy between the use of moral intuitions and

credibility of either scientific observation or moral intuitions independently of how they cohere with wider theoretical concerns, is unduly burdensome.

Daniels does, however, acknowledge that *some* account of why moral intuitions have credibility is necessary within his coherentist framework. The difference here is that this account for Daniels would be a relevant background theoretical consideration that is itself held in a state of wide reflective equilibrium, as opposed to justified independently.[53] Daniels sees no reason why such an account could not be generated. As far as assigning *any* credibility to moral intuitions, then, he takes the issue between him and Brandt to be a 'burden of proof' dispute, whereby Brandt is demanding that we not incorporate moral intuitions into an account of justification without first establishing their credibility, and Daniels arguing that we may do so provisionally in the hope that such an account will be forthcoming "once we know more about moral theory."[54] (Daniels 1979, p.33) As such, the garbage-in/garbage out objection in its no credibility guise, without considerable work concerning issues to do with theory-ladenness in science, is hardly a decisive consideration in Brandt's favour.

In its second guise, as the objection that (ethical) intuitions with very low credibility can nonetheless survive into a state of reflective equilibrium, Daniels has it that his characterisation of *wide* reflective equilibrium provides the necessary criteria to ensure that such cases are unlikely to occur. Daniels is keen to emphasise that wide reflective

---

scientific observation too literally. (Dworkin 1973, p.32) Thus, he acknowledges that "the extensive revisability of considered moral judgments does count against construing them as hard or privileged (I only mean completely reliable) observation reports." (Daniels 1980b, p.76) Daniels is keen to emphasise also that there are important dissimilarities regarding other aspects of the analogy: considered moral judgements are far more like 'theoretical considerations' than 'observation reports', they generally require inferential support, and do not obviously reflect 'simple properties of moral situations'. (Daniels 1979, pp.30-1). In general, then, we ought not expect considered moral judgements in reflective equilibrium to function in a manner akin to that of observation in scientific reasoning. The only relevant point of analogy is that both moral intuitions and scientific observation are assigned credibility values *internally,* i.e. via their coherence with broader theoretical concerns.

[53] This can give the appearance of vicious circularity in that one's 'account' of the credibility of moral intuitions is that they allow one to derive the principles one wanted. However, this is not the case since it would be subject to the proviso that any account of the credibility of one's intuitions satisfies Daniels' independence constraint. That is, where the reasons one regards moral intuitions as credible are, to some extent, independent of their role in allowing one to derive the moral theory one wanted.

[54] DePaul has it that this is a 'Wimpy' response (DePaul 1993, pp.30-1). Wimpy is a character in the Popeye comics who promises "to gladly pay you Tuesday for a hamburger today." DePaul, however, avers that a reflective equilibrium approach will "never produce an argument for its own reliability." (DePaul 1993, p.56)

I am myself unsure what to make of Daniels' 'burden of proof' point here. My inclination is to regard it as acceptable, but I'm not sure I can give any good reasons for this. However, since my ultimate aim is to argue that, even if we grant the reliability of moral intuitions, the sort of project Daniels has in mind will not be successful, the point is somewhat moot.

equilibrium allows for far more *extensive* revision of one's initial considered moral judgements than narrow reflective equilibrium: "Wide reflective equilibrium keeps us from taking considered moral judgements at face value, however much they may be treated as starting points in our theory construction." (Daniels 1979, p.28) If narrow reflective equilibrium merely systematises one's initial prejudices, wide reflective equilibrium provides the tools to remove those prejudices from one's reflective equilibrium altogether.

To see this, consider an example or two. Suppose one had, as part of one's moral conception, the belief that men ought, in general, to be paid more than women, deriving from a more general principle that people ought to be paid more the harder they work, together with the background belief that women work less hard. If pressed, one might attempt to justify this background belief with the thought that women talk more than men, and are thus likely, in general, to not work as hard. One can see that this sort of moral conception might well be upheld in a state of narrow reflective equilibrium, provided other relevant moral judgements and principles could be made to cohere with this viewpoint. However, it is plausible to think that such a moral conception could not be upheld in wide reflective equilibrium. This may simply be because the viewpoint is, at least as currently set out, woefully inadequately supported by evidence, or it may be because relevant background theories cannot be made to cohere with this moral conception. If we consider statistical evidence that suggests that there is virtually no difference in the average number of words used per day by men than by women,[55] for instance, the moral conception would surely stand in need of revision.

Consider again Stich's putative counter-example in the case of inferential rules. Here, a parallel wide reflective equilibrium strategy would be to deny that the 'law of averages' could pass into *wide* reflective equilibrium. What we need to think about here is just how we would show the gambler's fallacy *to be* a fallacy, and the 'law of averages' to be an unjustified inferential principle. Why is it not a piece of good inferential reasoning? Well, presumably because any predictions we might make on

[55] See 'Are Women Really More Talkative Than Men?' (Mehl et al. 2007). The investigators find that women use 16,215 words on average per day, whereas men use 15,669 words, which is not statistically significant enough to point to any salient difference between the sexes with regards to word use. Intriguingly, they found that whilst the total number of words used was roughly the same, they found that women tend to utilise a larger vocabulary than men.

the basis of using the 'law of averages' inferential principle will turn out to be unreliable. This may be fairly opaque in many spheres of everyday life, but becomes transparent in certain contexts. Following any degree of statistical analysis, for instance, the 'law of averages' would quickly be exposed as unfounded and false. We might demonstrate its falsity, for instance, in a mathematics classroom over an hour using coins. We could compile data on the next outcome given 3 heads in a row, the next outcome given 3 tails in a row, a control group etc., and then we will doubtless discover that previous outcomes have no effect whatsoever on the next outcome. Once we widen the range of considerations incorporated within a state of wide reflective equilibrium, we can see that the gambler's fallacy could not plausibly survive into wide reflective equilibrium.

We can generalise the point: any considerations that can possibly be adduced in order to demonstrate that a particular piece of reasoning is unjustified will be the *selfsame considerations* that ensure that it could not survive into a state of wide reflective equilibrium. Any argument of the form "Here's something that could be upheld in a state of reflective equilibrium, but look, it's not justified because of *x* and *y*" can be answered by "But *x* and *y* would be incorporated within *wide* reflective equilibrium." If one attends to the argument intended to demonstrate that a 'daffy' inferential or moral principle would pass into reflective equilibrium, one will see that the argument, if a good one, is *itself* the guarantee that the inferential or moral principle could not survive into a state of wide reflective equilibrium.

This, then, represents a general strategy that Daniels, as a *wide* reflective equilibrium theorist, could utilise when faced with any plausible counter-examples. Attaining a state of wide reflective equilibrium requires an evaluator to revise their initial intuitions to a far greater degree than the simple coherence considerations of narrow reflective equilibrium. The garbage-in/garbage-out objection, construed as a point concerning the very low credibility of moral intuition, then, is answered by turning the objection on its head. The question to the objector is: how do *you* know when the relevant moral intuitions are 'garbage'? Any considerations that can be adduced at this point to demonstrate the justificatory problems for some specific set of inputs will precisely be the sorts of considerations one should have introduced within wide reflective equilibrium. One can only design apparent cases where something appears

to be in wide reflective equilibrium whilst not being justifiably held by thinking of cases where one withholds considerations that would themselves need to be incorporated within a state of wide reflective equilibrium.

1.6 – The Usefulness Objection

In this remaining section, and in the next chapter, we shall make further refinements to our understanding of Daniels' wide reflective equilibrium approach, such that we can be in a position to give a preliminary indication of what sort of commitments it is appropriate to assign to any would-be standard reflective equilibrium theorist. I shall then elaborate upon these commitments in the following chapter, arguing that it makes sense to assign a certain 'explicative project' to someone adopting Daniels' or a similar, standard approach.

We start by considering potential difficulty for the wide reflective equilibrium approach as it becomes more and more sophisticated, what we might refer to as the 'usefulness objection', which runs as follows. First we note that Daniels' approach, as intended to tell us something about justification (in ethics) has a commitment to wide reflective equilibrium being a *tractable* affair. Daniels is optimistic about the prospects of wide reflective equilibrium making "problems of theory acceptance in ethics more tractable and... [perhaps] produce greater moral agreement." (Daniels 1979, p.34) By investigating which particular moral conceptions may be upheld in a state of wide reflective equilibrium, we thereby have some consideration that will help to advance problems concerning which moral theories ought to be discarded, and which are acceptable. Furthermore, these considerations are independent of either the truth or truth-aptness of moral theory.[56] The aspiration is that we may be able to settle questions of justification even though questions concerning the truth of some moral conception seem to be beyond us. As Elijah Millgram puts it, "Coherence looks good because it is taken to be a usable criterion in a way that truth is not." (Millgram 2000, p.86)

However, the price of the wide reflective equilibrium account canvassed is that it takes us farther away from a genuinely *usable* set of criteria. As Stich intimates when considering 'bells and whistles' *wide* reflective equilibrium approach to the problem of finding justified inferential principles, it is unlikely that one would ever be in a

[56] Daniels is explicit about (wide) reflective equilibrium reflecting "coherence constraints on theory acceptance or justification, not on truth." (Daniels 1979, p.36) This seems consonant with Rawls' policy of 'avoiding' such meta-ethical concerns: "We try to avoid the problem of truth and the controversy between realism and subjectivism about the status of moral and political values." (Rawls 1980, p.395)

position to *ascertain* whether one was in a state of (wide) reflective equilibrium on some matter:

> A dubious virtue of both the wide reflective equilibrium and the expert reflective equilibrium accounts is that they make clear-cut counterexamples harder to generate. That is, they make it harder to produce actual examples of inferential rules which the analysis counts as justified and we do not. In the case of wide reflective equilibrium, counterexamples are hard to come by just because it is so hard to show that anything is in wide reflective equilibrium for anyone. ("Would she really continue to accept that rule if she thought through her epistemological and metaphysical views and reached some stable reflective equilibrium position?" Well, God knows.) (Stich 1990, p.85)

Daniels' wide reflective equilibrium approach requires that, in order for us to determine whether an agent is a state of wide reflective equilibrium, we need to assess whether they hold their moral principles, moral judgements, and relevant background theories (together with relevant philosophical arguments) in a coherent manner. At this point, it becomes relatively clear that actually carrying out such an evaluation is immensely cognitively and computationally taxing.[57] Even with considerable division of labour amongst various assessors, it seems unlikely that we'd ever be in a position to actually ascertain whether an agent was or was in a state of wide reflective equilibrium.[58] This, of course, is before one even considers immensely difficult questions concerning how relevant background theories constrain our moral principles and judgements. One might ask how, for instance, evolutionary psychological theories ought to guide our choice of moral principles? How far should they constrain, via stability considerations, our various moral commitments? Such questions, and myriad others, will be the sort of questions one would need to be able to answer if one were truly to ascertain the wide reflective equilibrium status of any agent holding some moral conception of even minimal complexity.

The usefulness objection, then, is that Daniels' characterisation of wide reflective equilibrium is actually unusable in practice. It looks unlikely that we'd ever be able to

[57] As soon as one considers the argument put forward by Christopher Cherniak, (Cherniak 1986, p.143n.13) this becomes clear. Cherniak considers how long it would take, using a truth-table method, to check the *logical consistency* (which is presumably a necessary, though not sufficient, component of coherence) of 138 beliefs. He finds that, even using a 'supermachine', one nucleon in diameter, working at the speed of light, there would not be enough time in the entire history of the universe to determine whether those 138 beliefs were logically consistent.

[58] One might, however, be able to determine that an agent does *not* hold their moral conception in wide reflective equilibrium, at least in relatively trivial examples involving obviously contradictory moral conceptions.

know in practice if some moral conception is held in wide reflective equilibrium, and this would then mean that wide reflective equilibrium does not make questions of moral theory acceptance any more tractable than considerations of truth. It is therefore unclear why one would be interested, in the first place, in asking whether an agent holds their moral conception in a state of wide reflective equilibrium.

However, the following sort of reply, along lines suggested by John Arras (Arras 2007, p.57) seems here to be effective. We might distinguish between an ideal notion of 'wide reflective equilibrium' and a 'rough-and-ready' decision procedure for use in practical contexts. We can thereby enact a division of labour of sorts: as far as actual *usefulness* goes, the demand is that some rough-and-ready decision procedure be created via which we have a usable reflective equilibrium test that helps makes theory acceptance a more tractable matter. As far as characterising the ideal of 'wide reflective equilibrium' goes, however, the purpose is to clarify what an agent being in wide reflective equilibrium involves.

We saw earlier that Rawls' characterisation of the method of (Rawlsian) wide reflective equilibrium contained various idealisations. Rawls, in his rendering of (Rawlsian) wide reflective equilibrium, is careful to present matters such that the method described reflects a 'philosophical ideal'. On Rawls' account of wide reflective equilibrium, one could only reach a state of (Rawlsian) wide reflective equilibrium[59] once that state incorporated beliefs concerning all possible alternative moral conceptions (with supporting arguments). Practically speaking, such a state looks to be beyond our epistemic grasp, since the number of considerations involved are bewildering, and we are not even in a position to delineate the range of considerations that we are even supposed to incorporate within the state of (Rawlsian) wide reflective equilibrium. However, Rawls suggests that we might *approximate* to such a state by actively considering *current* alternatives: "The most we can do is to study the conceptions of justice known to us through the tradition of moral philosophy and any further ones that occur to us, and then to consider these... Thus justice as fairness moves us closer to the philosophical ideal; it does not, of course, achieve it." (Rawls 1999, p.43) In effect, Rawls employs a working

[59] This is to read Rawls in a manner consistent with the standard interpretation of reflective equilibrium, whereby the *state* of reflective equilibrium is of primary importance. One could nonetheless understand the relevant idealisation here as that of describing a method where, as it were, one's work is never done, which is consistent with the alternative reading.

conception of 'reflective equilibrium' such that he only compares his theory with a few current, salient alternatives. As such, Rawls estimates that, in practice, all he achieves is a *provisional* justification. (Rawls 1999, p.46)

We do not (as yet) have any reason to think that some working conception of 'wide reflective equilibrium' could not be created such that it could function as a usable test in matters of theory acceptance. It is a legitimate request, for the reason that we wish the reflective equilibrium account to make moral theory acceptance more tractable, that such an understanding should be forthcoming. However, provided our conception of the philosophical ideal of wide reflective equilibrium is in good order, it is likely to be possible to design a further, working understanding of wide reflective equilibrium such that it could actually be usable 'on the ground', and such that it is a reasonable approximation to our *ideal* understanding of wide reflective equilibrium. As far as our ideal of 'wide reflective equilibrium' goes, then, we need not be able to *know in practice* that some agent is in wide reflective equilibrium. Any estimation here will be a provisional judgement made using a working conception of 'wide reflective equilibrium'.

However, that is not to say that this ideal conception of 'wide reflective equilibrium' faces no epistemological commitments. We might interpret Stich's objection, contained in the epithet "Well, God knows", in a somewhat different manner. The complaint may be that wide reflective equilibrium accounts are unusable *in practice*, but that they are unusable *in principle*. The problem is not so much that it is cognitively or computationally beyond our actual abilities, but that we would have no idea how one might, even in principle, go about ascertaining whether a state of wide reflective equilibrium obtains. This might be, say, because a state of wide reflective equilibrium requires incorporating all sorts of philosophical arguments and background theoretical considerations, and we have no clear idea of how one might do this.

The way to get clear on the difference between these ways of reading Stich's objection is to ask as to the reason why an ideal conception of 'wide reflective equilibrium' is unusable. Is it unusable because, as a matter of contingent fact, the sheer array of computational tasks is such that it is far beyond our abilities to

calculate? Or is it unusable because we would not know how where to start in order to calculate whether a state of wide reflective equilibrium obtained? In the former case, the ideal conception of 'wide reflective equilibrium' is unusable because the task outstrips our contingent abilities. If we had enough time and resources, we would be able to carry out such calculations. In the latter case, however, no amount of time or resources will help, because we do not know how to proceed. It is not merely unusable in practice, but unusable in principle.

This way of reading Stich's objection, it seems to me, does yield a genuine constraint on Daniels' wide reflective equilibrium approach. It is premature to demand of one's ideal conception of 'wide reflective equilibrium' that it is of practice use, but it needs to be the case that the ideal conception of 'wide reflective equilibrium' is such that the reasons it is not of practical use are contingent limitations of time and resources, as opposed to unclarities in the conception itself meaning that, even if we had unlimited time and resources, we would still not know how to proceed.

Given the way 'wide reflective equilibrium' is understood as some function (to be defined) involving the examination of the logical and evidential relationships between the various moral and non-moral beliefs an agent happens to hold, it needs to be the case that, once we properly define this function, we would know how one could go about ascertaining whether an agent holds their moral conception in wide reflective equilibrium or not. That is, we'd need to know how to settle the question as to whether a state of wide reflective equilibrium obtains even if, in practice, we were unable to carry out such an investigation. The ideal conception of 'wide reflective equilibrium' need not be practically useful, but it needs to be, in principle, a verifiable matter.[60] We need to have some conception of what is involved in a state of wide reflective equilibrium obtaining such that, if we had the time and resources, we would be able to ascertain whether it obtained or not.

We have, then, two preliminary constraints on Daniels' wide reflective equilibrium approach that need to be fulfilled:

1. In order to answer the garbage-in/garbage-out objection, it needs to be sufficiently difficult to generate plausible counter-examples in which an agent

[60] Another way of putting the point is to say that the obtaining (or not) of a state of (wide) reflective equilibrium cannot be a potentially verification-transcendent matter.

holds an evidently unjustifiable moral conception in (wide) reflective equilibrium.

2. In order to be in principle useful, it needs to be possible, given enough time and resources, for us to ascertain whether a state of wide reflective equilibrium obtains (or not).

I conclude this chapter by reinforcing my main point: that the discussion of reflective equilibrium has overwhelmingly adopted the standard interpretation of reflective equilibrium. Daniels' account of wide reflective equilibrium provides the paradigm example of the standard interpretation of reflective equilibrium, whereby reflective equilibrium is of primary interest *qua* coherence theory of justification in ethics. In the next chapter, I shall expand upon these two preliminary constraints, and argue that the way to model these constraints is to regard Daniels' wide reflective equilibrium approach as committed to a certain 'explicative project'. This will enable us to get clear upon at least two points that are currently very vague. Firstly, we shall examine the connection between reflective equilibrium and justification. Thus far, I have employed the awkward locution that reflective equilibrium is intended to 'tell us something about' justification in ethics. This clearly needs to be explained. Secondly, we shall get a clearer understanding of what it means to say that a state of (wide) reflective equilibrium 'obtains'.

Chapter 2 – Explication and Objectivity

2.1 – Overview

My aim in this chapter is to clarify further the commitments for a standard approach to reflective equilibrium, i.e. in which it functions primarily as a coherence theory of justification in ethics. To do this, I shall assign a certain 'explicative project' to the would-be standard theorist, a project which, whilst being sufficiently charitable, allows us to make sense of the connection between reflective equilibrium and justification, as well as making better sense of what thinking of reflective equilibrium as an epistemic state that obtains (or not) given the set of moral and non-moral beliefs an agent happens to hold.

This will then put us in a position to see, over the course of chapters 3-5, that this 'explicative project', given a certain analysis of moral discourse in which problematic cases of moral disagreement are sufficiently prevalent, is misguided. The upshot is, as we shall see, that we ought to abandon the idea that reflective equilibrium could function as a coherence theory of justification in ethics, and consequently to abandon the standard view of reflective equilibrium in which a recommendation of the method of reflective equilibrium derives from the role of the method in attaining the distinctive epistemic state of being in reflective equilibrium.

In this chapter, I begin by elaborating upon Stich's understanding of reflective equilibrium approaches, since, although his concern is with justified inference and thus, does not count as a 'reflective equilibrium approach' on my definition (see §1.2), he outlines what I take to be an important and very helpful point: that reflective equilibrium approaches are best understood as attempts at *explicating* some everyday notion of justification, as opposed to a more traditional conception of 'conceptual analysis'. An explication is, very roughly,[61] where one replaces an existing, everyday notion (an agent's 'being justified' in believing *x* and *y*) with a more technical counterpart (an agent's believing *x* and *y* in reflective equilibrium). I then examine Carnap's remarks on explication, in order to get a clearer idea of what minimal conditions an explication needs to satisfy in order to count as 'satisfactory'.

[61] See §4.3 for a more accurate characterisation.

The overall point here is that the notion of 'explication' clarifies, in a charitable way, the intended relationship, hitherto gestured at with the awkward locution 'tells us something about', between being in a state of reflective equilibrium and being justified.

I focus on the notion of exactness, and relate it to a particular notion of objectivity, 'disagreement-entailing-error objectivity'. We note an important connection between explication and objectivity here: the point of an explication being exact is that it renders an area of discourse transparently objective. If we can replace an everyday notion that is inexact (in some sense we need to specify) with a technical counterpart that is exact, we have a method whereby we can agree what settles any relevant question in that area of discourse, which relates to the notion of objectivity I adopt.

In the final section, I explain how all of this pertains to Daniels' wide reflective equilibrium approach, or indeed any standard approach to reflective equilibrium in which 'reflective equilibrium' primarily refers to the obtaining of an epistemic state of coherence (in an agent's moral conception). Here the aim is to characterise, again in a charitable manner, what it means to regard a state of reflective equilibrium as 'obtaining'. Once we do so, we will be in a position to understand the commitments that the would-be standard reflective equilibrium theorist needs to uphold in order for their approach to remain viable.

2.2 – Stich on Explication

As we saw in the previous chapter, Stich's concern with reflective equilibrium approaches pertains to rules of inference as opposed to moral theories. Nonetheless, we saw that his mode of argument, making a garbage-in/garbage out objection through the use of plausible counter-examples (such as the gambler's fallacy), was such that one could easily make an analogous argument for ethical cases. Just as Stich came up with plausible counter-examples where 'daffy' (i.e. unjustifiable) inferential rules could be held by some people in (narrow) reflective equilibrium, we can come up with similar counter-examples in the moral case, i.e. 'daffy' moral principles that could be held by some people in reflective equilibrium.

It is worth pursuing Stich's characterisation of reflective equilibrium approaches further, since he connects them with broader meta-philosophical concerns that prove helpful. He sees reflective equilibrium accounts as part of a more general tendency within philosophy, a tendency he terms 'analytic epistemology'. This is defined as "Any epistemological project that takes the choice between competing justificational rules or competing criteria of rightness to turn on conceptual or linguistic analysis." (Stich 1990, p.91) So, an analytic epistemological project is one that attempts to make headway by taking an everyday concept[62] like *justification* and provide some sort of elaboration or analysis of what such a concept might mean. Stich understands Goodmanian or neo-Goodmanian[63] reflective equilibrium very much in this vein, as

[62] In what follows, our discussion shall, following Stich and later Carnap, sometimes be couched in terms of 'concepts', and sometimes be couched in terms of 'notions', 'ideas', or 'terms'. As is the normal convention, I indicate where a concept is being discussed by italicising it, e.g. the concept *justification*, and where a term or notion is being discussed by placing it in inverted commas, e.g. the notion of 'reflective equilibrium'.

One might think that eliding between 'notions' and 'concepts' represents something of a sleight of hand, because the term 'concept' is often understood in philosophy to refer to mental content, whereas a 'notion' typically refers to a semantic item. However, an explicative project, at least as conceived of by Stich and Carnap, does not concern itself with mental content as such. As far as I am able to discern, where Stich and Carnap use the term 'concept', they are not using the term to pick out *mental* contents, but are using the term in a metaphysically neutral way, roughly synonymous with 'the underlying idea picked out by terms like 'x' in language L_1, or 'y' in language L_2, etc.' Here, the utility in talking about 'concepts' as opposed to 'terms' is simply that concepts, unlike terms or notions, are not tied to a particular language. Stich, for instance, in speaking of the concept of *justification*, is not really referring to the mental content *justification*, but is just picking out the idea that one refers to in English as 'justification'. I take it, then, that one may, to all intents and purposes, use the notions of 'concept', 'notion', 'term', etc. interchangeably.

[63] The account of 'reflective equilibrium' Stich is here stalking comes from Goodman's *Fact, Fiction, and Forecast*. (Goodman 1983) There, Goodman offers the following guidance in coming up with justified inferential rules:

an intended 'conceptual analysis' of our everyday or 'folk' views on what it is to justify a general principle governing a certain practice, whether inferential or moral reasoning.

Stich offers us some intriguing guidance as to exactly what form this sort of 'conceptual analysis' might take. Stich puts it that, by regarding reflective equilibrium as a 'conceptual analysis' of *justification*, we might mean one of three things:

1. The claim is a *conceptual truth* – that it follows from the meaning of 'justification' or from the analysis of the concept of justification. Like other conceptual truths, it is both necessarily true and knowable *a priori*.
2. The claim is a nonconceptual necessary truth that is knowable only *a posteriori*. This would accord it the same status that some philosophers accord to the claim that water is H_2O.
3. The claim is being offered as a stipulative proposal. It is not telling us what our preexisting concept of justification amounts to, nor what is essential to the referent of that concept. Rather, in a revisionary spirit, it is proposing a new notion of justification. (Stich 1990, pp.78-9)

It is safe to say that the first option has fallen out of favour in philosophy. However, Stich maintains that 'conceptual analysis' has nonetheless survived in the form of explication, an amalgam of the first and third option:

> Actually, the divide between the first and last of these alternatives is not all that sharp. For one might start with an analysis of our ordinary notion and go on to propose modifications in an effort to tidy up the notion a bit here and there. As the changes proposed get bigger and bigger, this sort of explication gradually shades into pure stipulation. So long as the changes an explication urges in a preexisting concept are motivated by considerations of simplicity and don't result in any radical departure from the ordinary concept, I'll count it as a kind of conceptual analysis. I think a good case can be made that Goodman took himself to be providing such a conservative explication. (Stich 1990, p.79)

As Stich see it, then, there is a spectrum of ways in which we might be engaged in a method worth regarding as 'conceptual analysis'. On one end of the spectrum is a traditional conceptual analysis in which there is an intensional isomorphism between

---

"*A rule is amended if it yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend.* The process of justification is the delicate one of making mutual adjustments between rules and accepted inferences; and in the agreement achieved lies the only justification needed for either." (Goodman 1983, p.64)

A 'neo-Goodmanian' account, for Stich, is one that attempts to attach 'bells and whistles' to this approach, and includes any wide reflective equilibrium approach. (Stich 1990, pp.86-7)

the analysans and analysandum. "A triangle is a 2-dimensional polygon with 3 sides" might be such an example. 'Triangle' and '2-dimensional polygon with 3 sides' are held, not only to, in actuality, pick out the same objects, but to do so across all possible cases, since they are intended to mean the same thing. On the other end of the spectrum are purely stipulative concepts, where there is no corresponding everyday concept. "*i* is the square-root of -1" might be such an example. In between these two ends of the spectrum, however, lie a range of possibilities, ranging from conservative to eliminative,[64] in which the concept to be explicated, the *explicandum,* is captured using some *explicatum* (the explication itself).[65] As the explicatum tends towards stipulation, it functions as *eliminative* explication;[66] as the explicatum tends towards respecting the intension of the explicandum, it functions as *conservative* explication.

In terms of reflective equilibrium, Stich has it that "Goodman is proposing an explication of our ordinary notion of a justified inference: to be justified is to be sanctioned by a set of inferential rules that pass the reflective equilibrium test." (Stich 1990, p.79) As an attempt at conservative explication, the Goodmanian/neo-Goodmanian line with regards to 'reflective equilibrium', as Stich sees it, is to pursue a middle course between elimination of discourse involving the notion of 'justification' and traditional conceptual analysis in which one attempts to wholly preserve, whilst shedding light upon, the meaning of 'justification'. A conservative

---

[64] Stich acknowledges (Stich 1990, p.164n.6) that the contrast between 'conservative explication' and 'eliminative explication' is taken from Brian Loar's *Mind and Meaning*. (Loar 1986) Loar describes an 'explicative dilemma' in which one is faced with a certain area of informal discourse together with a theoretical framework where one has an 'imperialist inclination' to regard as being able to capture all relevant truths within that informal discourse. At this point, we can either *eliminate* the informal discourse along lines suggested by Quine (see ft.66), or we can achieve a 'conciliation' between the theoretically-informed explication and the informal discourse, which is referred to by Loar as *conservative* explication. For technical purposes, we accept the explication, but it does not thereby mean that we abandon the 'folk' discourse in such an area:

"Imagine two theorists early in the century who react to Russell's analysis of 'number' as "maximal set of conumerous sets", in these two ways. Both accept the replacement, but one feels his ontology has changed, while the second does not. Or imagine two philosophers who react to compatibilism about free choice (in terms of counterfactuals about wants and decisions) in the following ways. Both accept the replacement, but one takes his view of the human condition to have changed, and the other does not... For the latter member of each pair the replacement counts as a *conservative explication*, while for the former it is more or less radical." (Loar 1986, p.43)

[65] Here I follow Carnap's terminology. See §2.3 below.

[66] This position is taken by Quine. For Quine, "*Explication is elimination.* We have, to begin with, an expression or form of expression that is somehow troublesome. It behaves partly like a term but not enough so, or it is vague in ways that bother us, or it puts kinks in a theory or encourages one or another confusion. But also it serves certain purposes that are not to be abandoned. Then we find a way of accomplishing these same purposes through other channels, using other and less troublesome forms of expression. The old perplexities are resolved." (Quine 1976, p.260)

explication of 'justification' is instead intended as a subtle change of question that nonetheless sheds light on the original question. Instead of asking ourselves whether some set of inferential rules are justified, the idea is to ask ourselves instead whether those inferential rules are held in reflective equilibrium, and treat this outcome as thereby telling us whether they are justified.

One may wonder here as to the advantages of regarding a reflective equilibrium account as explicative, rather than a more traditional conceptual analysis (Stich's first option). Here we need to briefly remind ourselves of the constraints imposed by the garbage-in/garbage-out objection on a reflective equilibrium account. If one were to opt for regarding 'reflective equilibrium' as a traditional conceptual analysis of 'being justified in holding a moral conception',[67] then it would need to be *a priori* impossible to come up with counter-examples whereby some agent held their palpably unjustifiable moral conception in reflective equilibrium. It would be a constraint of *a priori* invulnerability to counter-examples, rather than sufficient invulnerability to plausible counter-examples.

However, this leaves it open to the standard reflective equilibrium theorist, pursuing a coherence theory of justification, to claim that the constraint imposed here, *a priori* invulnerability to counter-examples, is overly onerous. Suppose one found, for instance, that one's reflective equilibrium account (once suitably fleshed out) provided one with a large set of responses to particular moral conceptions that were identical to one's best intuitive judgements about whether an agent could justifiably hold such moral conceptions. The analysis, by any reasonable standard, seems to work. Suppose, however, that some counter-example were dreamed up that, despite being incredibly unlikely, perhaps even physically impossible, to occur, the responses to moral conceptions given via the reflective equilibrium account and via one's best judgements as to whether that moral conception is justifiably held part company. On the basis of this one unlikely counter-example, the *a priori* invulnerability constraint would require of us to reject the reflective equilibrium account. It is, however, absurd to claim that, in such a scenario, the reflective equilibrium account provides no conceptual guide of some sort to justification in ethics. Requiring that it be a traditional conceptual analysis, then, seems unduly restrictive.

[67] I return here to considering reflective equilibrium as a coherence theory of 'being justified' in *ethics*.

For this reason, we can now see why it is more appropriate to use the constraint that a standard reflective equilibrium account like Daniels' needs to be *sufficiently* invulnerable to *plausible* counter-examples. This represents a constraint that is genuinely charitable to the would-be standard reflective equilibrium theorist. Given that my intention is to build up an understanding of the project to which a standard reflective equilibrium theorist is genuinely committed, viewing 'reflective equilibrium' as (conservatively) *explicative* is preferable. An explicative account, in virtue of the fact that it occupies this middle ground between analysis and stipulation, need not preserve either the intension or extension of the everyday concept. All it need do, as we shall see in the next section, is be *sufficiently similar*. The explicatum needs to be sufficiently similar to the explicandum, but this does not entail that explicatum and explicandum need never part ways.[68]

[68] Curiously, Stich seems to have missed this point himself, when he argues that even the *possibility* of such counter-examples "poses a serious problem for the Goodmanian story", since "It is surely not an a priori fact that strange inferential principles will always fail the reflective equilibrium test for all subjects." (Stich 1990, p.84) However, by his own lights, as understanding the Goodmanian as putting forward a conservative explication (suggested merely five pages previously), the possibility of such counter-examples existing is not a serious problem at all. It is only once such counter-examples become both plausible and sufficiently prevalent that doubts as to the satisfactoriness of the explication emerge.

2.3 – Carnap on Explication

The connection between 'reflective equilibrium' and everyday notions of justification, then, is this: 'reflective equilibrium' is to be understood as *explicating* what it is for an agent to be justified in holding some moral conception. I take it that this construal of the (standard) reflective equilibrium approach already represents an improvement upon regarding it as one of traditional analysis. It is sufficiently tolerant such that isolated, implausible counter-examples do not automatically scupper the explication.

We need now to delineate the constraints that apply to explication, and gain a clearer understanding of how explication works. In order to do so, it is helpful to examine Rudolf Carnap's remarks on the nature of explication, found (primarily) in *Logical Foundations of Probability*.[69] (Carnap 1962) Carnap's exposition is exceptionally clear, utilises helpful terminology, and is furnished with effective examples that well-illustrate his intent. The nature of explication is summarised by Carnap in the following way:

> The task of ***explication*** consists in transforming a given more or less inexact concept into an exact one or, rather, in replacing the first by the second. We call the given concept (or the term used for it) the ***explicandum***, and the exact concept proposed to take the place of the first (or the term proposed for it) the ***explicatum***. The explicandum may belong to everyday language or to a previous stage in the development of scientific language. The explicatum must be given by explicit rules for its use, for example, by a definition which incorporates it into a well-constructed system of scientific either logicomathematical or empirical concepts. (Carnap 1962, p.3 [emphases in original])

In utilising the term 'explicatum', Carnap here deliberately avoids mirroring the corresponding Latinised terms regarding 'analysis', i.e. analysandum and analysans,

[69] According to Michael Beaney, (Beaney 2004, p.133) Carnap first used the notion of 'explication' in 'The Two Concepts of Probability'. (Carnap 1945, p.513) Carnap also briefly invokes the notion in *Meaning and Necessity* (Carnap 1970), published in 1947, three years prior to the publication of *Logical Foundations of Probability*. There he utilises the basic idea of explication as *replacing* an insufficiently 'exact' concept with a more 'exact' concept:

"The task of making more exact a vague or not quite exact concept used in everyday life or in an earlier stage of scientific or logical development, or rather of replacing it by a newly constructed, more exact concept, belongs among the most important tasks of logical analysis and logical construction. We call this the task of explicating, or of giving an explication for, the earlier concept." (Carnap 1970, pp.7-8)

In *Logical Foundations of Probability*, however, Carnap devotes an entire, albeit short, chapter to the clarifying the motivations behind and nature of explication. For this reason, like Beaney, I shall focus on this work.

for the reason that, in some cases at least, an explicatum "deviates deliberately from the explicandum but still takes its place in some way" (Carnap 1962, p.3). Thus, he wishes to signal the thought that what is intended by 'explication' *may* in some cases count as something worth calling 'analysis', but not necessarily. Consonant with Stich's use of the term, the notion of 'explication' is thus broader than that of traditional conceptual analysis. An explication is intended to serve as a *replacement* for an everyday notion, and it is not necessary for the explicatum to mirror, either intensionally or extensionally, the explicandum.

How, then, would one proceed, if one wanted to explicate some everyday notion? Here, the first step is to *clarify* the explicandum:

> There is a temptation to think that, since the explicandum cannot be given in exact terms anyway, it does not matter how much we formulate the problem. But this would be quite wrong. On the contrary, since even in the best case we cannot reach full exactness, we must, in order to prevent the discussion of the problem from becoming entirely futile, do all we can to make at least practically clear what is meant as the explicandum... An indication of the meaning with the help of some examples for its intended use and other examples for uses not now intended can help the understanding. An informal explanation in general terms may be added. (Carnap 1962, p.4)

Carnap illustrates what such clarificatory work involves via the use of examples. If one wished to explicate the notion 'true', for instance, we'd need to explain what sense of 'true' we were tracking. So, we might clarify that by 'true' we do not mean the sense of 'true' involved in being a 'true friend', having a 'true aim', etc., but rather mean true in the sense of veridical. Similarly, in explicating the notion of 'salt', we might clarify that we mean to talk about the stuff that people sprinkle on their food, as opposed to its broader use in chemistry. It is only once we are clear concerning what general sense of 'true' or 'salt' we wish to take as the explicandum that we are then in a position to attempt to supply some explicatum. For 'true', we then offer up the explicatum TRUE,[70] which might be supplied by a formal theory such as Tarski's understanding of truth as satisfaction. For 'salt', we offer up the explicatum N_AC_l, which needs to be supplied as part of chemical theory more generally, i.e. explaining

[70] Here I wish to adopt the following convention – when I refer to an explicatum, I shall use small capitals; when I refer to the explicandum, I shall place the term within quotation marks. Thus, TRUE explicates 'true', PRIME explicates 'prime number', LIFE-FORM explicates 'life-form', etc.

ionic bonding, atomic theory, etc. In both examples, we are restricting the scope of the explicatum to a particular area.

Once we have a clarificatory understanding of the explicandum, we then wish to provide a *satisfactory* replacement for the explicandum. Such a *satisfactory* explication, according to Carnap, needs to satisfy the following requirements:

1. The explicatum is to be *similar to the explicandum* in such a way that, in most cases in which the explicandum has so far been used, the explicatum can be used; however, close similarity is not required, and considerable differences are permitted.
2. The characterisation of the explicatum, that is, the rules of its use (for instance, in the form of a definition), is to be given in an *exact* form, so as to introduce the explicatum into a well-connected system of scientific concepts.
3. The explicatum is to be a *fruitful* concept, that is, useful for the formulation of many universal statements (empirical laws in the case of a nonlogical concept, logical theorems in the case of a logical concept).
4. The explicatum should be as *simple* as possible; this means as simple as the more important requirements (1), (2), and (3) permit. (Adapted from Carnap 1962, p.7)

Evidently, we need to discuss these requirements in detail, which I shall now do in reverse order. Regarding the fourth requirement, however, the requirement of simplicity, there may not be a great deal that can be said. Carnap makes it clear that it is the lesser of the four requirements, and takes it that it need only be invoked where one is faced with competing explicata that are indistinguishable on the basis of the other criteria. Noting the notorious difficulty of characterising considerations of simplicity, Carnap leaves this requirement at a fairly intuitive level.[71] Since such considerations do not play a key role in subsequent argument, I am content to follow Carnap in this regard.

The third requirement is also of lesser importance with regards to subsequent argument, but it is nonetheless worth elaborating upon. Carnap explains the fruitfulness requirement using an example, explicating 'fish' with the explicatum PISCIS:

[71] "The simplicity of a concept may be measured, in the first place, by the simplicity of the form of its definition and, second, by the simplicity of the forms of the laws connecting it with other concepts.... In general, simplicity comes into consideration only in a case where there is a question of choice among several concepts which achieve about the same and seem to be equally fruitful; if these concepts show a marked difference in the degree of simplicity, the scientist will, as a rule, prefer the simplest of them." (Carnap 1962, p.7)

> A scientific concept is the more fruitful the more it can be brought into connection with other concepts on the basis of observed facts; in other words the more it can be used for the formulation of laws. The zoologists found that the animals to which the concept Fish applies, that is, those living in water, have by far not as many other properties in common as the animals which live in water, are cold-blooded vertebrates, and have gills throughout life. Hence the concept Piscis defined by these latter properties allows more general statements than any concept defined so as to be more similar to Fish; and this is what makes the concept Piscis more fruitful. (Carnap 1962, p.6)

In order to be a genuinely fruitful explication, an explication needs to connect with a broader set of concepts other than the one being explicated. Putting forward PISCIS as explicating the notion of a 'fish' is not something that can be done in isolation of explicating various other concepts that are integral to the zoological language. Fruitfulness requires that PISCIS connects to other notions, for which we will then need explicata, e.g. 'gills', 'vertebrate', 'cold-blooded', etc. As an explicative project develops, it will connect with other notions that then stand in need of explication. An explication of 'salt' would involve connections to broader chemical theories, to use our earlier example. As Andre Carus, a prominent Carnap scholar, puts it, "Explications, then, do not just replace ordinary-language explicanda one by one; the entire system of interrelations holding them together is also replaced gradually by the provisionally canonical languages of science, giving rise to entirely new concepts (such as "cholesterol" or "tectonic plate") that have no obvious ordinary-language predecessors at all." (Carus 2007, p.41) Explications will tend towards *systematising* an area of discourse, rather than making piecemeal alternations to individual notions. In Carnap's view, we will tend towards the construction of artificial, systematic languages.

The requirement of fruitfulness also gives us some clue as to the form a particular explicatum might take, as Carnap requires that explicata facilitate universal generalisations, either as "empirical laws in the case of a nonlogical concept, [or] logical theorems in the case of a logical concept". Clearly, in the case of PISCIS, what we have is a definition that allows us, following empirical observation, to classify an animal as a PISCIS or not, which then allows us, with empirical laws such as "All PISCIS evolved in the following way…" (say), to make fruitful generalisations in a way that would not be possible with the everyday notion of a 'fish'. For purely logical

notions, however, an explicatum involves supplying a formal, logical theorem. One might explicate the notion of a 'number' using axiomatic set theory, for instance, where at no point need one engage in some empirical observation as such.[72]

This need for explicata to be systematic is also explicitly mentioned in Carnap's second requirement, that an explicatum be sufficiently *exact* "so as to introduce the explicatum into a well-connected system of scientific concepts." The reason for this connection is not hard to appreciate: if, at any point, it is unclear whether some animal ought to count as a PISCIS or not, the explicatum PISCIS is of lesser value to a scientific, classificatory endeavour. One could not, for instance, form any useful generalisations about PISCIS or hypotheses, say, about how certain features represent evolutionary adaptations or the like if there are sufficiently many cases where it is unclear just which animals are in fact PISCIS or not. If the explication is not itself sufficiently exact, then there is little advantage to be had in utilising the explicatum as opposed to the explicandum, and we would not be able to use the explicatum in the generation of universal laws or logical theorems.

This underlines what Carnap takes to be the central point regarding the purpose of explication. As Carnap sees it, the primary purpose of explication is that it allows us to avoid the *vagueness*[73] of everyday concepts.[74] Vagueness, for Carnap, is the antonym of exactness. Explication, as a method of making our terms exact, is a method for avoiding such vagueness. With such conceptual engineering, we can transform an

---

[72] This proves relevant for when we're thinking about how we might design a REFLECTIVE EQUILIBRIUM explicatum. Here, I take it that the most natural way to understand reflective equilibrium as explicative is to think of it as a formal function operating over the set of (moral and non-moral) beliefs that an agent happens to hold at the time of evaluation. See §2.5.

[73] I examine the notion of 'vagueness' in §2.4, because it is important to my overall argument that the requirement of exactness is spelled out as fully as possible. For now, we need to operate with a fairly intuitive understanding of vagueness in order to get a general sense of Carnap's criteria for satisfactoriness of explication.

[74] Carus emphasises the centrality of avoiding vagueness through the explicating an informal term, even including it in his definition of explication itself – "Explication, which in Carnap's view is the main task of conceptual engineering, consists in the *replacement* of a vague concept – the *explicandum* – by a more precise one, the *explicatum*." (Carus 2007, p.40) Throughout his book *Carnap and Twentieth-Century Thought*, Carus draws the reader's attention to the way Carnap diagnoses philosophical problems as arising from the vagueness inherent in everyday language. Particularly revealing is the following private reflection of Carnap's that Carus picks out:

> "I do not share the apparently widespread view that the vagueness and ambiguity of most words in everyday language do not much interfere with human communication. It is hard for me to understand how someone could really believe this, in view of the countless misunderstandings and failures to get something across that we observe daily. I would have thought that there could be no disagreement about the damage done by this vagueness." (Carnap, quoted in Carus 2008, p.276)

area of discourse which is unduly vague into a sufficiently precise and exact form of discourse, such that it can be used in the formulation of universal laws and theorems.

More needs to be said here about vagueness and exactness, which I address in the next section, but for now we can address one issue with regards to *how exact* an explicatum needs to be. Need an explicatum, in order to count as 'satisfactory', be so exact that no vagueness at all is present, or is some vagueness allowable? As we've seen, Carnap describes the process of explication as "transforming a given more or less inexact concept into an exact one", which suggests that an explicatum is 'exact' once it reaches a certain threshold. Carnap also tells us that "the explicandum is more or less vague and certainly more so than the explicatum" (Carnap 1962, p.5), which suggests that *some* degree of vagueness may be allowable in the explicatum.

Carus suggests that the way to understand Carnap here is to regard particular explications as falling on a spectrum where we have some *ideal* standard of exactness in mind. The ideal is that the explicandum is replaced by some explicatum where the application of that notion is entirely governed by rules internal to some constructed language. That is, where every application of the explicated term is exact, and where the rules internal to the constructed system of explicated terms allow us to ascertain, on any given occasion, whether the term is to be applied or not. However, actual explications may well fall short of this ideal:

> [An explicatum is] assumed to be defined within the context of a formal language which, ideally (though perhaps not in every actual case), establishes its meaning under all possible *internal* uses, within the (constructed) explication language. In Carnap's view this is just what the important difference between explicandum and explicatum consists in: the one is vague and ill defined, while the other is more precise and *better* defined. The dynamic feedback process of explication makes the scientific vernacular as a whole more self-conscious and deliberate; fewer of its concepts are passively accepted from fashion cascades or folk wisdom. (Carus 2008, pp.286-7)

The ideal standard of 'exactness', then, is such that the explication allows us to make the application of the explicatum completely transparent. All possible uses within that constructed language are such that there are no borderline cases, and that the correct application of the term is always obvious. There should be no 'clots' of vagueness or indeterminacy within an explicated language. (Carus 2008, p.282)

Nonetheless, this remains an *ideal*. In practice, some vagueness may nonetheless be present.

Having such *exactness* as a regulative ideal means that Carnap is relatively permissive when it comes to his first requirement for satisfactoriness of explication, that there be *similarity* between the explicandum and the explicatum: "Since the explicandum is more or less vague and certainly more so than the explicatum, it is obvious that we cannot require the correspondence between the two concepts to be a complete coincidence." (Carnap 1962, p.5) One may, as we've seen, *elucidate* or *clarify* the explicandum to some extent such that we're clearer about exactly what it is we're intended to explicate, but it will always prove difficult to assess just how 'similar' the explicandum and explicatum are, since the everyday concept is, *ex hypothesi*, vague.[75] Where we have vagueness, it will be possible that there will be many *interpretations* of some explicandum that are equally acceptable. This is one reason why we need to regard explications as *proposals* rather than *claims*,[76] and explications as being satisfactory or not, as opposed to correct or not.[77]

Carnap illustrates this point using an example: explicating everyday notions of 'cold', 'colder', 'warm', 'warmer', etc. using TEMPERATURE, (Carnap 1962, pp.12-15) defined procedurally as just whatever reading a functioning thermometer gives us. Rather than using some explicata like WARM, COLD, etc., we replace such notions with the qualitative notion of TEMPERATURE, which we can measure using a particular

---

[75] This relates to avoiding the problem often referred to as the 'paradox of analysis', whereby it is unclear how an analysis could be both informative and yet preserve the meaning of the analysandum. As Beaney notes:

> "What is important is the tension that underlies *any* reconstructive project. On the one hand, the work of analysis is to elicit and clarify what we already know, and we cannot depart too radically from our ordinary understanding on pain of clarifying nothing at all. On the other hand, there must be a certain amount of reconstruction and revision, since our ordinary understanding is frequently confused and unreliable. (Beaney 2004, p.124)

An explication, then, avoids the problem by weakening the constraint of preserving the meaning of the explicandum to that of there being sufficiently similarity between the explicatum and the explicandum.

[76] The ideal of explication, then, connects with Carnap's *principle of tolerance*, that philosophers and scientists should be free to propose any sort of constructed language, and that these proposals should be judged essentially on grounds of their *utility* for particular purposes. As Carus puts it, "The transformation of philosophical doctrines into proposals regarding the form of language rests on the principle of tolerance. If there were a 'correct' language, then there could be no proposals, but only claims." (Carus 2008, p.261)

[77] Hence also the emphasis on the *satisfactoriness* of an explication, as opposed to its 'correctness': "Strictly speaking, the question whether the solution is right or wrong makes no good sense because there is no clear-cut answer. The question should rather be whether the proposed solution is satisfactory, whether it is more satisfactory than another one, and the like." (Carnap 1962, p.4)

instrument, e.g. a mercury thermometer. There are often cases where we will be unsure whether some object is 'warmer' than another, for instance, due to the vagueness of 'warmer'. Not so, however, with the TEMPERATURE of the object, since we have a method (using the thermometer) that allows us to ascertain the TEMPERATURE of the object. Crucially, however, and as Carnap notes, the explicatum will not precisely mirror our ordinary use of 'warmer' or 'colder', say. Our ordinary sensations with regard to 'colder' and 'warmer' are subject to various perceptual changes. We find things to be unbearably hot when we've come from a cold environment, internal body chemistry affects our perception, our perception of 'warm' weather is highly dependent on wind-speed, etc. It seems reasonable to assume that there will be differences between our everyday notions, 'colder', 'warm', etc., and the corresponding explicata. Nonetheless, "Experiences of this kind do not at all lead us to the conclusion that the concept Temperature defined with reference to the thermometer is inadequate as an explicatum for the concept Warmer." (Carnap 1962, p.12) We accept that TEMPERATURE is in various ways *dissimilar* to our everyday notions of warmth, but we nonetheless accept the explication because it is sufficiently similar, and it proves useful to us through its exactness and fruitfulness.

2.4 – Exactness and Objectivity

We have some idea, then, of the criteria an explicatum needs to satisfy in order to count as satisfactory. However, we need to say much more about the requirement of exactness, since it plays a central role in my overall argument. We've already seen that for Carnap the central motivation for an explicative project is to remove vagueness, but we need to be clearer as to the importance, and nature, of doing so.

We need to be clearer, first of all, about the notion of 'vagueness'. According to Mark Sainsbury and Timothy Williamson, we may have one of two things in mind here, depending on whether we are construing vagueness as a semantic idea, or an epistemological one. On a semantic reading, "It is of the nature of a vague predicate to draw no sharp boundary between the things to which it applies and those to which it does not." (Sainsbury & Williamson 1998, p.470) Here, it belongs to the meaning of a term, say 'heap', that it admits of borderline cases where the term is indeterminate in nature. Let's say we're examining cases involving *x* grains of sands. Where *y* indicates whether it counts as a 'heap', the (correct) extension of 'heap' [*x,* y] would look something like this: {[0,N], [1,N], [2,N] … [25,…], [26,…],[27,…] … [100,Y], [101,Y], …}. Here, around 25 grains of sands represents a borderline case where it is neither a heap nor not a heap. There is no standard of correctness here. There will be occasions where it is clearly appropriate to attempt to apply the term, and yet the meaning of 'heap' is such that no standard of correctness exists in such cases. The notion of a 'heap' is indeterminate in nature since there are gaps in its (correct) extension.

From the indeterminacy of 'heap', it follows the correct extension of 'heap' is not *effectively decidable*.[78] That is, we have no method available that could, given enough time and resources, allow us to ascertain correctly whether any given collection of grains of sands constitutes a 'heap' or not. Since there are certain numbers of grains

[78] "A property/relation is *effectively decidable* iff there is an algorithmic procedure that a suitably programmed computer could use to decide, in a finite number of steps, whether the property/relation applies in any given case." (Smith 2007, p.9)

Clearly, the notion of 'effective decidability' is geared towards logical cases where one is attempting to decide something purely in formal terms, but one could have a similar notion for cases where the relevant notion turns upon empirical observation. Essentially, one would be working with a verifiability principle. The extension of 'badger', for instance, is 'effectively decidable' if and only if it is in principle possible to verify whether any given animal counts as a 'badger' or not.

of sands where there is no standard of correctness available for whether it counts as a 'heap', it follows that there is no method available that could decide, even under ideal circumstances, any given case. If the explicatum is indeterminate, it will not be possible to effectively decide the correct application of the explicatum. By contraposition: if we do have such a method available, 'heap' is a determinate notion.

On the semantic understanding of vagueness, then, if a predicate is vague, it will be both indeterminate and (consequently) not effectively decidable. However, one may alternatively understand 'vagueness' solely as an epistemological concern, where one only makes a claim about whether any given competent language-user (and thus in command of the predicate) is in a position to effectively decide whether, e.g., some collection of grains of sand counts as a 'heap' or not. That is, 'heap' is epistemologically vague if and only if we have no method available that would, given enough time and resources, would allow us to ascertain correctly whether any given collection of grains of sands constitutes a 'heap' or not.

On an epistemological understanding of vagueness, it is open as to whether one has semantic vagueness. That is, being unable to effectively decide whether something counts as a 'heap' does not imply that the meaning of 'heap' is indeterminate (due to vagueness). It is possible that the (correct) extension of 'heap' looks like this: {[0,N], [1,N], [2,N] … [25,N], [26,Y],[27,Y] … [100,Y], [101,Y], …}. Or it could look like this: {[0,N], [1,N], [2,N] … [25,N], [26,N],[27,Y] … [100,Y], [101,Y], …}. In either case, there are no gaps in the respective extensions (they are determinate) and so there is always a standard of correctness for whether 'heap' is applied correctly or not. However, if there is epistemological vagueness, we lack a method that allows us to decide which determinate extension is picked out by 'heap'. 'Heap' may be determinate, but we are not in a position to *know* which determinate extension ought to count as the correct application of 'heap'.[79] We lack the required fineness of discrimination.

How does this relate to Carnap's ideal requirement of exactness for an explicatum? Carnap's preoccupation is with vagueness in its epistemological guise, where we have everyday notions that are vague in the sense that any given competent language-user

[79] This is the position put forward by Williamson in his *Vagueness*, (Williamson 2005, pp.185-247) and Roy Sorensen in *Blindspots* (Sorensen 1988, pp.246-52).

is left unsure how to apply such notions in certain borderline cases. Whether this is because the predicate is vague at a semantic level is something of a moot point for Carnap. The point of explication is to construct a system of terms where we are confident in applying our terms, i.e. where such epistemic borderline cases ideally do not arise at all. Carnap requires, ideally, that an explicatum be exact in the sense that it allows us to effectively decide whether, for all appropriate cases (i.e. all items in the domain of the explicatum), the explicatum (correctly) applies. There ought to be no occasions such that we are unable to decide, given enough time and resources, whether such-and-such is a PISCIS, or whether α is of higher TEMPERATURE than β. The explicatum needs to be in the form of an effective procedure. It needs to set out rules, in the definition of the explicatum, such that we have a procedure such that we can effectively decide whether the explicatum applies or not. Since an effectively decidable explicatum will be semantically determinate, Carnap's ideal for an explicatum is that it is be also a determinate notion. The ideal explicatum, for Carnap, will be both determinate and its correct application effectively decidable.

Take again Carnap's example of an everyday, informal concept that suffers from (epistemological) vagueness – 'warmer'. If we ask ourselves whether α is warmer than β, there are situations in which we run into difficulties. There will be borderline cases where any given competent language-user will be unsure as to whether it is correct to say "α is warmer than β" or not. In particular, we will be unsure just how much 'warmer' α needs to be than β for it to be correct to apply the notion of 'warmer', or we may be unsure whether our particular estimation of whether α is warmer than β is the subject of some perceptual illusion. As such, the extension of 'warmer' will not be effectively decidable, and is thus vague in (at least) an epistemological sense. This vagueness would then create difficulties further down the road in creating useful logical or empirical generalisations utilising the notion of 'warmer'.

The explicatum designed to replace 'warmer', however, will ideally admit of no such cases, and will be both a determinate and effectively decidable notion. In this case, we replace the explicandum, 'warmer', with the explicatum, HIGHER TEMPERATURE, and define HIGHER TEMPERATURE in an operational way, such that we follow a particular physical procedure in order to determine whether α causes the mercury in

a thermometer to expand more than β. Being able to apply this procedure, at least in principle, i.e. given enough time and resources, for any two items appropriate to HIGHER TEMPERATURE, entails that the extension of HIGHER TEMPERATURE will be determinate and decidable in a way that 'warmer' is not. The use of the explicatum TEMPERATURE renders this area of discourse exact.

Now, as we've seen, one reason that we desire this kind of determinacy because it allows for the construction of fruitful, universal generalisations, whether logical or empirical. This is not, however, the only reason why we desire this sort of determinacy: there is an important connection between the requirement of exactness in an explicative project and *objectivity*. Roughly speaking, (satisfactorily) explicating an area of discourse (such that the explicatum is sufficiently similar, fruitful and simple, but, most importantly, exact) allows us to make that area of discourse *more objective*. I shall now explain what I mean by this.

Since the notion of 'objectivity' is itself a matter of considerable attention and dispute in its own right,[80] and that I have in mind a particular conception of 'objectivity', I shall now elaborate exactly what I mean by this. The relevant notion of 'objectivity' here I refer to as 'disagreement-entailing-error objectivity'.[81] In an area of

[80] The *locus classicus* here is Crispin Wright's *Truth and Objectivity*, (Wright 1994) where Wright outlines at least four distinct notions of 'objectivity' that supposedly represent improvements in objectivity from the baseline provided by an area of discourse fitting his 'minimalist' conception of truth: verification-transcendence, cognitive command, width of cosmological role, and judgement-independence. (c.f. Miller 2007, pp.332-5).

The most relevant notion of objectivity in Wright's work is that of 'cognitive command'. According to Wright, an area of discourse exhibits 'cognitive command' if and only if "It is a priori that differences of opinion formulated within a discourse, unless excusable as a result of vagueness in a disputed statement, or in the standards of acceptability, or variation in personal evidence thresholds, so to speak, will involve something which may properly be regarded as a cognitive shortcoming." (Wright 1994, p.144)

There is an underlying problem here, however, concerning the placement of 'cognitive command' within Wright's overall alethic minimalist project. Given Wright's commitment to regarding areas of discourse that exhibit *discipline* (such that there are standards within the discourse such that some sentences count as properly applied and others not) and *syntax* (such that sentences can be embedded, negated, etc.) as minimally truth-apt, it seems to follow that it will be *a priori* that a minimally truth-apt sentence is either true or false (at the very least, under a 'classical' non-intuitionist, non-relativised notion of 'truth'). This entails that any disagreement on such a truth-apt matter will necessarily involve one of the relevant parties to a disagreement believing something *false*. This can itself be plausibly regarded as a 'cognitive shortcoming'. (Kölbel 2002, pp.24-8; Shapiro & Taschek 1996; Hale 1998, p.299). The result would be that any minimally truth-apt discourse would *automatically* exhibit cognitive command, and thus cognitive command would thereby fail to pick out any distinctive notion of objectivity.

Evidently, much more could be said here, but I take it that the notion of 'cognitive command' is less helpful than the notion of 'disagreement-entailing-error objectivity' I use below.

[81] This notion of objectivity I have taken from Max Kölbel's *Truth Without Objectivity*:

discourse that is disagreement-entailing-error objective, it will be *a priori* that (at least) one of the parties to a disagreement will have made a mistake or error of some sort. For instance, if we imagine disagreements (where one party asserts a proposition, and the other asserts its negation) as to whether some triangle is right-angled, it is *a priori* that (at least) one of the parties to the disagreement will have asserted something incorrectly, which counts as an error. There is an operative standard of correctness for the assertion "This triangle is right-angled", and going against this standard represents an error.[82] We know in advance that, should a dispute arise, one of the parties to the disagreement will have made a mistake of some sort, i.e. that their belief or judgement will be incorrect.

In areas of discourse which are not disagreement-entailing-error objective, however, we get cases where, if a disagreement arose, it would be a case of *faultless disagreement*.[83]

"(CO) For all thinkers *A* and *B*: it is a priori that if *A* believes that *p* and *B* believes that not-*p* then either A has made a mistake or *B* has made a mistake." (Kölbel 2002, p.31)

The notion is similar to that of Wright's 'cognitive command' constraint. (See ft.80) However, Wright's 'cognitive command' requires that some sort of *cognitive shortcoming* is involved in the case of disagreement over some issue exhibiting cognitive command, which in turns creates difficulties for Wright's overall minimalist project since it is unclear whether believing something (minimally) false counts as a 'cognitive shortcoming'. Kölbel's definition, however, is broader in that it allows for any sort of 'mistake', whether or not it is a 'cognitive shortcoming'. As such, it is easier to apply, and seems to accord well with our intuitive understanding of 'objectivity'.

[82] At this point, we need to note that one could understand 'correctness' here as referring to *truth-claims* or *warranted assertibility-claims*. Where we understand 'correctness' as truth, in the case of disagreements about (truth-apt) propositions, it is obvious that (on a classical understanding of truth, at least) any disagreement will involve an error, since one party asserts the negation of a true proposition.

On a warranted assertibility view, however, this is less obvious. The warranted assertibility view of correctness, then, only commits us to claiming that if P is correct (i.e. it is warranted to assert P), then it is incorrect to assert ¬P. It does not commit us to claiming that either P or ¬P is correct (i.e. that it is either warranted to assert P or warranted to assert ¬P).

On the occasions where it is clearly warranted to assert some proposition P, then, it is not warranted to assert ¬P, and thus one party will be in error in asserting ¬P without warrant. Taking the example of "This triangle is right-angled", where the triangle is right in front of us, clearly drawn, etc., and it is demonstrably right-angled, it is correct (warrantedly assertible) that P, and thus it is incorrect that ¬P.

However, there are also occasions whether neither P nor ¬P is warrantedly assertible. Suppose, again taking the example of "This triangle is right-angled", that the triangle referred to is a triangle drawn by a person 300 years' ago where all we know about it is that it still exists somewhere and is definitely a triangle. On this occasion, it is neither warranted to assert P or to assert ¬P, and thus *both* parties to the disagreement will have made some sort of error.

Either sort of occasion still counts as disagreement-entailing-error objective, since it remains *a priori* that at least one party to a disagreement has made an error of some sort. One of the occasions is unusual, however, in that *both* parties to a disagreement are in error.

[83] The notion of 'faultless disagreement' comes from Kölbel's paper of the same name, (Kölbel 2003, pp.53-4) and plays a key role in his *Truth Without Objectivity* (Kölbel 2002). A faultless disagreement exists where two thinkers A and B form diametrically opposed beliefs (or judgements) concerning some proposition, and yet neither A nor B is at fault or has committed some sort of error. If one has a faultless disagreement in some area of discourse, it will follow that that area of discourse fails to be

Imagine that we have a disagreement between two competent language-users A and B, about whether it is correct to apply some term that is semantically vague. Let's say that A and B are disagreeing about whether the weather is warmer today than yesterday, and that there is a very marginal difference in the temperature on the two days. Because it is (arguably) unclear at a semantic level, however, *how much* of a difference in temperature (or even whether temperature is the sole factor in 'warmth' levels), is needed for one day to constitute being 'warmer' than the last, we might regard it as fine to assert that "Today is warmer than yesterday", but also as fine to assert that "It is not the case that today is warmer than yesterday". A and B disagree over whether it is correct to apply the term 'warmer' in this case, but neither of them has made anything worth calling an 'error' or a 'mistake'. As a result, we can say that the issue as to whether α is warmer than β is not disagreement-entailing-error objective.

The problem here is caused by the indeterminacy of the notion of 'warmer', which came about as the result of some semantic vagueness where there exist borderline cases where there is no apparent standard of correctness as to whether the notion applies. Without such a standard of correctness, one cannot convict any party to the disagreement of some 'error', and thus there exist areas within discourse using the notion 'warmer' that fail to be disagreement-entailing-error objective.

Here we see a core motivation for explication. If we introduce exact explicandum HIGHER TEMPERATURE, we can avoid such cases of faultless disagreement and transparently render such areas disagreement-entailing-error objective. We introduce some explicandum such that what counts as the (correct) extension of that explicandum is defined in an operational manner. That is, α is a HIGHER TEMPERATURE than β if and only if α, upon introduction of a mercury thermometer, causes the mercury to expand more than β. We introduce a determinate explicatum through the supplying of an effective procedure and stipulating that the extension of that notion *just is* the outcome of following such a procedure. Providing the procedure is an effective one, there is no possibility of faultless disagreement

disagreement-entailing-error objective, since it is not *a priori* that a disagreement means that some error has been made by (at least) one of the parties to the disagreement.

concerning the application of the (exact) explicatum, since the standard of correctness just is whatever outcome following the procedure gives you.

The correct application of an explicatum, then, is not merely a disagreement-entailing-error objective matter, but *transparently* so. That is, because the explicatum is defined via an effective procedure, we *know* in advance that cases cannot arise in which there is no standard of correctness for the application of the explicatum. Not only can we satisfy ourselves that the extension of the explicatum is determinate, i.e. without gaps, we have a method through which we can actually ascertain, given enough time and resources, what the correct application of the explicatum is in any given case. Thus, faced with any disagreement about the application of the explicatum, we not only know *a priori* that one of the parties to the disagreement is in error, but we can (given enough time and resources) locate *which* party to the disagreement is in error.

If we then accept HIGHER TEMPERATURE as explicating the notion of 'warmer' (as presumably we do), we then have a method that removes areas of faultless disagreement concerning claims about one thing being warmer than another. We use the explicatum HIGHER TEMPERATURE as a way to settle the question as to whether α is warmer than β. The question as to whether any particular locality at a time is warmer than another, then, becomes a disagreement-entailing-error objective matter. A satisfactory explication takes us from a situation in which we have areas of faultless disagreement as to how to apply some term correctly to a situation where we have an agreed method of determining whether or not the term, or rather its explication, correctly applies. In this situation, because we accept this explicatum as explicating the explicandum, any disagreement will be such that it will be *a priori* that one party has made some sort of mistake, and, furthermore, we can (in principle) identify *which* party has made an error.

It is worth stressing here that whether α is warmer than β only becomes a transparently disagreement-entailing-error objective matter here where we accept HIGHER TEMPERATURE as explicating 'warmer' *to the exclusion of other possible explicata.* There may, in fact, be many potential procedures that could satisfactorily explicate 'warmer', such as HIGHER VOLTAGE FROM A THERMOCOUPLE or HIGHER THERMAL

RADIATION. It is likely that all three explicata, whilst distinct procedures that would return different values for different occasions, would all be *sufficiently similar* to the explicandum 'warmer', whilst also being highly exact. In order to render the extension of 'warmer' (in this area) transparently disagreement-entailing-error objective, however, one such explicatum would need to be chosen to the exclusion of others.

This example concerns using a physical procedure of empirical observation such that the extension of the explicatum *just is* the outcome of carrying out this procedure of observation. Another example, one where the explicatum is given using a set of logical procedures, may be useful at this point. Take the everyday notion of a 'prime number', a number that is divisible by only itself and 1. Informally, we typically utilise the following method to see if a given natural number is prime. We look at the number and rule out obvious ways in which it could fail to be prime: if the number ends in an even digit, we immediately conclude it's not prime, since any number ending in an even digit will be divisible by 2. Similarly for any number ending in '5': it will be divisible by 5. If the number ends in '3', we immediately try to divide it by 3 and see if the result is an integer. We then try other plausible candidates. After a while, we feel that we've checked enough possibilities and can't see any other possibilities, so we conclude that it is prime.

At least as informally understood (i.e. without further clarification), the notion of a 'prime number' is arguably not a determinate notion. There is an issue, for instance, about whether negative integers can be prime: is -7 a prime number? It is also the case that our informal understanding of 'prime number' does not supply us with an *effective* procedure for determining whether some number is prime. That this informal method is not effective becomes clearer once we start to deal with very large numbers. Without some step-by-step guide here, it is unlikely that the informal method adumbrated will ever generate a sufficient degree of confidence or reliability that some number is prime or not. No matter how much time we invest in the matter, without some further systematisation of our methodology, we will be unable to decide whether certain very large numbers are prime or not.

By contrast, a logical, effective procedure is not subject to the same limitations. Let us imagine a procedure PRIME, operating across the domain of natural numbers $\mathbb{N}$ (thus ensuring that negative integers cannot be PRIME) that takes the following form (or some equivalent):

1. Let $i = 2$
2. Let $k = n / i$
3. If $k \in \mathbb{N}$, STOP, $n \notin$ PRIME
4. If $i > n/2$, STOP, $n \in$ PRIME
5. $i = i + 1$
6. GOTO 2

At no point in this procedure is there any unclarity about how to proceed. One could easily program a computer to carry out this procedure, such that whether any natural number is PRIME will simply be the outcome of the computation. For any natural number (i.e. all inputs in the domain of PRIME), the effective procedure will return, after a finite number of steps, an answer as to whether the number is PRIME. We have a method available by which we can effectively decide whether some number is PRIME or not.

Once again, the (correct) application of PRIME is understood in a stipulative fashion: whether a natural number x is PRIME or not *just is* whether the procedure returns the value 'PRIME' when taking x as an argument. Here it is built into the definition of the explicatum PRIME that a PRIME number is whatever the effective procedure returns as PRIME, similar to the way in which TEMPERATURE was defined instrumentally as the result a functioning thermometer gave. Because PRIME is an effective procedure, we know that, given enough time and resources, it will decide any appropriate input, and thus we know in advance that PRIME is a determinate notion. Faced with any natural number, the algorithm will return, after a finite number of steps, a verdict as to whether the number is indeed PRIME or not.

It seems appropriate here to accept the explicatum PRIME, and no other, as replacing the everyday notion of a 'prime number'. It is a satisfactory explicatum: it is sufficiently similar to the explicandum, it is fruitful (we can use the procedure within

other mathematical methods), it is sufficiently exact,[84] and it is simple. Because we adopt PRIME (as opposed to some logically-distinct, alternative procedure) as explicating 'prime', we thereby have a method through which we have eliminated potential sources of faultless disagreement, e.g. concerning whether -7 is prime. We accept that the way to settle the question is by asking whether -7 is PRIME, and we conclude that -7 is not an appropriate input and so could not be PRIME.[85] If there was a disagreement about whether -7 was prime, we could then point to the accepted explicatum to establish that some sort of error had been made, securing the disagreement-entailing-error objectivity of this area of discourse. We can then go further and identify which party to the disagreement was in error.

The method of explication, then, is a tool, not just for the purposes of constructing universal generalisations, but for rendering areas of discourse transparently determinate and disagreement-entailing-error objective. If the explicatum fits our ideal of exactness, we know that there will be no areas of indeterminacy in its extension, since its extension is operationally defined as just whatever is the outcome of following through some procedure. If there are no such areas of indeterminacy in the extension of the explicatum, it follows that any disagreement about whether the explicatum will involve some error, i.e. the failure of one of the parties' judgements to accord with the extension of the explicatum. It becomes a disagreement-entailing-error objective matter. An ideally exact explicatum, accepted to the exclusion of alternatives, transparently secures the disagreement-entailing-error objectivity of some area of discourse.

[84] Actually, the exactness of PRIME here seems to achieve our ideal notion of exactness, provided the various notions utilised in the algorithm itself (e.g. '∈', '>', 'STOP' etc.) are not themselves indeterminate or not effectively decidable in some way.

[85] We might, of course, have defined an explicatum PRIME-I such that it accepts *integers* as opposed to *natural numbers*, which would of course have altered the pattern of correct application for 'prime number'. Here, I just wish to reinforce the point that, in order to secure disagreement-entailing-error objectivity, we need to accept PRIME as explicating 'prime number' to the exclusion of alternatives such as PRIME-I.

2.5 – The Explicative Project

It is worth taking stock a little. We began by examining Stich's proposal for understanding neo-Goodmanian accounts of reflective equilibrium in the area of justified inference as attempts to explicate what it is to hold a justified set of inferential principles. We stressed that, as an explicative project, as opposed to a traditional analytic project, we ought not expect that an explicatum preserves the meaning of the explicandum, and thus that one can overplay the importance of *a priori* counterexamples in which the explicandum and explicatum part company. We then saw that this was just to rehearse a point Carnap himself makes about explicata: that they only need to be sufficiently *similar* to the explicandum in order to be satisfactory. Added to this criterion of satisfactoriness were fruitfulness, simplicity, and exactness. We've then spent a good while examining the importance of Carnap's ideal of exactness, connecting it with a specific notion of objectivity, disagreement-entailing-error objectivity. By putting forward an effective procedure for deciding the extension of one's explicatum, one transparently reveals the extension of the explicatum to be determinate, and we can then be confident that the application of the explicatum is a disagreement-entailing-error objective matter.

In all of this, we need to bear in mind that determinacy and effective decidability represents the *ideal* for exactness in one's explicatum. In practice, an explicatum might nonetheless have residual areas of unclarity such that we are unsure how to apply the procedure in certain circumstances. As the explicative project develops, however, one would hope to remove these residual 'clots' of (epistemological) vagueness. One would hope to have an explicatum that allows us to effectively decide whether any given appropriate item falls under the explicatum or not, thus rendering it a fully determinate notion. The more we approximate to this ideal, the more we can secure objectivity in our judgements.

The obvious question now is how this connects to the standard interpretation of reflective equilibrium in which reflective equilibrium *qua* coherence theory of justification in ethics takes centre stage. The proposal, adapted from Stich, is that we understand reflective equilibrium as an attempt to explicate what it is for an agent to be justified in holding some moral conception. In order to do so, we need to supply

some explicatum REFLECTIVE EQUILIBRIUM that will be sufficiently similar to our judgements about whether an agent is justified in holding their moral conception, is sufficiently exact (such that we can effectively decide whether an agent holds their moral conception in reflective equilibrium), is fruitful (such that it connects with other issues, e.g. moral knowledge), and is simple enough. I'll now spell that out in more detail, and in particular examining whether this proposal is a charitable and appropriate way to characterise what we've already seen concerning the standard interpretation of reflective equilibrium.

At the end of the last chapter, I argued that it was premature to demand, as the usefulness objection would have it, that the standard reflective equilibrium theorist produce a conception of 'reflective equilibrium' that is of immediate use in terms of separating moral conceptions into those that may be justifiably held, and those that may not. What is needed, first of all, is to garner a clear conception of our ideal, and to understand what is involved in a state of (wide) reflective equilibrium obtaining, and to understand how it relates to an agent's being justified in holding their moral conception. When it comes to the actual business of moral theory acceptance, and examining in practice which agents are justified in holding their moral conception, what is needed are working methods that allow us to approximate to this ideal. It is legitimate to request that such methods are forthcoming, but it is an inappropriate criticism if it is directed against accounts that are attempting to clarify the relevant philosophical ideal.

However, it is legitimate to request of our philosophical ideal that it is something where it is in principle possible for us to ascertain whether a state of (wide) reflective equilibrium obtains (or not). This is due to the fact that a state of (wide) reflective equilibrium (on the standard view) obtains or not as the result of some function operating over the various moral and non-moral beliefs an agent happens to hold. If the various beliefs 'fit' together sufficiently, then a state of (wide) reflective equilibrium obtains. What this then implies is that our philosophical ideal conception of 'wide reflective equilibrium' is something that may be, to all intents and purposes, beyond our epistemic grasp *in practice* (since it is, in all likelihood, far too cognitively or computationally taxing), but that nonetheless it must be *in principle* possible to ascertain whether a state of wide reflective equilibrium obtains (or not). With the

appropriate reflective equilibrium function,[86] then, and a full specification of the belief set of the agent involved, it must be in principle possible to ascertain whether a state of (wide) reflective equilibrium obtains (or not).

We are now in a position, by invoking notions of 'explication', 'correctness', and 'disagreement-entailing-error objectivity', to clarify what it means to say that "it must be in principle possible to ascertain whether a state of (wide) reflective equilibrium obtains". First of all, regarding the notion of 'obtaining', we might take this to mean many things. We might take it to be a *factual* claim, i.e. as a true/false proposition. Furthermore, we might take it to involve some metaphysical claim, whereby such a fact is, as it were, part of the fabric of reality independently of any human investigation, a position often referred to as 'Platonism'.[87] Clearly, we do not want to attribute to a standard reflective equilibrium theorist any tendentious metaphysical or semantic claims that they might regard as optional commitments. For this reason, I propose understanding it as a claim about the disagreement-entailing-error objectivity of ascriptions of 'being in (wide) reflective equilibrium', which is neutral on questions of truth-aptness[88] and the metaphysical status of the relevant proposition. On this view, a state of wide reflective equilibrium 'obtains' (or not) if and only if, faced with a disagreement where A asserts 'Agent x is in wide reflective equilibrium' and B asserts 'It is not the case that agent x is in wide reflective equilibrium', at least one of the parties to this disagreement is in error. To put it another way, if a state of wide reflective equilibrium is something that 'obtains' (or not), attributions of 'reflective equilibrium' have to be such that *faultless disagreement* is impossible.

A state of wide reflective equilibrium 'obtaining' (or not) then, is a matter of ascriptions of 'being in wide reflective equilibrium' being disagreement-entailing-

[86] Clearly, we do not currently have such an appropriate reflective equilibrium function. What we have is some indications of what an appropriate function might look like: namely, that it would involve testing sets of propositions for whether they are held in a particular set of logical and evidential relationships characteristic of 'coherence'. Nonetheless, in thinking of reflective equilibrium as a distinctive epistemic state, the hunch of the standard reflective equilibrium theorist is that some such function is available, and that its discovery is a matter of theoretical ingenuity, time and resources. It ought to be possible, *ultimately*, to supply some function that could, with enough time and resources, decide any given case of an agent holding some moral conception in conjunction with various background beliefs.

[87] See, e.g. (Wright 1994, pp.9-10), where Crispin Wright characterises 'Platonism' (in the philosophy of mathematics) as the view we "Have the capacity to understand a specification of circumstances under which a statement is true irrespective of whether we know how it might be determined to be true." (Wright 1994, p.10)

[88] See ft.82

error objective matters. Our requirement, however, was that it is in principle possible to ascertain whether a state of wide reflective equilibrium obtains (or not), and it needs to be the case that it is in principle possible to *ascertain* that ascriptions of 'being in (wide) reflective equilibrium' are disagreement-entailing-error objective. That is, some characterisation of '(wide) reflective equilibrium' must exist such that we know in advance that, faced with any disagreement about whether an agent is in (wide) reflective equilibrium or not, at least one of the parties to the disagreement is in error. It is not good enough, in other words, for ascriptions of 'reflective equilibrium' to be disagreement-entailing-error objective, but *unknowably* so. It must be possible to make ascriptions of 'being in (wide) reflective equilibrium' *transparently* disagreement-entailing-error objective.

It is at this point that the notion of 'explication' is helpful. In order to model the requirement that it must be possible to make ascriptions of 'being in (wide) reflective equilibrium' transparently disagreement-entailing-error objective, we need to say that it ought to be possible to design an *ideally exact* explicatum. That is, it ought to be possible, given enough time, resources, and theoretical wherewithal, to design some effective procedure, REFLECTIVE EQUILIBRIUM, that would effectively decide any appropriate case as to whether an agent holds their moral conception in REFLECTIVE EQUILIBRIUM.

This ideal explicatum REFLECTIVE EQUILIBRIUM would take the following form. Whether it is correct to assert that an agent is in REFLECTIVE EQUILIBRIUM *just is* whether the explicatum REFLECTIVE EQUILIBRIUM, in the form of an effective procedure, returns a positive value. If following through the procedure, which would be some function (to be specified) that checks whether the logical and evidential relationships between the various beliefs the agent happens to hold are held in a manner characteristic of coherence returns a positive value, then it *just is* correct to assert that 'this agent is in REFLECTIVE EQUILIBRIUM'. We would thus have a standard of correctness for all appropriate ascriptions of REFLECTIVE EQUILIBRIUM, and thus we would have an explicatum that is transparently determinate. There could be no gaps in the extension of REFLECTIVE EQUILIBRIUM, since all values are decided by the effective procedure. This entails that we know in advance that any ascription of REFLECTIVE EQUILIBRIUM is a disagreement-entailing-error objective manner.

Faced with any disagreement concerning whether some agent is in REFLECTIVE EQUILIBRIUM or not, we would have a method that decides (given enough time and resources) what counts as the correct answer. Whether an agent holds their moral conception in REFLECTIVE EQUILIBRIUM would be *transparently* disagreement-entailing-error objective.

The commitment, then, is the following: some such ideally exact explicatum, REFLECTIVE EQUILIBRIUM, taking the form of an effective procedure must be, in principle, within our epistemic grasp. We can now add in our other principal constraint: that this ideal explicatum needs to be *sufficiently* similar to our everyday notion of 'being justified'. That is, the results of applying the REFLECTIVE EQUILIBRIUM procedure need to accord sufficiently well with our intuitive judgements as to whether the agent is justified in holding their moral conception. This means that the REFLECTIVE EQUILIBRIUM explicatum will be sufficiently invulnerable to plausible counter-examples.

Provided this constraint is fulfilled,[89] we would then have an explicatum that provides a standard of correctness for whether the agent is justified in holding some moral conception. Just as PRIME (and not some alternative procedure) gives us a standard of correctness for whether '*x* is a prime number', this ideal REFLECTIVE EQUILIBRIUM explicatum (and not some alternative procedure) would give us a standard of correctness for whether '*A* is justified in their moral conception'. If we wanted to know whether some agent was justified in holding some moral conception, we know in advance what counts as the standard of correctness, i.e. whether the procedure REFLECTIVE EQUILIBRIUM (to the exclusion of alternative procedures) would return a positive value. This would render the area of discourse ('being justified', or matters of moral theory acceptance more generally) transparently disagreement-entailing-error objective.

None of this implies that such an explicatum could be made available *in practice*. We may simply lack the theoretical imagination to come up with a procedure that truly captures what is involved in various moral and non-moral beliefs 'fitting together'. The idea is to capture what is involved in the commitment that whether a state of

[89] I assume that the REFLECTIVE EQUILIBRIUM will be able to meet the fruitfulness and simplicity requirement.

(wide) reflective equilibrium obtains or not must in principle be ascertainable. We have in mind an *ideal* form of explication where the explicatum is ideally exact (and sufficiently similar to 'being justified'), such that any appropriate case is effectively decidable. That ideal may *in practice* turn out to be unattainable, but it is a commitment on the standard view that it must be attainable under ideal circumstances (i.e. given sufficient time, resources, and theoretical wherewithal).

It is helpful, then, to regard the standard understanding of reflective equilibrium as an *explicative project* where we have in mind a certain in principle attainable ideal to which any proposed explicata, due to limitations of time, resources, and theoretical imagination, can only approximate. In terms of creating *working* explicata, it is satisfactory to produce procedures that are *sufficiently* exact, which may well mean that they do not take the form of *effective* procedures, such that they can decide any appropriate case. The commitment is, however, that the ideal is in principle attainable. It must be in principle possible to design some explicatum, REFLECTIVE EQUILIBRIUM, that could effectively decide any given case in a manner that is sufficiently similar to our intuitive judgements about 'being justified'. Ideally, reflective equilibrium, qua coherence theory of justification in ethics, would render all questions of whether an agent 'is justified' in holding their moral conception, transparently disagreement-entailing-error objective.

This completes my characterisation of the *target* of my subsequent argument. On the standard interpretation of reflective equilibrium, 'reflective equilibrium' is understood primarily *qua* coherence theory of justification in ethics, where it picks out an epistemic state that obtains (or not), that is intended to tell us something about whether an agent is justified in their moral views. The way to understand this is to see the standard approach as committed to an explicative project in which an ideal explicatum is in principle attainable. Over the next few chapters, however, I shall argue that, on plausible assumptions about the nature of moral discourse, this is not the case, and thus that the standard approach is committed to an explicative project that is unlikely to be successful.

Chapter 3 – Kripke's Sceptical Paradox

3.1 – Overview

We have, then, a clear conception of the commitments of the explicative project involved in a standard understanding of reflective equilibrium. Some ideally exact explicatum, sufficiently similar to 'being justified', must be available such that it could be, in principle, used as a procedure for effectively deciding the correctness of any particular judgement as to whether an agent holds their moral conception in REFLECTIVE EQUILIBRIUM, which, as we've granted from the outset, would then allow us to answer as to whether they are justified in holding their moral conception. If such an ideal is in principle available, we may have some confidence in the explicative project of rendering attributions of 'being justified' (more) disagreement-entailing-error objective.

Thus far, I take it that a standard reflective equilibrium theorist is entitled to cautious optimism with regards to this explicative project. Their hunch is that a reflective equilibrium approach *qua* coherence theory can illuminate the question as to whether an agent is justified in holding their moral conception. Some ideally exact explicatum, sufficiently similar to 'being justified', is attainable (given enough time, resources, and theoretical imagination). From here, it looks to be a viable project to create working explicata that approximate to this ideal.

However, I shall argue, over the course of chapters 3-5, that it is not possible, given a plausible reading of a certain sort of moral disagreement, to supply such an ideal explicatum, even in principle. In contrast to, say, areas of mathematics, where such an ideal explicatum is in principle (and, often, in practice) attainable, moral cases (at least as we currently find them) create insuperable difficulties. I shall argue that it is not possible, to design a REFLECTIVE EQUILIBRIUM explicatum that could be accepted to the exclusion of alternatives in the form of an effective procedure that would decide whether the various moral and non-moral beliefs held by an agent 'fit' together, as required by reflective equilibrium *qua* coherence theory of justification in ethics. As a result, the standard approach to reflective equilibrium has a commitment that it cannot hope, given the nature of the moral sphere, to satisfy.

The purpose of this chapter, however, is to lay some of the groundwork for the argument. The argument I use is adapted from Wittgenstein's rule-following considerations, but more particularly is adapted from a particular strand of argument utilised by Kripke's sceptic in *Wittgenstein on Rules and Private Language* (Kripke 1982). Evidently, these concerns in philosophy have attracted an inordinate amount of attention, so I can only mount a partial survey of the terrain.[90] The intention here is to give an indication of the lay of the land, so that when I isolate the considerations that are relevant to the argument (as I will do in chapter 4), it will be easy to locate these considerations within the overall logical space surrounding Kripke's sceptical paradox/the rule-following considerations.

I proceed by outlining the 'sceptical paradox' that Kripke's sceptic argues confronts any would-be meaning factualist. The meaning factualist is seemingly committed to the view that they will be able, under ideal epistemological conditions, to uniquely identify some meaning-fact that rules out various ersatz alternative interpretations that the sceptic dreams up (or could dream up). However, they are unable to do so, which creates a *reductio* for the meaning factualist, and then the sceptic's suggestion is that the faulty premise is that there are meaning-facts. I consider a selective range of possible 'straight solutions' to this sceptical challenge, concentrating in particular on dispositionalist attempts, before examining Kripke's Wittgenstein's supposed 'sceptical solution', at least as it is standardly understood.

[90] Even Martin Kusch, in his *A Sceptical Guide to Meaning and Rules,* a book-length treatments of Kripke's celebrated work, notes that his review of relevant material is necessarily selective. (Kusch 2006, p.94) In what follows, I shall maintain that certain strand of Kripke's sceptic's argument is successful, and survey the secondary literature insofar as it pertains to this strand of argument. Even with this limitation in scope, however, my use of secondary review is selective.

3.2 – Meaning Factualism

In this section, I shall provide a brief overview of the sceptical paradox as formulated by Kripke's sceptic. The standard reading[91] of how Kripke understands Wittgenstein[92] is to see him as putting forward a sceptical paradox intended to undercut a seemingly innocuous semantic thesis, namely, that when a language-user, S, asserts a sentence, 'p', there is a corresponding metalinguistic fact of the matter[93] expressed in the form 'S means that *p*'.[94] To use the exemplary case, if Smith utters the sentence '68 + 57 = 125', there is a fact of the matter (Smith means *addition*) about what Smith means by '+'. The concept *addition* supplies the relevant metalinguistic fact about what Smith means by '+'. Similarly, when Jones utters the sentence 'This is a dog', there is a fact of the matter as to what Jones means, and that by 'dog' he means *dog*. Call this the 'meaning factualist thesis':

---

[91] This reading may be found in, for example, Wright's work (Wright 1984, pp.768ff.; Wright 1987, p.27; Wright 1989), and Paul Boghossian's highly influential paper 'The Rule-Following Considerations (Boghossian 1989). The standard reading is contrasted with the 'revisionary' reading (Miller 2007, p.191), originating from George M. Wilson (Wilson 1994), but also shared by Alex Byrne (Byrne 1996), and Kusch (Kusch 2006, ch. 5). Essentially, the difference concerns the relevant stances of the voices in Kripke's dialectic. The standard reading has it that Kripke endorses the radical sceptical view, attributed to Wittgenstein, that there are no facts of the matter (understood in a truth-conditional manner) concerning the claim that 'S means that *p*'. Kripke's (and Kripke's Wittgenstein's) attention then turns to saving such claims about meaning from the thought that our practice of ascribing meaning is radically in error, through assigning *assertibility conditions* to claims about meaning: the so-called 'sceptical solution'.

The revisionary reading has it that, whilst Kripke's *sceptic* holds this radical sceptical view (thus abandoning any form of factualism regarding metalinguistic claims), Kripke's *Wittgenstein* is rather a therapeutic voice, viewing such non-factualism as intolerable, and regarding the sceptical paradox as a *reductio ad absurdum* of the *particular form* of factualism on offer. Wilson suggests that Kripke's Wittgenstein's position is that genuine facts about meaning are secured, not via truth-conditions, but via assertibility conditions. (Wilson 1994, p.254) According to the revisionary view, Kripke's Wittgenstein should be regarded as putting forward a 'factualist' understanding of meaning, but as disagreeing with Kripke's sceptic as to how one should understand such 'factualism'. On this reading, then, the 'sceptical solution' offered should not be read as an attempt to get by without meaning-facts, but rather as a competing account of what such 'meaning-facts' consist in.

[92] It is, one should note, highly controversial to what extent Kripke's exposition of the rule-following considerations remain faithful to Wittgenstein's original intentions. McDowell argues (McDowell 1993) that Wittgenstein's aim, far from presenting a 'sceptical paradox' to the would-be meaning factualist, is to avoid a certain presupposition upon which the 'sceptical paradox' relies. McDowell insists that Wittgenstein's aim is to prevent the assimilation of understanding to interpretation (McDowell 1993, p.260; c.f. McDowell 1984, p.270). On McDowell's view, it would therefore be a mistake to regard Wittgenstein as implicitly accepting the 'sceptical paradox' in offering a 'sceptical solution', since Wittgenstein should be read as attempting to demonstrate why the seeming 'paradox' arises from a misunderstanding in the first place.

[93] The 'metalinguistic fact' picks out purported facts expressed in the meta-language, as opposed to the object-language. If I assert 'The cat is on the mat', the object-linguistic fact is that the cat is on the mat; the metalinguistic fact is that, by 'The cat is on the mat', I mean that *the cat is on the mat*.

[94] Here I follow Miller's exposition (Miller 2007, pp.166ff.).

Meaning Thesis	Factualist	There are meaning-facts, metalinguistic facts of the matter, expressible in the form 'S means that *p*'.

Adopting the meaning-factualist thesis, in conjunction with what is generally regarded as a platitude concerning the *normativity* of meaning,[95] leads us to a further claim. These metalinguistic facts about what S means by 'p' provide a standard of correctness against which particular uses of 'p' can be compared in order to identify whether S has used the term 'p' correctly. S has, if you like, a 'contractual'[96] commitment to continue to use 'p' in accordance with its meaning. We can spell out this commitment in the following way: granting that S currently means that *p*, (there is a metalinguistic fact of the matter about what S currently means by 'p'), any future usage of 'p' will only be *correct* if it conforms to *p*. Having established that Smith means *addition* by '+', should Smith in the future consistently perform '+' calculations that run contrary to *addition*, Smith has made a linguistic error in his use of '+'.[97] The idea, then, is that the S's sincere use of 'p' is answerable to a standard of correctness

[95] I ought to make it explicit here that I'm only interested in the notion of the 'normativity' of meaning in terms of the platitude that there exists a distinction between correct and incorrect usage, and that there is a possibility applying a term incorrectly. As Paul Boghossian puts it:

"The normativity of meaning turns out to be, in other words, simply a new name for the familiar fact that, regardless of whether one thinks of meaning in truth-theoretic or assertion-theoretic terms, meaningful expressions possess conditions of correct use." (Boghossian 1989, p.148)

[96] Here I follow McDowell (McDowell 1993, p.221) in endorsing this intuitive notion of correctness-normativity. Wright expresses this intuitive, 'contractual' understanding of normativity thus: "It is usual to make free use of the idea of an application of a concept according with, or failing to accord with, its content. It is in accordance with the meaning of 'red', as we understand it, that it should be applied to red things rather than blue ones. We think of giving the meaning of an expression in contractual terms. Once the meaning has been fixed in a certain way, we are all obliged to make a certain kind of use of the expression; only that kind of use conforms with the sense of the expression that was fixed. We are, so to speak, constrained by our understanding. If we are to use the expression in conformity with the way we understand it, or the way the community at large has generally used it in the past, we ***have*** to use it in certain sorts of ways." (Wright 1994, pp.19-20)

[97] It is important to be clear here that the relevant error is a linguistic error. If Smith, for instance, asserts that '94 + 23 = 107', the relevant error here is an arithmetical error: he has *miscalculated*. Thus, a one-off miscalculation is hardly likely to be considered problematic in terms of meaning. A linguistic error, by contrast, is where one uses a term in a way that is not faithful to the correct meaning. If someone consistently miscalculated, or insisted that '94 + 23 = 5', for instance, it would be more plausible to think that they were using '+' incorrectly as opposed to miscalculating.

I note this because Åsa Wikforss (Wikforss 2001) argues that it is inessential to meaning factualism that there exists a possibility of linguistic error. Standard arguments to this effect, she maintains, are typically based on a confusion between making *false* judgements (i.e. miscalculating) and making linguistic errors.

It seems, however, that the possibility of linguistic errors, of using a concept contrary to its meaning, is a simple consequence of what everyone regards as a platitude – that meaning-facts entail conditions for correct use. A meaning-fact divides potential applications of the term into two groups: correct and incorrect usage. If one uses a term in a way that belongs to the second set, one has made a linguistic error. One may, or may not, have made a false statement *in addition to this linguistic error*, depending on whether someone can charitably understand what you were *intending* to say. Nonetheless, one has still used a term contrary to its meaning, and this is a linguistic error.

provided by the meaning-fact, and it is on this basis we can say whether S makes some linguistic error. Facts of the matter about what S means by 'p', then, supply a standard against which S's future usage of 'p' may be assessed for correctness. Intuitively, we seem to be committed to the idea that there are straightforward facts of the matter as to what S means when they use the term 'p', namely *p,* and that this fact of the matter supplies a correctness-condition against which S's particular uses of the term 'p' may be assessed. Call this the 'meaning-facts supply correctness-conditions thesis':

Meaning-facts Supply Correctness-Conditions Thesis	If there are meaning-facts, they supply correctness-conditions. S uses 'p' correctly just in case S's usage conforms to *p*. S makes some linguistic mistake just in case S's usage fails to conform to that standard.

Adopting the meaning factualist thesis, then, would seem to require of us to adopt this second commitment. Kripke's sceptic then presents us with another seemingly innocuous commitment once one adopts the meaning factualist thesis: if there is a fact of the matter about what S means when they use 'p', it should be possible for S, given unlimited epistemic access to all possible mental and behavioural facts, to identify such meaning facts. This is the sceptic's challenge to the would-be meaning factualist: cite some mental or behavioural fact about yourself (that would be available if there were no restrictions on your epistemic powers) that allows you to identify the relevant meaning-fact.

What does Kripke's sceptic have in mind here? First of all, we need to note the sort of facts that one might cite in response to the sceptic that would satisfy them. Kripke takes it that, in contrast with the 'argument from below' in Quine's *Word and Object*, there are no significant restrictions on the kind of 'fact' one might cite in responding to the sceptic's challenge.[98] However, as many commentators have noted

[98] Quine's challenge is for us to cite some piece of *behavioural* evidence such that one can establish that 'gavagai' refers to a *rabbit* as opposed to an *undetached-rabbit-part, time-slice-of-rabbit* etc. For Quine, such meaning-facts ultimately had to be couched in the language of so-called 'occasion sentences'. (Quine 1976, pp.35ff.)

Kripke notes (Kripke 1982 pp.10-1) the similarity of his sceptic's argument and Quine's 'argument from below', but insists that there are no such behaviourist restrictions in his sceptical challenge:

(Boghossian 1989, pp.178ff.; Goldfarb 1985, p.95; McGinn 1984, p.81; Wright 1984, pp.775ff.), the considerations cited do need to be *non-semantic* and *non-intentional* in nature. It is no response to the sceptic's challenge that the relevant fact that identifies that 'S means that *p*' is that S means that *p*, or that S intends *p,* since such a response, at least in such a simple form,[99] evidently invites a regress (cite some possibly available fact of the matter that would allow you to identify that S intends *p*). Kripke's sceptic is evidently requiring that meaning-facts are reducible to non-semantic, non-intentional items.

Furthermore, by 'identify', Kripke's sceptic has something rather specific in mind, namely that we should be able to *uniquely* pick out some meaning-fact to the exclusion of possible alternatives. It would not suffice to say that, when S uses the term 'p', the relevant meaning-fact is that, by 'p', S means that *p* if we cannot rule out the possibility that S actually means *q* or *r* by 'p'. Granting unlimited epistemic access, we ought to be able to cite considerations that demonstrate that all alternative interpretations of what S means by 'p' are ruled out, and that only *p* remains as the unique interpretation of what S means by 'p'.

Why ought we accept this demand? At this point, Kripke's sceptic claims that failure to *uniquely* identify the relevant meaning-fact would result in a certain justificatory lacuna. Kripke's sceptic's claim is that that the existence of meaning-facts, combined with unlimited epistemic access, should be such that one would thereby be in a position to *justify* one's meaning *addition* by '+'. (Kripke 1982, p.11) The claim then is

---

"Another important rule of the game is that there are no limitations, in particular, no *behaviorist* limitations, on the facts that may be cited to answer the sceptic. The evidence is not to be confined to that available to an external observer, who can observe my overt behavior but not my internal mental state." (Kripke 1982, p.14)

[99] In terms of attempts to meet Kripke's sceptic's challenge in a way that is non-reductionist, Wright's 'judgement-dependent' account (Wright 1989a) may fare better than such obviously viciously circular accounts. Roughly, Wright's view is that, only within (non-trivially specified) ideal conditions or 'C-conditions', 'S means that *p*' if and only if S *judges* that they mean that *p*. That is, in ideal conditions, a person's judgement that they mean such-and-such *constitutes* the fact that they mean such-and-such.

A consequence of this, as noted by Jim Edwards (Edwards 1992, pp.24-5), is that the distinction collapses between it *seeming* to S that they meant *p* and it *being the case* that S meant that *p*. Any case of linguistic error necessarily has to be explained via the non-obtaining of ideal conditions (C-conditions). At this point, however, it becomes unclear whether one could provide a list of C-conditions (without using a condition such as 'provided S doesn't make a mistake', which would render the account trivial) that could rule out all the possible ways in which one might make a linguistic error. (Miller 2007a)

that, by the lights of meaning factualism, [100] one would only justified in claiming that S means *p* by 'p' if one is able to uniquely identify the relevant meaning-fact *p*. In the absence of being able to demonstrate that alternative, ersatz interpretations of what S means by 'p' are ruled out, Kripke's sceptic has it that, on a meaning-factualist approach, one cannot be justified in claiming that S means *p* by 'p'.

There are a couple of possible confusions concerning this justificatory claim that we need to anticipate. The first confusion is created due to a failure to appreciate that the intended force of Kripke's sceptic's argument is *constitutive* in nature, not epistemological. (c.f. Kush 2006, pp.14-5) It is tempting to reply, as Goldfarb initially suggests (Goldfarb 1985, p.94), that whilst there is a 'bare possibility' that one might mean, say, *grue* (*grue* $=_{def}$ *green* before 2042, *blue* thereafter)[101] by 'green', it hardly seems at all plausible to think that this may in fact occur. It does not seem to undermine the claim that one is *justified* in claiming that one means *green* by 'green', say since the alternative hypothesis, that one means *grue* by 'green' is not, as it were, a relevant alternative.[102] One might maintain, for instance, that one is perfectly justified in one's claim that 'S means *green* by green' in much the same way that one may be justified in saying "That's a zebra" when confronted with zebra-like objects, and where one has ruled out all the relevant alternatives (e.g. has the animal been painted?). One would be justified or warranted in asserting that 'S means *green'* if one had ruled out, say, the possibility that S was, having just read Goodman, performing a witty subterfuge, or other relevant possibilities.

---

[100] The conditions for being justified in claiming 'S means that *p*' here are the conditions Kripke's sceptic regards as appropriate to a meaning factualist account. Ultimately, (again going along with the standard interpretation of Kripke's work) Kripke's Wittgenstein wishes to maintain that we are justified in *asserting* 'S means that *p*', but that we are not justified in asserting this in virtue of having identified a relevant metalinguistic fact. On Kripke's sceptical solution, what we see is a replacement of a truth-conditional, factualist understanding with a warranted assertibility understanding. (Kripke 1982, p.77) It is part of Kripke's sceptical solution that assertions of the form 'S means that *p*' are warranted/justified independently of any meaning-facts.

Kusch puts the matter in the following, helpful way: in the formulation of the sceptical paradox, Kripke utilises a 'metaphysical' understanding of justification, such that it is appropriate to think that, in order to be justified in maintaining that I mean *addition* by '+', I need to be able, in principle, to uniquely identify the relevant meaning-fact. However, the sceptical solution relies on a 'functional' understanding of justification, tied to conditions of assertibility. (Kusch 2006, pp.34-8)

[101] This is, note, a subtly different definition than that of Goodman's 'grue'. (Goodman 1983, p.74)

[102] Here I have in mind Fred Dretske's 'relevant alternatives' reply to radical epistemological scepticism (Dretske 1981), whereby knowledge claims ought to be understood as claims about having eliminated relevant alternative explanations for the way things appear.

As Goldfarb is aware, however, such an objection arises through a misapprehension of the epistemological idealisation at work in Kripke's sceptic's argument (c.f. Boghossian 1989, pp.150-1). It is of course the case that, given everyday informational constraints, the available evidence will often lead to scenarios where alternative ways of capturing the facts are available, i.e. where there is a certain degree of empirical underdetermination of theory by evidence in play. Here meaning-facts would be in the same position as any other theoretical item, whereby there may be, in practice, no epistemically-available considerations which help us decide which theoretical claim to make given the evidence. However, as soon as one grants unlimited epistemic access, the situation looks somewhat different: *if all the facts are in*, and we still are unable to determine whether, say, S means *green* or *grue* by 'green', it looks far more persuasive to claim that there is no fact of the matter as to what S means by 'green'. As Goldfarb puts it, "If nothing in the world settles an issue between one or another possibility, then we may conclude that there is nothing to be settled; issues of whether there are any particular grounds for doubting an ascription [of meaning] simply do not enter." (Goldfarb 1985, p.95) The claim is that, if meaning-facts exist, then, given unlimited epistemic access, it ought to be possible to justifying one particular metalinguistic claim[103] to the exclusion of all alternatives. If, however, once all the facts are in, we are still faced with alternative ersatz interpretations that are equally consistent with the facts, it would be essentially arbitrary to claim that one particular interpretation supplies the relevant fact of the matter. That is, we would not, under such conditions, be justified in claiming that one interpretation constituted the meaning-fact.

The second confusion with regards to the justificatory claim is that, as various commentators argue,[104] Kripke's sceptic might be seen to require that S *himself or herself* has to be in a position to uniquely identify the relevant meaning-fact via introspection. As such, one might easily reply that a relevant meaning-fact *is* available, but is not available via introspection. However, the claim is not that any person in particular needs to be in a position to uniquely identify the relevant meaning-fact in order to warrant the assertion 'S means that *p*', but merely that *someone* needs to be. The reason Kripke's sceptic tends to couch things in terms of

[103] In line with ft.93, a 'metalinguistic claim' is a claim about the meaning of an object word or sentence made in the meta-language. e.g. 'By 'cat', John means *cat*'.
[104] See, for instance, (Miller 2000, p.172; Zalabardo 1997, pp.288ff.)

the speaker himself or herself being able to uniquely identify the relevant meaning-fact is simply that he wishes to *include* introspection amongst the possible avenues of investigation. It does not, for instance, preclude the idea that the speaker might *also* utilise a highly-sophisticated dispositionalist theory. There is no implicit informational restriction, then, in Kripke's claim that, if meaning-facts exist, then, under ideal epistemological circumstances, S should be able to locate them. S would not only have access to all their behavioural and mental facts, but would have access to theoretical considerations, even facts about *everyone else's* mental states.

The claim then, is that a commitment of the meaning factualist position is that, under ideal epistemic circumstances, one would be able to cite considerations that uniquely identify (to the exclusion of all possible alternatives) the relevant meaning-fact. Attaching this third thesis, the 'Unlimited Epistemic Access Yields Unique Identification Thesis', allows us to give a summary representation of the Kripke's sceptic's target:

Meaning Factualist Thesis	There are meaning-facts, metalinguistic facts of the matter, expressible in the form 'S means that *p*'.
Meaning-facts Supply Correctness-Conditions Thesis	If there are meaning-facts, they supply correctness-conditions. S uses 'p' correctly just in case S's usage conforms to the standard provided by the meaning-fact. S makes some linguistic mistake just in case S's usage fails to conform to that standard.
Unlimited Epistemic Access Yields Unique Identification Thesis	If there are meaning-facts, we should be able, given unlimited epistemic access, to uniquely identify the relevant meaning-fact. A corollary of this is that one will be able, given unlimited epistemic access, to rule out all logically possible, ersatz, alternative interpretations.

3.3 – The Sceptical Paradox

At this point, Kripke's sceptic alleges that the conjunction of these three theses attributed to the meaning factualist creates a 'sceptical paradox'. To illustrate the paradox, Kripke's sceptic begins by imagining a competent language-user S (where such competence includes competence in everyday arithmetic) faced with a novel calculation[105] to perform, namely 68 + 57. Our hypothesised language-user then is imagined to assert that '68 + 57 = 125'. Intuitively, not only do we want to say that there is a straightforward fact of the matter as to what would constitute the correct answer to this sum (i.e. 125), but we want to say, following the meaning factualist thesis, that there is a fact of the matter as to what this competent language-user S *means* when they use the term '+'. The relevant meaning-fact here is that when S uses the term '+', what they mean is *addition*.

Kripke's sceptic requires of the meaning factualist that, given unlimited epistemic access, S should be able to uniquely identify the fact that they mean *addition* by '+'. From this it follows that, granting unlimited epistemic access, if there is a fact of the matter about what S means by '+', S should be able to rule out all possible alternative interpretations of what they might mean by '+'. However, Kripke's sceptic alleges, this is precisely what S cannot do: it turns out that it is equally compatible with all the facts, even granting unlimited epistemic access, that what S has means[106] by '+' is *quaddition*, defined as the function 'Perform *addition* if the numbers involved are less than 57, else return the answer '5". If so, i.e. if S in fact means *quaddition* by '+', the answer to a '68 + 57' query, consistent with this previous usage, would in fact be '5'.

---

[105] Since it is the case that any particular competent language-user's quotidian arithmetical calculations are finite in number, it is always possible to confront a given language-user with a calculation they have not encountered before. For the purposes of simplicity, however, Kripke's sceptic assumes it that the language-user in question has never before calculated with numbers higher than 56.

[106] A complication here is that Kripke sets up the sceptical paradox by challenging the meaning factualist to cite some fact about themselves that enables them to conclude that they have *meant* '+' in the past. However, as Kripke explains, one should not thereby get the impression, that the sceptical paradox concerns only what one *meant* as opposed to what one *means* currently:

> "If the sceptic is right... there can of course be no fact about which particular function I meant, and if there can be no fact about which particular function I meant in the *past*, there can be none in the *present* either. Before we pull the rug out from under our own feet, we begin by speaking as if the notion that at present we mean a certain function by 'plus' is unquestioned and unquestionable. Only *past* usages are to be questioned. Otherwise, we will be unable to *formulate* our problem." (Kripke 1982, pp.13-4)

In other words, Kripke formulates the sceptical paradox in this way so as to avoid self-referential problems familiar to sceptical arguments. The sceptic cannot legitimately assert, for instance, that "There is no fact of the matter about what anyone means on any given occasion", since this would either be false or meaningless.

Kripke's sceptic's challenge to S is thus: granting unlimited epistemic access, cite some fact about yourself that enables you to rule out the possibility that what you mean by '+' is actually *quaddition*, and thus that the 'correct' way to respond to this calculation request is with '5', not '125'.

Kripke's sceptic then considers various candidates for something to supply the relevant consideration that would rule out the ersatz interpretation of '+', *quaddition*. An obvious candidate is the set of S's previous applications of '+'. However, since, by hypothesis, S has never performed a calculation involving numbers higher than 56 before, all of S's previous calculations involving '+' are consistent with both S meaning *addition* and S meaning *quaddition*. Any finite number of previous calculations can provide us with no way of ensuring that some suitably-altered *quaddition*-like interpretation be ruled out as a legitimate interpretation of what S means by '+'. Access to past usages of a term will not succeed in uniquely identifying (under ideal epistemic circumstances) the relevant concept picked out by '+' by S.

The next candidate discussed (Kripke 1982, pp.15-7) is the notion of some internalised *rule,* or more specifically *algorithm,* that S is self-consciously following. For example, when faced with '$x + y$' on previous occasions, I followed a rule that I had explicitly 'engraved' on my mind through following a certain procedure. Let's say that this procedure is as follows: I counted x marbles in one pile, y marbles in another pile, joined the piles, and then counted the number of marbles in the final pile. I then undertook to understand the notion of '+' in line with the results of this counting procedure. To be consistent, then, with this previous, 'engraved' rule, I should perform '+' by carrying out this counting procedure. This would then ensure that I could rule out *quaddition* as a legitimate interpretation. In effect, performing '+' calculations is to be understood in terms of some more primitive rule, *counting*. Through so doing, one finds that performing *quaddition* would be incompatible with the results of following this more primitive procedure.

Kripke's sceptic's response is to point out that one can re-run the sceptical paradox for this new rule of 'counting'. The challenge to S is now to cite some fact that enables them to rule out the possibility that one was in fact *counting* when one invoked the notion of 'counting', as opposed to *quounting*, defined as 'perform *count* if

the numbers involved are less than 57, else return the answer '5'. We now find that every previous performance of 'counting' is consistent with both *counting* and *quounting*, and thus that appealing to 'counting' as settling whether one means *addition* or *quaddition* by '+' can only succeed in delaying the onset of the sceptical paradox, and is of little help. In fact, one can generalise this point without too much difficulty: if one attempts to rule out ersatz alternative interpretations as to what constitutes the correct usage of 'p' through appeal to some further rule, one faces the difficulty that one needs to be able to rule out ersatz alternative interpretations of that further rule.[107] In short, citing further instructions will only yield a regress, since one needs to be able to demonstrate that these instructions themselves cannot be followed in an alternative way.

Similar considerations apply if one attempts to appeal to particular items regarding our *mental* history, available through introspection. Suppose one attempted to maintain, in response to Kripke's sceptical paradox that facts about one's intentions fixed the relevant metalinguistic fact. The evident difficulty here is that the sceptical paradox can simply be re-run for the appropriate intention. Even granting that S's intention to *add* when confronted with '+' entails that S means *addition*, Kripke's sceptic will ask what considerations are available that enable us to uniquely identify the intention to *add* as opposed to *quadd*. The sceptical paradox is merely delayed by such a response to the sceptic. An appeal to intention to supply the relevant

[107] Neil Tennant here alleges that this move is suspect. (Tennant 1997, pp.108-24) He maintains that Kripke has failed to demonstrate that this process of re-interpreting related rules *could* continue indefinitely. For instance, he notes that one of the features of addition is that, when it comes to integers, the following rule always applies: for any *n*, there are *n+1* ways of reaching that number by adding two integers. Note that the this rule would not be the case if 'adding' were understood as meaning *quaddition*.

It's hard to know what to make of this point. After all, it does not seem to be *constitutive* of the notion of '+' that it respect this rule, (Kusch 2006, p.132) so one wonders why the quadder cannot simply accept that this rule is not true of *quaddition*. If, however, one decided that it was constitutive of '+' that it obey this rule, then the quadder can simply utilise a different interpretation of this rule such that it only applied to numbers below 57, at which point the rule 'bends'.

The general point Tennant is making, however, is that it becomes harder and harder to construct a consistent picture of 'bent' rules once one starts thinking through the various analytic entailments of our standard understanding of some arithmetical function. At this point, Tennant alleges that the burden of proof is on Kripke's sceptic to supply a coherent, fully-worked through proof that an ersatz alternative understanding of '+' (say), such that it deals with all the required analytic entailments and remains consistent with our previous usage, is possible. (Tennant 1997, p.123)

This seems to me to mislay the burden of proof. After all, Kripke's sceptic has already provided *inductive* grounds for thinking that the business of supplying ersatz interpretations of relevant connected rules could continue indefinitely. To be sure, as more connections are added, it requires more theoretical imagination, and of course at some point it is reasonable to suppose that we will lack such theoretical wherewithal, but then that would only imply that we cannot think of further 'bent' rules that are equally compatible, not that such rules are not in principle available.

meaning-fact unless one can uniquely identify that the particular intention is to *add* as opposed to perform some ersatz alternative.

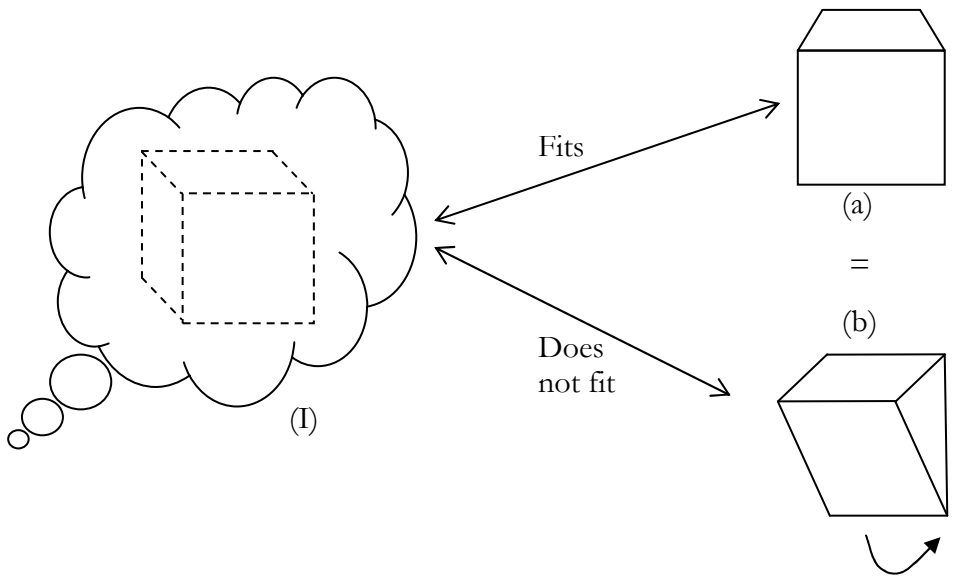
It may be maintained that forming the intention to *add,* say, has its own distinctive phenomenology, such that one has a particular mental image or *quale* in mind, an image that naturally guides us in performing one particular pattern of application of some term to the exclusion of alternatives. The difficulty here is that a quotidian mental image, much like an explicit rule one 'engraves' on one's mind, can still be rendered compatible with non-standard interpretations. Wittgenstein's example of applying the term 'cube'[108] seems apposite here. If we take the view that some schematic mental image of a cube comes before S's mind on hearing or using the word 'cube', for instance, we might think that this mental image guides S in their usage of the term 'cube' to accord with *cube*. However, as represented in the figure below, having some distinctive mental image (I) come before the mind seems equally compatible with *qube,* understood as apply the term 'cube' where one is faced with either cubes or triangular prisms at a particular angle of projection:

108 "What really comes before our mind when we *understand* a word? – Isn't it something like a picture? Can't it *be* a picture?

Well, suppose that a picture does come before your mind when you hear the word "cube", say the drawing of a cube. In what sense can this picture fit or fail to fit a use of the word "cube"? – Perhaps you say: "It's quite simple; – if that picture occurs to me and I point to a triangular prism for instance, and say it is a cube, then this use of the word doesn't fit the picture." – But doesn't it fit? I have purposely so chosen the example that it is quite easy to imagine a *method of projection* according to which the picture does fit after all.

The picture of the cube did indeed *suggest* a certain use to us, but it was possible for me to use it differently." (Wittgenstein 2001, §139)

***Philosophical Investigations* §139**



Here, the same object appears to fit the mental image of a cube at one angle (a), but not at another (b). One wants to say that the mental image (I) would uniquely identify *cube* as the meaning of 'cube' and thus, in accordance with *cube*, it would be incorrect to apply the term 'cube' to this object. However, it is mysterious how the mental image could allow us to uniquely identify *cube* as the relevant meaning-fact, since *qube* seems to accord with the mental image just as well. If we suppose that S had never encountered triangular prisms at this particular angle of projection before, S's use of 'cube' would be equally compatible with *cube* and *qube*. If we also suppose that the mental image (I) was present in all cases of applying 'cube', this is of no help in enabling us to identify 'S means *cube*' as the relevant meaning-fact, since *qube* is equally compatible with the mental image. The mental image does not enable us to uniquely identify the metalinguistic fact.

Kripke also considers the idea that, rather than conceiving of mental items as bringing to mind an image within our mental experience, we might conceive of them as *sui generis* states of mind that *just do* guarantee a unique pattern of application such that they cannot be interpreted otherwise.[109] Such mental items are not

[109] Kripke's characterisation of the position under attack is vulnerable to the complaint that he assimilates the notion of 'introspectable' to 'able to be called upon within mental experience', and that

'introspectable' in the sense that we can experience them in our mental lives, but nonetheless they play a role in guiding our application of the term. So, for instance, in using the word 'cube', S cannot call to mind something within experience, but nonetheless there is a *sui generis* mental item that guides S in using the term 'cube' in accordance with *cube*. McGinn suggests that we think of such *sui generis* mental states along the lines of beliefs, thought, intentions, and hopes, where there is no real *experience* of believing or intending something, but nonetheless we accept first-personal authority on such matters. (McGinn 1984, p.89) One might say similarly of meaning *cube* by 'cube'. S has no experience of meaning *cube*, but S could reflect upon his mental states and conclude, with first-personal authority, that they were guided by a *sui generis* mental state of meaning *cube* by 'cube', and this would entitle S to claim that they had identified the relevant meaning-fact.

The problem with this suggestion, noted by Wright, is that it seems to have missed Kripke's sceptic's point.[110] Even if we grant that there are many occasions where a person is able to introspect and enjoy first-personal authority about, say, their intentions, one would still need to know *how* this is achieved before the sceptical challenge is met. At the very least, we'd need a sketch of the relevant first-personal epistemology of such *sui generis* mental items. How can S determine, even in ideal epistemological circumstances, that the relevant *sui generis* mental state is that of picking out meaning *cube* as opposed to *qube*, say? The present proposal seems to amount to little more than the claim that S, without bringing this mental state within experience as such, just is able to do so. As Wright puts it, "Perhaps Kripke's use of 'desperate' in response to the idea of meaning, etc., as *sui generis* states was uncalled-for. But the characterisation, 'mysterious'… was not." (Wright 1989, p.114) Without further work, appealing to *sui generis* mental items that fix the relevant meaning-fact, only succeeds in raising further questions about how this is possible.

---

this is unduly narrow an understanding. Colin McGinn argues that 'introspectable' should be understood more along the lines of there existing a first-/third-personal asymmetry. (McGinn 1984, p.88-9) That is, even if someone had access to the sum total of S's behaviour, they would still be in a lesser epistemological position than S, who has access to their mental states.

As a result, we need to digress somewhat into points made by McGinn and Wright in order to buttress Kripke's argument here.

110 "This is about as flagrant an instance of philosophical stone-kicking as one could wish for." (Wright 1989, p.113)

We have, as it were, a 'first run' at the sceptical paradox. At least as far as the candidates canvassed thus far, the meaning factualist position Kripke's sceptic is attacking is internally inconsistent. The meaning factualist is committed to the claim that, given unlimited epistemic access, we should be able to uniquely identify the relevant meaning-facts such that we are able to distinguish correct from incorrect usage of a particular term and justify us in our metalinguistic claims. We ought to be able to rule out ersatz alternative interpretations of what S means by 'p'. However, as far as the candidates we have so far considered, no item in S's previous usage of the term, nor any item in S's mental states or mental history, is able to provide a basis for an elimination of these ersatz alternatives. It is always logically possible that some ersatz alternative, equally compatible with all possibly available facts yet nonetheless logically distinct, is available.

Assuming that there is not some other candidate we have failed to consider that might perform this task, Kripke's sceptic would have it that the meaning factualist position is internally inconsistent. It leads us to the view that metalinguistic facts of the matter exist, and that, if such facts exist, they will be uniquely identifiable under ideal epistemological circumstances, and yet they are not uniquely identifiable under ideal circumstances. The conclusion to take away from this *reductio*, Kripke's sceptic maintains, is the 'intolerable' conclusion that there are no meaning-facts, no facts of the matter expressible in the form 'S means that *p*'.

3.4 – Arguments Against Dispositionalism

Hitherto, we have not considered dispositionalist responses to Kripke's sceptical challenge. We need to consider what difference it might make if we expand the range of available facts to include facts, not necessarily about the actual world, but facts about possible worlds. Rather than looking at, as it were, present and historical facts, one might appeal to *dispositions* to respond in a particular way. Given the epistemological idealisation invoked in the 'Unlimited Epistemic Access Yields Unique Identification Thesis' attributed to the would-be meaning factualist, set-up of the sceptical paradox, it is quite reasonable to include such hypothetical facts within S's possible grasp. On a simple-headed dispositionalist account, then, S means *p* by 'p' if and only if S *would* use 'p' in accordance with *p*. For example, S means *addition* by '+' if and only if S would return the *sum* of the relevant numbers involved when queried. This dispositional state is incompatible with *quaddition*, since with *quaddition* one is disposed to return the value '5' with queries where one of the numbers is above 57. One could thereby point to the relevant disposition, the disposition to return the *sum* of the numbers involved, i.e. consistent with *addition* but not *quaddition*, as ruling out the ersatz alternative interpretation of '+', *quaddition*. As such, it appears to answer the sceptic's challenge.

Kripke invokes three main arguments[111] designed to defeat a dispositionalist account – the finiteness argument, the normativity argument, and the mistake argument. I shall give some indications as why the first two arguments here are inconclusive. However, as we shall see, the mistake argument succeeds in providing reasons to think that a dispositionalist solution to the sceptical paradox is not available, and it is this argument that shall prove most relevant to my subsequent argumentation with regards to the explicative project involved in the standard interpretation of reflective equilibrium.

Kripke's finiteness argument against dispositionalism runs as follows. The sort of candidate Kripke's sceptic is looking for is such that it is able to rule out all possible ersatz alternatives before one can properly conclude that one has uniquely identified the relevant meaning-fact. However, Kripke argues (Kripke 1982, pp.26-7) that a

[111] Here I follow Kusch (Kusch 2006).

disposition, i.e. a fact about the ways in which a speaker is disposed to respond, will not be able to do so, since the totality of their dispositions are finite[112] whereas a meaning-fact covers an indefinite number of applications. Kripke maintains, for instance, that I am only really disposed to perform *addition* when faced with a '+' query where the numbers are relatively manageable: my disposition to *add* only applies to a finite number of cases, not to the indefinite number of cases over which '+' may apply. My dispositional state with regards to '+' is captured just as well, if not better, by *skaddition* (Miller 2007, p.173), defined as 'perform *addition* if the numbers involved are small enough to handle, else return the answer '5". Needless to say, this function will extensionally diverge from *addition* in cases where the respective numbers involved are enormous. Citing my dispositional state does not allow me to rule out meaning *skaddition* by '+' as opposed to *addition*.

I am liable to think, following Blackburn (Blackburn 1984, pp.35-7) and Boghossian (Boghossian 1989, pp.164-6), that this argument in its current form is not decisive because it confuses individual dispositions (e.g. regarding '+' queries) with one's overall dispositional state. When it comes to describing my overall dispositional state, it is plausible to think that the relevant disposition to *add* when faced with a '+' query is nonetheless still in play, but is countervened by other relevant dispositions. It is entirely akin, as Blackburn points out, to the brittleness of glass being constituted by its disposition to break when struck forcefully, but there being situations where there are countervening factors, such as creating a glass object that won't break at certain points. The relevant individual disposition is still present, but the total dispositional state is such that it will not always break when struck forcefully. Similarly, the disposition to *add* when faced with '+', is still present, but that does not imply that I will actually perform *addition* in all cases. As such, it is unfair to maintain that some individual dispositional fact only extends finitely, and thus the finiteness argument is inconclusive.

Kripke's second argument against dispositionalist theories of meaning centres around his requirement that meaning-facts supply 'normativity', understood in the sense that

[112] Here there is an ambiguity between dispositions being finite in the sense that the disposition itself, which is presumably encoded in the brain, has to be composed of a limited series of instructions, and dispositions being finite in the sense that they only apply to a finite number of cases. It is the latter idea that does the work in Kripke's finiteness argument. However, whilst the former idea seems highly plausible, obvious even, this latter claim would appear to be false.

if S means that *p*, it follows that there is a semantic prescription (for S) to utilise 'p' in accordance with *p*. However, a dispositionalist theory, however sophisticated, will only list what S *will* do under certain circumstances, and this can in no way supply the relevant prescription alleged to be essential to any meaning factualist account:

> Suppose I do mean addition by "+". What is the relation of this supposition to the question how I will respond to the problem "68 + 57"? The dispositionalist gives a *descriptive* account of this relation: if "+" meant addition, then I will answer "125". But this is not the proper account of the relation, which is *normative*, not descriptive. The point is *not* that, if I meant addition by "+", I *will* answer "125", but rather that, if I intend to accord with my past meaning of "+", I *should* answer "125". Computational error, finiteness of my capacity, and other disturbing factors may lead me not to be *disposed* to respond as I *should*, but if so, I have not acted in accordance with my intentions. The relation of meaning and intention to future action is *normative*, not *descriptive*.

If S means that *p*, there is a semantic prescription for S to use 'p' in accordance with *p*. Should they fail to do so, they are liable to semantic sanction and criticism. A dispositionalist account, however, is ill-equipped to explain from where such a prescriptive implication derives. If it turns out, for instance, that S starts to answer '5' when faced with large sums, it seems that the only conclusion one could derive would be along the lines of: S didn't mean *addition* by '+' after all, or, other dispositions are interfering with S's disposition to apply *addition* when confronted with '+' queries, or something along these lines. At no point does it follow, Kripke argues, that S's has acted in a way that, in some sense, they *ought not* have done. Dispositional facts do not yield semantic prescriptions; meaning-facts do; therefore, dispositional facts do not yield meaning-facts.

This argument, however, is vulnerable to the complaint that it is inessential to meaning-facts that they supply semantic prescriptions. We have hitherto relied solely on the thought that, if there are meaning-facts, they supply us with *correctness-conditions* for the use of that term.[113] However, it does not follow that there are any particular semantic prescriptions on how we ought to proceed. Anandi Hattiangadi (Hattiangadi 2006), Kathrin Glüer & Åsa Wikforss (Glüer & Wikforss 2009), and Paul Boghossian (Boghossian 2005) all argue that the simple platitude that a term's

[113] As I indicated earlier (see ft.95), I have only attributed to the meaning factualist the Meaning-Facts Supply Correctness-Conditions Thesis, in which, if there are meaning-facts, they supply conditions of correct and incorrect usage.

possessing meaning implies correctness-conditions does not, of itself, mean that any particular speaker is *obliged* to act in a particular way. It simply means that typically we think of the notion of 'meaning' as implying a standard of correct use. There is no semantic prescription, for instance, to use 'horse', say, only to things that are horses: trivially, if someone holds a gun to my head and commands me to say 'horse' when I am confronted with biscuit-tins, I am under no semantic obligation to refrain. We seem to only have action-guiding semantic prescriptions if we happen to have the desire to use our words correctly (Boghossian 2005, p.207). However, this makes the relevant normative force operating on the speaker entirely hypothetical, akin to 'If you don't want to get your hair wet, take an umbrella'. There does not appear to be any categorical semantic prescription present such that one *just ought* to use one's terms correctly.[114] It is contentious, then, whether Kripke's sceptic's normativity argument succeeds as an argument against dispositionalism, since the dispositionalist need not regard meaning-facts providing semantic prescriptions as a commitment they need to uphold.

Kripke's third argument against dispositionalism, the mistake argument,[115] I regard as successful. We start by noting the intuitive commitment that meaning-facts supply correctness-conditions. As such, the meaning-fact allows us to identify correct usage and mistaken usage, i.e. linguistic error. The dispositionalist theory, then, ought to be

---

[114] Hattiangadi's, Glüer & Wikforss's, and Boghossian's way of presenting the matter is not without opposition, however. Daniel Whiting (Whiting 2007), in response, makes the following point: The fact that a meaning claim implies a correctness-condition has a very obvious implication about what one may, may not, ought, or ought not, do. Just as, using Hattiangadi's own example (Hattiangadi 2006, p.224), whether a child is tall enough to go on a ride at a theme park is itself a straightforward, non-prescriptively-normative fact, i.e. whether the child meets the appropriate standard, it is likewise the case that whether a person uses a term correctly is itself a straightforward, non-prescriptively-normative fact, i.e. whether their usage meets the appropriate semantic standard. However, Whiting points out (Whiting 2007, p.136), in both cases, these facts both have prescriptive, action-guiding implications: if the child is tall enough, they *may* go on the ride without sanction or criticism; similarly, if the usage of some term meets the standard of correctness, they *may* use it without sanction or criticism (on semantic grounds, at least).

On Whiting's view, there may be countervailing norms in operation for why one *ought* to use one's terms in ways that, semantically-speaking, do not meet the respective standard (e.g. one ought to use one's words incorrectly if doing so prevents you getting shot), but that is quite consistent with the thought that some semantic (prescriptive) norm is still in force. That the semantic norm to use one's words may be overridden is itself no reason to think that there is no semantic norm at work.

This debate is ongoing, and space precludes a full examination of this intriguing dispute. Nonetheless, and as Glüer & Wikforss note (Glüer & Wikforss 2009), one need not rely on meaning supplying semantic prescriptions in order to formulate the Kripke's sceptical challenge. As we shall see, one only need rely upon what is consensually regarded as a platitude: that meaning-facts imply correctness-conditions.

[115] The term is Kusch's (Kusch 2006, p.97, c.f. p.121). The mistake argument may be found in (Kripke 1982, pp.28-32).

such that it would allow us to identify such mistaken usage through identifying the relevant meaning-fact. The mistake argument now runs that, even if a sophisticated dispositionalist theory can overcome the various difficulties already noted,[116] it will be unable to demonstrate if and when a particular use of a term is mistaken.

Consider how a dispositionalist theory might discharge the commitment to be able, granting unlimited epistemic access, to identify mistaken applications of a given term. To use Kripke's own example (Kripke 1982, p.29), how might a dispositionalist set about establishing that someone who systematically miscarries[117] in applying '+', whilst insisting that they are actually adding, is committing some linguistic mistake?[118] The dispositionalist will typically respond that what S means by '+' is revealed by their disposition, under ideal conditions, to *add* when faced with '+' queries. Waiving the problem of non-intentionally, non-semantically specifying such conditions, the question may be raised at this point: why should we suppose that their disposition is to *add* under ideal conditions? Why not say instead that their disposition is to *madd*, whereby they are disposed to *madd* (i.e. *add-but-miscarry*) when faced with '+' under ideal conditions?

Consider the following analogy, proposed by Kusch (Kusch 2006, p.121): suppose we have two machines, A and B. Machine A is a standard adding machine, but due to some accidentally loose wire, fails to carry. Machine B, by contrast, is a machine that is designed to *madd*. It was produced by taking a normal adding machine and deliberately loosening the wire. If we consider the total dispositional state of the machines, we see that they are functionally-isomorphic: both will produce the same output given a certain input. However, one intuitively wants to say that one machine

---

[116] Kripke waives, for instance, the finiteness problem (Kripke 1982, p.30) in his formulation of the argument.

[117] By 'miscarrying', we simply mean that the adder fails to carry over surplus from the units column to the tens column, from the tens column to the hundreds column etc. Thus a miscarrier will say '68 + 57 = 15' because they haven't carried over the 1 from the units to the tens column, and the 1 from tens column to the hundreds column.

[118] Again, here we need to be careful to point out that the relevant error here is that of linguistic error, as opposed to a calculation error. This is why we need to stipulate that the mistake here is systematic and insisted upon: it is not the case that the person is disposed to miscarrying under conditions of, say, bad lighting, drunkenness, etc. but adds perfectly well under ideal conditions. The problem is that the person seems to miscarry when performing '+' on a regular basis, whilst insisting that they are 'adding' correctly. At this point, we do not wish to say that they mean *addition* by '+' and are miscalculating, but rather that they've misunderstood '+'. They have failed to use the term in accordance with its meaning, and have thus committed some linguistic, as opposed to calculative, error.

is carrying out *madd* and working properly whereas the other machine is carrying out *add* but is faulty. The question is, however: on what basis could a dispositionalist make such a claim?

The point here is that the dispositionalist faces a degree of theoretical choice as to how to interpret the matter when it comes to constructing the relevant theory: they could interpret things such that the speaker means *addition* but, for some countervening reason, perhaps the disposition to miscarry coming into play, is systematically disposed to miscarrying when applying the term, thus committing a linguistic mistake. However, they could also interpret things such that the speaker actually means *maddition*, whereby they just mean something else by '+' than we do, and no linguistic mistake is involved. Both theories account for the facts as to how S responds when faced with '+', but one regards it as a systematic error due to some countervening disposition, whereas the other regards S as meaning something else by '+' than we do.

Now, we've argued that in order for some consideration to count as meaning-determining, it needs to supply us with criteria that, given unlimited epistemic access, allow us to distinguish correct from incorrect usage, and thus to identify linguistic mistakes. On the dispositionalist's account, in order to determine whether S is genuinely committing a linguistic mistake, one would need to know which theory supplies the correct understanding of S's dispositional state: is it that they mean *add* but other dispositions are getting in the way, or is it that they really mean *madd*? The problem here is that the dispositionalist has already exhausted the relevant dispositional facts – S is identically disposed (overall, that is) under both theories. Thus, the dispositional account is unable to tell us whether S's use of '+' should count as a mistake or not, and thus fails to secure meaning factualism.

In essence, the objection is very similar to the Quine-Duhem problem.[119] There is always an element of theoretical choice in how one arranges the theory to account

[119] Here the problem is the underdetermination of theory by evidence: "The totality of our so-called knowledge or beliefs, from the most casual matters of geography and history to the profoundest laws of atomic physics or even of pure mathematics and logic, is a man-made fabric which impinges on experience only along the edges. Or, to change the figure, total science is like a field of force whose boundary conditions are experience. A conflict with experience at the periphery occasions readjustments in the interior of the field. But the total field is so underdetermined by its boundary conditions, experience, that there is much latitude of choice as to what statements to reevaluate in the

for the relevant facts. The *overall* dispositional state of a speaker may be fully-specified (since we have unlimited epistemic access), as it were, at the edges, perhaps in the language of possible stimuli and responses. Thus, we can stipulate that machine A and machine B are functionally-isomorphic over all possible cases. However, if one then wishes to capture an individual meaning-facts, say a fact about what S means by '+', one is faced with an indefinite number of theoretical choices as to how one does so given S's overall dispositional state: one could operate with what we'd regard as a standard understanding *p*, such that S is disposed, under ideal conditions, to use 'p' in accordance with *p*; or one could operate with an ersatz alternative understanding *q*, such that S is disposed, under ideal conditions to use 'p' in accordance with *q* but countervening dispositions, such as a disposition to miscarry, interfere. Whether it is correct to then say that S is making some linguistic error in the use of a *particular* term will depend upon our theoretical choices.

The dispositionalist has, at this stage of the argument, two options: they can either argue that all totally-extensionally-equivalent dispositionalist theories *just are* equivalent theories, or they can maintain that, *once all the facts are in*, differences in the internal theoretical apparatus will be reflected in extensional differences, and thus it just could not be the case that a faulty machine and a working machine be functionally-isomorphic. The first option may be easily dismissed if one takes at all seriously the need to able to identify when S has made a *mistake* in their application of some individual term. Whether S has made a mistake, as opposed to whether S has simply meant something else, depends upon how the given term is theoretically-construed, i.e. in terms of the relevant dispositions. Without any way of settling which theory to adopt, which a dispositionalist cannot do since the total dispositional state (which is available given the ideal epistemological circumstances) is, *ex hypothesi*, identical under the two theories, we cannot determine whether a mistake, a linguistic error, has been made.

---

light of any single contrary experience. No particular experiences are linked with any particular statements in the interior of the field, except indirectly through considerations of equilibrium affecting the field as a whole." (Quine 1951, pp.42-3)

One may regard Kripke's 'sceptical paradox' as, if you like, a limiting case of the Quine-Duhem problem, for the reason that the Quine-Duhem thesis pertains to situations of limited epistemic access. The Kripkean problem, however, is that we have reason to think that no improvement in our epistemic situation will remove this difficulty.

The second option, denying that there could be extensionally-equivalent theories that return different answers as to whether S is in error, may be a more successful avenue to take, but we await further argument. It amounts to the claim that hypothetical situations such as the functionally-isomorphic machines yet divergent-in-faultiness will not carry over into sufficiently-complex cases. Once a certain degree of complexity is involved, such that our reactions to myriad situations are taken into account, considerations will emerge that privilege one particular dispositionalist theory over its competitors. At this point, however, one may insist that the responder has missed the point: at no point need Kripke's sceptic insist that actual, practical questions of theory choice are intractable, and unable to be resolved.[120] The point is that it remains a logical possibility that there are ersatz theoretical alternatives available that return divergent answers as to whether S has committed some linguistic error with regards to a particular term. Unless the dispositionalist can demonstrate that situations analogous to Kusch's two machines example are logically impossible, Kripke's sceptic has shown that the dispositionalist is unable to uniquely identify the relevant meaning-fact such that one can determine if a linguistic mistake is being committed.

I conclude, therefore, that Kripke's sceptic does indeed supply us with sufficient considerations to refute dispositionalist responses to the sceptical paradox. Any attempt to specify the relevant meaning-facts will still face the problem of eliminating alternative ersatz proposals that, with a bit of theoretical jiggery-pokery, can be specified such that the overall dispositional response is identical to the standard proposal with different theoretical jiggery-pokery. The claims about what constitutes the meaning of some particular term 'p' will, however, be different, and will consequently return different answers as to whether any particular application of 'p' is correct or involves a linguistic mistake. The dispositionalist, even under epistemologically ideal circumstances, will not be able to uniquely identify the meaning-fact. Kripke's sceptical paradox, having considered the plausible candidates for identifying such meaning-facts, remains unanswered.

[120] Indeed, if it were acceptable to claim that Kripke's ersatz alternatives aren't *really* possible, but just bizarre logical possibilities, one could answer the sceptical paradox with a wave of a hand. Kripke's sceptic isn't committed to the thought that there is anyone out there who actually does perform *quaddition* by '+'; the point is merely that such a situation is logically possible, and this undermines the intuitive meaning factualist position.

3.5 – The Sceptical Solution

Responses to the paradox are immensely various and complicated, and necessarily any overview of Kripke's sceptical paradox retains a degree of provisionality. However, having considered the plausible candidates, it is fair to say that no 'straight solution',[121] a solution that cites some candidate that fixes the relevant meaning-fact, has emerged to answer Kripke's sceptical challenge.[122] Such a solution may be forthcoming, or there may be space for replies that question the formulation of the sceptical paradox itself. One might question, for instance, the underlying characterisation of 'meaning factualism', such that one regards Kripke's argument as a *reductio ad absurdum* of a particular conception of 'meaning factualism', and not meaning factualism *per se*. (Kusch 2006, Wilson 1994) Or it may be that Wright's 'judgement-dependent' account (Wright 1987a, Wright 1989a), which challenges the inherent reductionism of the sceptic's challenge, can be finessed so as to secure the relevant facts through claiming that, under ideal conditions, one's judgement that one means *p* by 'p' *constitutes* the relevant meaning-fact.

In the next chapter, I shall outline exactly what lesson I am taking from the sceptical paradox, because I do not wish to endorse the sceptical paradox as such, but rather wish to adopt a strand of the sceptic's argument for my own purposes. In this last section, however, I wish to consider what is standardly regarded as Kripke's own response to the sceptical challenge: his 'sceptical solution'.[123] Aside from reasons of

---

[121] The distinction between 'straight solutions' and 'sceptical solutions' is Kripke's:

"Call a proposed solution to a sceptical philosophical problem a *straight* solution if it shows that on closer examination the scepticism proves to be unwarranted; an elusive or complex argument proves the thesis the skeptic doubted... A *sceptical* solution of a sceptical philosophical problem begins on the contrary by conceding that the sceptic's negative assertions are unanswerable. Nevertheless our ordinary practice or belief is justified because – contrary appearances notwithstanding – it need not require the justification the sceptic has shown to be untenable." (Kripke 1982, p.66)

[122] Here I follow Kusch in his estimation that "All of the investigated versions of meaning determinism fail." (Kusch 2006, p.147) At the very least, if there is such a solution, it does not, as yet, enjoy especial prominence in the secondary literature.

[123] Here, I follow Miller's exposition (Miller 2007, pp.175-7), which is consonant with Boghossian's (Boghossian 1989) and Wright's (Wright 1994) view that Kripke offers a 'communitarian' solution to his sceptic's paradox, in which the correctness of metalinguistic claims is established via utilising communal criteria of warranted assertibility.

However, it is not clear that Kripke intended his sceptical solution to be understood as a piece of constructive explanation as to what constitutes the correctness of metalinguistic claims. If one consider's what Boghossian calls (Boghossian 1989, p.155) Kripke's 'official' stance, it looks like Kripke intends his 'solution' as a purely *descriptive* account of when everyday language users *do* regard metalinguistic claims as correct, as opposed what makes it the case that they are correct:

"Following Wittgenstein's exhortation not to think but to *look*, we will not reason *a priori* about the role such statements [concerning attributions of meaning] *ought* to play; rather we will find

purely surveying the relevant discussion, there are two reasons, as can only become clear in the next chapter, why investigating the sceptical solution is helpful for our concerns. Firstly, it helps us to appreciate that the strand of argument I am adopting from Kripke's sceptical paradox can broadened to incorporate considerations of the *correctness* of a meaning claim, as opposed to the *truth* of a meaning-fact. That is, the strand of argument I am using applies to considerations of the warranted assertibility of some meaning claim just as much as considerations of meaning-facts. , Secondly, it may seem that the sort of response I recommend in reaction to the strand of argument I'm adopting from Kripke's sceptic, a 'Wittgensteinian' response modelled on Wright's Wittgenstein in his recent 'Rule-Following Without Reasons' (Wright 2007), is itself a version of the sceptical solution. I shall stress (in §4.4) that this is not the case, so it pays to be clear in advance about the nature of the sceptical solution.

Kripke's sceptical solution, as standardly understood, results from accepting that the *correctness* of 'S means *addition* by '+" cannot be understood in terms of a truth-apt meaning-fact, but that nonetheless it can be understood in terms of *warranted assertibility*. Utilising 'criteria'[124] for the warranted assertion of a metalinguistic claim of the form 'S means that *p',* we are thus able to distinguish between correct (warrantedly assertible) and incorrect (not warrantedly assertible) metalinguistic

---

out what circumstances *actually* license such assertions and what role this license *actually* plays." (Kripke 1982, pp.86-7)

However, Kripke, *malgre lui*, introduces *substantive* assertibility conditions. Consider, for instance, his claims that "If Jones does *not* come out with '125' when asked about '68 + 57', we **cannot** assert that he means addition by '+'." (Kripke 1982, p.95, emboldening added) and "If the individual in question no longer conforms to what the community would do in these circumstances, the community **can** no longer attribute the concept to him." (Kripke 1982, p.95, emboldening added) Strictly speaking, a purely descriptive account would run that we *will* not assert that Jones means addition by '+' and that the community *will* no longer attribute the concept to him. These claims clearly invite the communitarian reading.

[124] Again, along the same lines considered in ft.123, it is unclear whether it is Kripke's intention is to utilise the notion of 'criteria' as a way of avoiding a constructive philosophical explanation. Again, the 'official' stance is that 'criteria' are tools of description: "Roughly speaking, outward criteria for an inner process are circumstances, observable in the behaviour of an individual, which, when present, **would** lead others to agree with his avowals." (Kripke 1982, p.100, emboldening added)

However, Kripke also utilises 'criteria' in a way that suggests a constructive explanation is being offered. When considering the Robinson Crusoe example, for instance, Kripke slides between 'criteria' and 'Wittgenstein's *theory*' of 'assertibility conditions': "*If* we think of Crusoe as following rules, we are taking him into our community and applying our criteria for rule following to him. The falsity of the private model need not mean that a *physically isolated* individual **cannot** be said to follow rules; rather that an individual, *considered in isolation* (whether or not he is physically isolated), **cannot** be said to do so. **Remember that Wittgenstein's theory is one of assertibility conditions**." (Kripke 1982, p.110, emboldening added) Strictly speaking, if one were to consider 'criteria' as simply tools of description, the most that could be said here is that an individual, considered in isolation, *would* not be said to be following rules (and even this claim looks somewhat suspect).

claims. The resultant constructive philosophical explanation would be such that one would eschew any talk of the truth/falsity of the metalinguistic claim; rather, one would only be interested in whether one could warrantedly assert the relevant metalinguistic claim. However, there would still be some basis for saying that, say 'S means *addition* by '+" is correct (warrantedly assertible), and yet that 'S means *quaddition* by '+" is not.

Kripke's substantive explanation of why we ought to regard some metalinguistic claim as correct, as standardly interpreted, runs on 'communitarian' lines. A communitarian would have it that community agreement over the application of 'p' *constitutes* conditions of warranted assertibility with regards to the usage of 'p'. S correctly applies '+' if they demonstrate, over a period of time, that their usage of '+' conforms to the communal expectations as to how one ought to apply '+'. At this point, they are, as it were, accepted into the linguistic community, and it is taken that their usage of '+' can be treated as reliable. Whilst, for Kripke, there is no *fact of the matter* as to whether S has used 'p' correctly, nonetheless there are conditions under which the community is justified in asserting that 'S has used 'p' correctly', and it seems acceptable to regard such a statement, once it meets these conditions, as correct (if not *true*).[125]

From this basic communitarian view, a couple of things follow. Firstly, there is no logical space for what philosophers often insist must be a logical possibility, i.e. that an entire linguistic community may come to mistakenly apply some particular concept. Here, whatever represents the settled community verdict on the matter *constitutes* what is to count as correct usage. Since correctness of application amounts to whether a given language user is warranted in applying their terms in the way that they do, and these conditions of warranted assertibility amount to whether they accord with the community usage, it follows that community usage cannot possibly

[125] As such, the standard approach here is to treat assertibility conditions as giving necessary and sufficient conditions for warranted assertibility. Again (see ft.123 and ft.124), we face difficulties here as to whether this is the best way to read Kripke, since Kripke's 'official' stance is that they are not necessary and sufficient conditions. (Kripke 1982, p.111) Kusch suggests that Kripke's 'appropriateness conditions' (Kusch 2006, p.28) should be understood along the lines of "It is often necessary and frequently sufficient for S to mean *p* that *a*-conditions a_1 … a_n are fulfilled." (Kusch 2006, p.27)

The problem, as I've already noted, is that if it is only *often* necessary for S to mean *p* that it, for instance, agree with community-usage, Kripke could not legitimately claim things like "If the individual in question no longer conforms to what the community would do in these circumstances, the community **can** no longer attribute the concept to him." (Kripke 1982, p.95, emboldening added)

diverge correct usage. As McDowell puts it, "We cannot hold, then, that the community 'goes right or wrong', by the lights of its understanding... 'rather, it just goes'." (McDowell 1993, p.223)

Secondly, there is no logical space for the idea that a 'language user', considered in isolation,[126] can give any content to the idea of their following a rule 'correctly'. Thus, Kripke's constitutive account is integrally linked to the 'anti-private language argument' he finds in Wittgenstein.[127] Kripke puts the point in the following way:

> If one person is considered in isolation, the notion of a rule as guiding the person who adopts it can have *no* substantive content... The situation is very different if we widen our gaze from consideration of the rule follower alone and allow ourselves to consider him as interacting with a wider community. Others will then have justification conditions for attributing correct or incorrect rule following to the subject. (Kripke 1982, p.89)

The reason why Kripke holds that a person, considered in isolation, cannot properly be regarded as following a rule (i.e. applying a term correctly), is that there can apparently be no basis for a distinction between it *seeming* to the isolated 'rule-follower' that they have applied the rule correctly, and actually following the rule correctly. (c.f. Wittgenstein 2001, §202) Since there is no relevant linguistic community here, conditions of warranted assertibility have no grip. All we can really say, according to Kripke, "is that our ordinary practice licenses him to apply the rule in the way it strikes him." As Kripke immediately notes, however, "this is *not* our usual concept of following a rule." (Kripke 1982, p.88) Genuine *justification* for meaning ascription only enters the picture once we have a relevant community, whose usage we can compare with that of the S's in order to ascertain whether we may warrantedly assert that they mean *p* by 'p'.

---

126 As Kripke notes (Kripke 1982, p.110), the point is not that a *physically isolated* individual cannot properly be said to be following rules, and thus using concepts correctly/incorrectly, but rather that it is only by considering the individual *in the context* of communal standards that notions of correctness/incorrectness gain traction.

127 Kripke's analysis of the structure of *Philosophical Investigations* has it that "The sections following §243 – the sections usually called 'the private language argument' – deal with the *application* of the general conclusions about language drawn in §§138-242 to the problem of sensations." (Kripke 1982, p.79)

There is, however, to be a quite basic difficulty with this communitarian 'sceptical solution'.[128] Construed as a constructive account whereby rough assertibility conditions are given for when it is correct (i.e. warrantedly assertible) that 'S means that *p*', we seem to be in something of a quandary when it comes to how to understand the satisfaction/non-satisfaction of those conditions themselves. Is it to be regarded as *true,* i.e. a *fact*, that S's usage of '+', say, accords with the usage found in the linguistic community? Here, Wright argues that:

> If so, that truth did not consist in any aspect of his finite use of that sentence or of its constituents; and, just as before, it would seem that his previous thoughts about that sentence and its use will suffice to constrain to within uniqueness the proper interpretation of the assertion conditions he associated with it only if he is granted correct recall of the content of those thoughts – exactly what the sceptical argument does not grant… So the case appears no weaker than in the sceptical argument proper for the conclusion that there *are* no such truths; whence, following the same routine, it speedily follows that there are no truths about the assertion conditions that any of us presently associates with a particular sentence, nor, *a fortiori*, any truths about a communal association. (Wright 1984, p770)

Recall that the condition for S's correct (i.e. warrantedly assertible) use of 'p' was roughly that it accorded with community usage. Wright's point here is that, if we construe the satisfaction of this condition as a factual matter, Kripke's sceptic's argument would resurface, apparently forcing upon us the conclusion that there is no fact of the matter concerning whether S's usage of 'p' conforms to community usage. What makes it the case that community usage is captured by *addition* as opposed to *quaddition* (or some suitably altered concept of *quaddition* such that it 'bends' after the point at which there are historical facts about how members of the linguistic community have utilised '+')? As such, there would be no fact of the matter as to whether S is warranted or not in using 'p' as they do, and thus again we have no means of distinguishing correct from incorrect usage of 'p'.

On pain of inconsistency, then, Kripke would be forced to grant that the correctness of regarding S's use of '+' as meeting the relevant assertibility conditions is itself a matter of warranted assertibility. The problem here is that we are faced with a regress when it comes to the characterisation of the assertibility conditions. Imagine that S has performed all sorts of '+' problems, and that S's teacher, T has examined S's

[128] This problem is noted in (Miller 2007, p.185-7, Wright 1984, pp.770-1; Zalabardo 1987, pp.37-41.).

responses. Here, for T to be warranted in asserting that 'S means *addition*' requires that it would be warranted for T to claim that 'S's use of '+' accords with community usage'. However, for T to be warranted in this claim would require (in part) that T's usage of their own terms, e.g. 'accords', itself accords with community usage. Again, we'd need to understand this in terms of warranted assertibility, so we imagine some further character, U, claiming that "T is warranted in asserting that 'S's use of '+' accords with community usage". Here U is only warranted in asserting this if their usage of relevant terms accords with community usage. But again, whether their usage accords with community usage is a matter of warranted assertibility, so we imagine another character, V, asserting that 'U is warranted in asserting that "T is warranted in asserting that 'S's use of '+' accords with community usage"'. Here V is only warranted in asserting this if their usage of relevant terms accords with community usage, etc. At some point, then, for warranted assertibility conditions to work as a constructive explanation of the correctness of some metalinguistic claim, it would need to be the case that someone's use of relevant terms *just did* accord with community usage.

As far as being a constructive piece of philosophy, then, i.e. some sort of explanation as to what makes 'S means that *p*' correct, Kripke's sceptical solution (as standardly understood) looks to be a 'non-starter'. (Miller 2007, p.187) There is an inherent instability in a communitarian account of what makes 'S means that *p*' correct in that, for any assertibility condition to obtain, there needs at some point to be facts concerning community usage against which S's use could potentially be compared. Whilst it may be the case that any given member of the linguistic community, T, having observed S's usage of '+' on repeated occasions, *regards themselves* as justified in maintaining that S means *addition*, to say that T *is* warranted in asserting 'S means *addition*' in virtue of the coincidence between S's usage and community-usage is something that at some point presupposes that there are facts concerning community-usage.

This concludes my selective survey of relevant material for my argument. I shall indicate in the next chapter how I'm adapting a strand of Kripke's sceptic's argument without endorsing the sceptical paradox *per se*. What we've seen, however, is that the plausible candidates for supplying a consideration that allows us to uniquely identify

the relevant meaning-fact are not successful, or at least, if there is a successful candidate, it has eluded our grasp. Proposed candidates for straight solutions merely relocate the point at which the sceptical challenge kicks in. As for the sceptical solution, as far as it is standardly understood, it at some point relies on the presupposition that meaning-facts are available. Understanding the sceptical solution, however, will allow us to clarify the nature of Wright's Wittgenstein's response to the particular strand of argument I shall adapt from these considerations.

Chapter 4 – Rule-Following and Explication

4.1 – Overview

My intention in this chapter is to draw out one strand of argument found within Kripke's sceptical paradox, and to relate it to matters of explication. I start off by separating out what is essentially an underdetermination argument in Kripke's sceptic's argument from its role within the constitutive sceptical challenge. Its import is epistemological; not constitutive or semantic. This argument, as we shall see when we reflect upon the sceptical solution, need not be restricted to claims about uniquely identifying relevant meaning-facts, but can be extended to considerations of warranted assertibility. Not only is not possible (given unlimited epistemic access) to uniquely identify any given claim about what S means by 'p' as true, I argue, neither is it possible to uniquely identify it as warrantedly assertible. As such, metalinguistic claims, claims of the form 'S means that *p*', are not 'rationally demonstrable' (a term I shall explain in §4.5), and point to an inherent limitation to our justificatory powers in that we cannot demonstrate the superiority of one metalinguistic claim over another.

I then follow Wright in reading Wittgenstein in the rule-following considerations as primarily concerned with this epistemological issue, as opposed to intending to formulate some constitutive sceptical paradox. Wright utilises a distinction between 'basic' and 'non-basic' cases of rule-following, a distinction which brings into sharp relief that we ought to expect it to be the case that one cannot always demonstrate the superiority of one metalinguistic claim over another, since there is an inherent limit to our justificatory powers. Wright argues that there must be cases of rule-following, 'basic' cases, where it is inappropriate to think that we might justify or explain why it is that we regard the correct pattern of application of 'p' to be *p*. There does not appear to be any inferential move being made, say, in our regarding 'red' as correctly applied in *this* particular way. Searching for constructive philosophical explanations that would enable us to uniquely identify particular metalinguistic claims as correct is something of a fool's errand. We know in advance that rule-following is going to bottom out in 'basic' cases where the most that can be said about why we apply the concept in the way we do is that *we just do*. Wright's Wittgenstein, by

contrast, eschews any attempt at such constructive explanation of why it is correct to apply 'red' like *this*, and merely describes our practice, where we regard metalinguistic claims as '*prima facie* legitimate', that is, where we (at least initially) make such claims without any call for justification or explanation, subject to future revision as difficulties become apparent.

My guiding idea here is that lessons garnered from Wright's Wittgenstein's approach as a response to the underdetermination strand of argument found within Kripke's sceptical argument can be carried over to matters of explication. We face a similar problem: In order to render a particular area of discourse transparently disagreement-entailing-error objective, we wish to single out one particular explicatum as replacing the explicandum. However, the underdetermination strand of Kripke's sceptic's argument establishes that there is an inherent limitation to our ability to justify favouring one particular explicatum over another, and this entails that our choice of a particular explicatum (to the exclusion of others) is not itself transparently a disagreement-entailing-error objective matter.

Nonetheless, it seems entirely appropriate to maintain that particular explications, as opposed to alternatives, are accepted. On an approach mirroring that taken by Wright's Wittgenstein, we may note here that it our practice is to adopt particular explicata with *prima facie* legitimacy in spite of this inherent limitation to our justificatory powers. So long as we are inclined to apply the explicandum universally or near-universally, we may, with *prima facie* legitimacy, adopt some explicatum that coincides with these inclinations, again subject to future revision should it become necessary. The underdetermination problem at this point becomes a hyperbolic worry: *we just do* adopt some explicatum as opposed to various possible alternatives, and accept that it supplies the pattern of correct application for the corresponding explicanda.

However, this raises the possibility of there being genuinely problematic cases where it would not be regarded as *prima facie* legitimate to adopt some explicatum to the exclusion of alternatives. These cases appear where we are faced with an explicandum involving basic concepts that are themselves applied non-universally, resulting in the explicandum itself being applied variably. Such cases are problematic

because we have no method, even under ideal epistemological circumstances, which would allow us to single out one particular explicatum, to the exclusion of alternatives, as superior to alternatives. In such problematic cases, I take it that any given explicatum would not be *prima facie* legitimate.

4.2 – Rational Demonstrability

We have from chapter 3 a partial overview of Kripke's sceptical paradox and the sceptical solution, at least as they are both standardly understood. As I have already indicated, my interest in the sceptical paradox extends only as far as harnessing a particular strand of the argument, without thereby endorsing the sceptical paradox itself. In this section, I shall draw out the strand of argument I am adopting from Kripke's sceptical paradox, translating it into terms of *correctness* as opposed to *truth*, before drawing out some implications with regards to correct/mistaken usage of any particular everyday concept.

First of all, we need to separate out what I shall refer to as the 'underdetermination strand' of the sceptical paradox from Kripke's sceptic's broader concern of attacking meaning factualism. The sceptical challenge was to cite some consideration that would potentially allow us to uniquely identify 'S means that *p*' as the relevant meaning-fact. The claim then, was that, even once we idealise our epistemological circumstances, our choice of which particular metalinguistic claim is true cannot be uniquely identified using all the available behavioural, mental and dispositional facts. Having examined all the plausible candidates, facts about historical usage, mental images, *sui generis* mental states that guarantee a particular application, and dispositions, we find that none succeed in allowing us, under ideal conditions, to uniquely identify the relevant meaning-fact. We always faced the same basic underdetermination problem: even once all the (non-semantic, non-intentional) possibly-available facts were in, one is always able to construct some ersatz alternative, equally compatible with the facts. All the possibly-available facts underdetermined the corresponding metalinguistic fact.

To formulate Kripke's sceptical paradox itself, however, one needs to invoke the Unlimited Epistemic Access Yields Unique Identification Thesis (if meaning-facts exist, we would, given unlimited epistemic access, be able to uniquely identify them). After noting that, having exhausting the list of plausible candidates, we have been unable to cite any candidate that would allow us to uniquely identify (given unlimited epistemic access) meaning-facts, we conclude that we are not able to uniquely identify the appropriate meaning-fact. We then conjoin this conclusion with the

claim that, if meaning-facts exist, we would be able to uniquely identify them, and yield the sceptical conclusion that there are no meaning-facts.

That the relevant claims are distinct may be emphasised by noting that one way of responding to the sceptical paradox here is to claim that the Unlimited Epistemic Access Yields Unique Identification Thesis is an optional commitment for the would-be meaning factualist. One might insist, for instance, that there being a fact of the matter, that S means that *p*, does not imply that such a fact is knowable, even in principle. The Unlimited Epistemic Access Yields Unique Identification Thesis assumes that truth is epistemically-constrained,[129] which is philosophically contentious. It is, on the face of it, a reasonable philosophical position to hold that there exist meaning-facts, but that they are not necessarily within our epistemic grasp, no matter how one idealises the epistemological circumstances. Once one separates out the constitutive/semantic concern, i.e. whether there is a fact of the matter expressible in the form 'S means that *p*', and the epistemological concern, i.e. whether 'S means that *p*' is knowable (and, if so, under what circumstances), one can see that the Unlimited Epistemic Access Yields Unique Identification Thesis is not something that the meaning factualist need necessarily endorse.

As when we discussed issues of semantic as opposed to epistemological vagueness,[130] we need to be careful about drawing semantic or constitutive conclusions from a problem that looks to be epistemological in nature. A predicate may be vague in the sense that we are unsure as to what determinate concept ought to count as providing the correct pattern of application (i.e. epistemologically vague), but that

[129] The Unlimited Epistemic Access Yields Unique Identification Thesis rules out *a priori* the idea that there could be metalinguistic facts that in principle outstrip our epistemic capacities. If there are such meaning-facts, they are assumed by Kripke's sceptic to be in principle identifiable (and uniquely so). As such, 'truth' is understood in an epistemically-constrained manner.

Epistemically-constrained notions of 'truth' are found within verificationist and Dummettian anti-realist trains of thought. The verificationist requires of synthetic sentences, in order to be meaningful, they need to be in principle verifiable. As such, one could not have a true synthetic sentence where there is no in principle means of verification. (Ayer 1986, p.48) 'There is a beetle in this box that disappears when you try to observe it' is, by the lights of verificationism, meaningless, and thus neither true nor false. The Dummettian anti-realist, by contrast, requires of synthetic sentences, in order to be either true or false, that they are in principle verifiable. (Dummett 1991, pp.4-8) As such, 'There is a beetle in this box that disappears when you try to observe it' is meaningful, but nonetheless neither true nor false.

On either of these approaches, one would be compelled to accept the Unlimited Epistemic Access Yields Unique Identification Thesis. One could not have a meaning-fact that in principle evaded our means of verification (which is understood along lines of unique identification).

[130] See §2.4.

consideration, without further argument, does not entitle us to claim that a predicate is vague at a semantic/constitutive level. We may be left unsure whether S means *addition* or *quaddition* by '+', and would remain unsure no matter how much we idealise the epistemological circumstances. However, it does not, strictly speaking, follow that there is no fact of the matter about what S means, since one would have to assume that the relevant metalinguistic fact is in principle knowable, which is far from uncontentious. As such, a meaning factualist could maintain that there are metalinguistic facts of the matter, but that at least some of them remain beyond our epistemic grasp.[131]

Nonetheless, the basic underdetermination strand of Kripke's argument can be detached from this meaning sceptical conclusion, which is precisely what I shall now do. In actuality, we can even broaden the argument. As we've seen, Kripke's argument is couched in terms of our inability, under ideal conditions, to uniquely identify 'S means that *p*' as a meaning-fact, i.e. as a *true* proposition. However, the same underdetermination point extends to considerations of whether it is *correct* (where 'correct' is understood as neutral between 'true' or 'warrantedly assertible')[132] that 'S means that *p*'.

We can see this by recalling Kripke's sceptical solution (as standardly understood). Roughly speaking, the sceptical solution is to maintain that it is warranted to assert that 'S means that *p*' if and only if S's usage of 'p' conforms to community usage. We saw that the sceptical solution is vulnerable to the criticism that it only works as an explanation of when certain metalinguistic claims are correct (warrantedly assertible)

[131] Space precludes a detailed examination of this important issue. However, it is worth at least noting a salient difficulty with this sort of response. Wright (Wright 1989a) argues that postulating metalinguistic truths that are, in principle, beyond our epistemic access renders the notion of a meaning-fact useless when it comes to explaining any given person's linguistic performances.

Typically, one thinks that there is a connection between meaning and understanding: we typically hold understanding a term is to know its meaning. In *understanding* something, a language-user displays what Wright refers to as a 'tracking epistemology', whereby our linguistic performances track these meaning-facts. The difficulty now is that, if the relevant meaning-fact eludes our grasp, even under ideal epistemic circumstances, it renders our intuitive conception of *understanding* a concept quite mysterious.

[132] As noted earlier, (see ft.82) I am utilising the notion of 'correctness' in what I take to be a neutral manner. It may be interpreted as meaning 'true' or along warranted assertibility lines. Kripke's sceptic's argument is directed explicitly against those seeking meaning-facts, where such meaning-facts are understood as truth-apt. My point here is that this basic underdetermination strand of his argument applies more generally. Whether it is warranted to assert that 'S means that *p*' is likewise underdetermined by any potentially available considerations. As such, this strand of argument can be couched in terms of 'correctness' rather than 'truth'.

if it is assumed that, at some stage, that it is a fact that some language-user's usage of their terms accords with community usage, which is incompatible with its status as a *sceptical* solution. As an attempt at constructive explanation of what constitutes the correctness (construed in terms of warranted assertibility) of a given metalinguistic claim, the sceptical solution is inherently unstable.

However, even if we waive concerns about the stability of the sceptical solution, there is still a basic underdetermination problem.[133] Let's assume that there is a fact about how the community has historically utilised some term 'p', and thus a fact about whether S's usage accords with community usage. If we construe 'correct' along warranted assertibility lines, and then understand 'warranted assertibility' roughly as a matter of the agreement between the speaker's usage and community usage, there is no reason why asserting 'S means *addition*' is warrantedly assertible, and yet it is not warranted to assert 'S means *quaddition*' (using some suitably altered version of *quaddition* such that it 'bends' after historical facts about community usage have been exhausted), since we have no way of uniquely identifying that it is *addition* that captures community usage.

Whether or not we construe 'correct' along the lines of 'true' or 'warrantedly assertible', then, we are unable to cite some consideration, given ideal epistemological conditions, that allows us to uniquely identify the relevant claim about meaning as correct. As such, the unique identification of which particular metalinguistic claim is correct is underdetermined by the behavioural, mental and dispositional facts. The underdetermination strand of the argument, then, is not restricted to considerations of whether 'S means that *p*' is true, but also runs for warranted assertibility.

At this point, it is useful to introduce a term of art: that of 'rational demonstrability'. To say that some claim is *rationally demonstrable* means that, given ideal epistemological circumstances, we would be able to uniquely identify that claim as correct. To say that 'S means that *p*' is rationally demonstrable means that, given ideal

[133] Not that Kripke denies this, of course. Kripke regards his solution as a *sceptical* solution precisely because he does not believe that there are any considerations available (under ideal epistemological conditions) that uniquely identifies S's meaning *addition* by '+', or, for that matter, the community's meaning *addition* by '+'. It would always be equally warranted to assert that 'S means *quaddition*' on some suitably altered version of *quaddition* such that it 'bends' at a point where the historical community usage of '+', which is finite, offers no guidance.

epistemological conditions, we would be able to uniquely identify 'S means that *p'* as being the correct metalinguistic claim. Using this notion, the underdetermination strand of Kripke's argument amounts to the claim that 'S means that *p'* is not rationally demonstrable: there are no considerations that could be available that would enable us to uniquely identify 'S means that *p*' as being correct, i.e. such that we can rule out 'S means that *q'* or 'S means that *r'*. No matter how much rational thought we invest in the matter, the relevant metalinguistic claim, 'S means that *p'*, cannot be uniquely identified as correct.

We then note some implications of this failure of rational demonstrability for metalinguistic claims. We note the connection between the metalinguistic claim 'S means that *p*' and conditions of correct use for the relevant term 'p'. If, when S uses the term 'p', 'S means that *p'* is correct, *p* gives the correct pattern of application for 'p'. Given this connection, from the lack of rational demonstrability of 'S means that *p'*, it follows that regarding *p* as supplying the correct pattern of application for 'p' is likewise not rationally demonstrable. Because one cannot rule out alternative, ersatz interpretations of S's use of the term, one cannot uniquely identify the set of correctness-conditions that the relevant metalinguistic claim identifies. Uniquely identifying that 'S means that *p*' is correct enables us to uniquely identify correctness-conditions for 'p'. Hence, if we cannot uniquely identify the relevant metalinguistic claim as correct, we cannot uniquely identify correctness-conditions for the application of that term. Since this failure is not merely a contingent failure but rather a failure that remains in place even when we idealise the epistemological circumstances, that *p* supplies the correct pattern of application for 'p' is not rationally demonstrable.

From this, it follows that, if 'S means that *p*' is not rationally demonstrable, neither is it rationally demonstrable when S has made some linguistic error. If we imagine a person who insists on using some term in a systematically mistaken way, such as a person who systematically miscarries when adding, whilst insisting that they are *adding*, we wish to say that they have made some sort of linguistic mistake, that they are mistaken about the meaning of '+'. However, given that it is not rationally demonstrable that *addition* supplies the correct pattern of application for '+' (as opposed to *quaddition*, *maddition* etc.), it is not rationally demonstrable that S is

mistaken in their application of '+'. Alternative interpretations in which S is not mistaken have not been ruled out.

We have three interrelated conclusions involving the failure of rational demonstrability: 1. That 'S means that *p*' is correct is not rationally demonstrable, 2. That *p* supplies the correct pattern of application for 'p' is not rationally demonstrable, and 3. That any insistent usage of 'p' systematically contrary to *p* involves a linguistic mistake is not rationally demonstrable. All three are epistemological conclusions, and do not of themselves entail any constitutive or semantic conclusions. I do not wish to claim that nothing constitutes correctness for metalinguistic claims of the form 'S means that *p*', or that nothing constitutes the pattern of correct (and thus mistaken) usage for 'p'. The claim is rather that there is an inherent limit to our abilities to provide definitive grounds for the relevant claims about meaning. There are no improvements one could possibly make to one's epistemological circumstances such that definitive (i.e. uniquely identifying) grounds for making the relevant claims may be found. From this inherent limitation to our justificatory powers, however, semantic or constitutive conclusions do not automatically follow.

With this caveat in mind, the underdetermination strand of Kripke's argument allows us to establish that there is an inherent limitation to our ability to demonstrate in a definitive manner, that one particular meaning claim is correct. Under ideal epistemological circumstances, the relevant mental, dispositional and behavioural facts underdetermine the relevant metalinguistic claim, and thus what counts as fixing the correct pattern of application for some term 'p', allowing us to identify linguistic mistakes. These matters are not rationally demonstrable.

4.3 – Wright on Basic/Non-Basic Rule-Following

Evidently, we still have a problem. Putting aside the question of whether there are facts of the matter expressible in the form 'S means that *p*', how are we to make sense of the fact that identifying a unique metalinguistic claim as correct, thus establishing a pattern of correct application (and thus being able to identify mistaken application) for 'p', is something that looks to be, in principle, beyond our grasp? In the following two sections, I shall elaborate Wright's Wittgenstein's view that, in order to make sense of our everyday practice, we are forced to accept the non-rationally demonstrable nature of metalinguistic claims. This amounts to regarding metalinguistic claims, and judgements about the correct application of some term 'p', as enjoying '*prima facie* legitimacy', where this '*prima facie* legitimacy' does not amount to any constructive philosophical claim, but is rather *descriptive* of our practice. I take this view, or something similar, to be a plausible way to understand Wittgenstein's response to the underlying rule-following problematic.[134]

In order to illustrate this particular 'Wittgensteinian' position, Wright's recent reflections found in 'Rule-Following Without Reasons' (Wright 2007, c.f. Wright 1989a) on how to understand Wittgenstein on rule-following are helpful. We wish to understand how to make sense of how it is we can legitimately regard S as meaning *p* in spite of the lack of rational demonstrability. In this section, I shall outline Wright's diagnosis of the central problem at the heart of the rule-following considerations (and thus the sceptical paradox), making clear his distinction between 'basic' and 'non-basic' cases of rule-following. In the next, I shall endorse Wright's reading of Wittgenstein on this matter, illustrating how one ought to respond to the rule-following considerations.

One central preoccupation found within Wittgenstein's rule-following considerations, according to Wright at least, is making sense of how it can be that we

[134] It almost goes without saying that any substantive exegetical claim concerning the later Wittgenstein's ambitions and intentions in *Philosophical Investigations* and elsewhere must be provisional in nature. In what follows, I rely upon Wright's recent interpretation of Wittgenstein, and make only the claim that it provides one plausible understanding of Wittgenstein's remarks, especially in the light of the notes found in notebooks from 1933 to 1944 and collected in *Remarks on the Foundations of Mathematics* (Wittgenstein 1989). That said, I shall at various points indicate obvious points of textual support for Wright's reading such that the reader can at least see why the reading has some intuitive plausibility.

may, with entitlement, continue to regard our practice of applying our terms in particular ways as conforming to, and answerable to, antecedent rules. As we've seen, the underdetermination strand of Kripke's argument reveals that it is not rationally demonstrable which particular rule should be regarded as providing the proper standard for correct application of some concept, although it often seems *highly obvious* to us. Given Wittgenstein's tendency to think that our practice is in order as it is,[135] he would want some way to preserve the idea that it is legitimate (in some sense) to apply rational categories to the business of applying some concept, such that we can fail to apply our concepts correctly (i.e. make a linguistic mistake), but not such that it requires *rational demonstrability*. Wittgenstein wants to maintain that it is quite appropriate, for instance, to treat my applications of '+' as answerable to the standard provided by the rule *addition*, even though it is not rationally demonstrable that it is *addition* that supplies the relevant standard.

Wright maintains that we really ought not find it terribly surprising that the considerations enlisted in the sceptical paradox lead to an 'intolerable' sceptical conclusion, since these considerations rely upon an understanding of what it is to follow a rule or apply some term that is inappropriate for all cases. Wright notes that at the heart of Wittgenstein's discussion of rule-following, which is then adapted by Kripke's sceptic, is what he calls the *'modus ponens* model' of rule-following. (Wright 2007, pp.490-1) Here, understanding the relevant concept is modelled as amounting to adopting a rule in the form of a general conditional. When this is combined with a relevant minor premise stating that the appropriate conditions are met, the conclusion (applying the concept) follows via *modus ponens*. Thus, to use Wright's own example of how one might apply the concept of *castling* in chess:

> *Rule*: If neither King nor one of its Rooks has moved in the course of the game so far, and if the squares between them are unoccupied, and if neither the King nor any of those squares is in check to an opposing piece, then one may Castle
> *Premise*: In this game neither my King nor this Rook have yet been moved, the squares between them are unoccupied, and . . .

[135] Wittgenstein's methodological presumption is that our practice is in order as it is. (Wittgenstein 2001, §124) Thus, if it turns out that our philosophical reflections fail to make sense of the practice, the presumption is that it is our philosophical theories that are mistaken, as opposed to the practice itself.

In this context, this scarcely seems controversial. Few philosophers *seriously* entertain the thought that all metalinguistic claims are illegitimately made. Nearly everyone here assumes that *something* must be amiss in the sceptical paradox, but it is not so easy to pinpoint exactly *what*.

Conclusion: I may castle now. (Wright 2007, p.490)

Evidently, the concept of *castling* in chess, much like the concept of *prime* in mathematics, depends on our utilising and understanding other concepts. They are 'non-basic' concepts, requiring our understanding other concepts in order to apply the relevant rule correctly. To apply the term 'prime' correctly, I need to understand how to apply the terms 'divide', 'integer', as well as understanding when a number is 'itself', 'the same', etc. To apply the term 'castling' correctly, I need to understand the relevant notions of a 'king', a 'move', a 'square', etc. I explain or justify my regarding some particular pattern of application for 'castling' as correct by invoking some *modus ponens* piece of reasoning involving other concepts.

Obviously, such a process of breaking down rules into component rules, as is required by the *modus ponens* model, cannot continue indefinitely: one either ends up with a regress or circularity.[136] Clearly, then, some concepts need to be such that they do not rely on inferences from further concepts in their definition, i.e. they need to be *basic*.[137] At this point, however, the *modus ponens* model of rule-following breaks down. Again, to use Wright's example, take the case of applying the term 'red':

Rule: If . . . x . . . , it is correct to predicate 'red' of x
Premise: . . . x . . .
Conclusion: It is correct to apply 'red' to x. (Wright 2007 p.495)

The difficulty here, Wright argues, is that it is hard to see what possibly could supply the relevant premise except simply the fact that x is red. According to the *modus ponens* model, the understanding of the relevant concept is supplied by the concept-rule, with the minor premise simply claiming that the conditions for its application are met. However, in applying this model to a basic case, the application of 'red', the model *presupposes* that one already understands the relevant concept *red*: the model

136 Wittgenstein indicates this problem in the oft-quoted passage that "It can be seen that there is a misunderstanding here from the mere fact that in the course of our argument we give one interpretation after another; as if each one contented us at least for a moment, until we thought of yet another standing behind it. What this shows is that there is a way of grasping a rule which is *not* an *interpretation*, but which is exhibited in what we call "obeying the rule" and "going against it" in actual cases... We ought to restrict the term "interpretation" to the substitution of one expression of the rule for another." (Wittgenstein 2001, §201)

137 "Basic cases – where rule-following is 'blind' – are cases where rule-following is *uninformed by anterior reason-giving judgement*." (Wright 2007, p.496)

"calls for a conceptual repertoire *anterior* to an understanding of any particular rule." (Wright 2007, p.496)

Wright's argument, then, runs that, in non-basic cases, it makes sense to regard the application of a concept as rule-governed on the *modus ponens* model, since one can trace the various inferential moves that underlie what counts as the correct application of the concept. In basic cases, however, the model breaks down. Here, understanding the relevant minor premise presupposes the very conceptual mastery that the rule, expressed in the form of a general conditional, was itself intended to explain. In order to know when to apply 'red' correctly, I need to know the relevant rule concerning 'red', but it seems I also need to already understand the very concept of 'red' in order to know when the rule applies. The *modus ponens* model is designed for situations where one makes inferences based on applying more basic concepts within the concept-rule and, as such, it is inappropriate for the basic case.

One can see that, in our discussion of the underdetermination strand of the sceptical paradox and the failure of rational demonstrability, we have implicitly relied upon this *modus ponens* model. The difficulty we found there was that it was always possible to come up with ersatz alternative rules where the application would not differ when adjoined to the circumstances encountered previously (expressed by the minor premise), but would differ when adjoined to novel circumstances. Underlying the sceptical paradox, then, was the presumption that, in order to explain the difference between correct and incorrect application for '+', we would need to supply a consideration that fitted this *modus ponens* model.

Wright contends that this move is suspect, and that we have to regard the basic cases (perhaps, such as '+') as *blind* rule-following, not governed by any inferential moves. By this, Wright does not intend some psychological or phenomenological notion of 'blindness' (Wright 2007, p.490), but 'blind' in the sense of non-inferential, and whereby attempting to find some underlying inferential structure to the application of the concept is inappropriate. As the basic case shows, not all cases of rule-following exhibit such an underlying rational structure:

> In any basic case, the lapse of the modus ponens model means that we should not think of knowledge of the requirements of the rule as a state

> which *rationally underlies* and enables competence, as knowledge of the rule for castling rationally underlies a chess player's successfully restricting the cases where she attempts to castle to situations where it is legal to do so. In basic cases there is no such underlying, rationalising knowledge enabling the competence. (Wright 2007, p.498)

Such a conclusion maps neatly onto our previous conclusion regarding the sceptical paradox: if one searches for considerations that uniquely identify *red* as what one means by 'red', one will come up empty-handed, no matter how much one idealises the epistemological circumstances. Using the distinction between basic/non-basic cases of rule-following, allows us to come to terms with why. In non-basic cases, we utilise other concepts in delineating rules for how one should apply a certain concept – e.g. a 'cube' is understood in terms of 'squares', and 'squares' are understood in terms of notions of 'line', 'length', 'equal' etc. On such occasions, it is appropriate to apply the *modus ponens* model. Ultimately, however, we arrive at basic cases of rule-following, where the model breaks down, and where it is inappropriate to apply the *modus ponens* model, to treat the concept as potentially exhibiting some underlying rational structure. Either way, at some point in our considerations we arrive at a point where we simply take some pattern of application as correct, without being able to uniquely identify it as such.

One might think that here we have the materials for a 'semantic primitivist' response to the sceptical paradox. For instance, we might maintain that *addition* is understood in terms of a primitive, basic concept of *counting*.[138] We might then block the follow-up made by Kripke's sceptic, i.e. that one may multiply interpret the notion of 'counting', by arguing that *counting* is not a case of rule-following on the *modus ponens* model, but is rather a primitive response.

However, this will clearly not satisfy the sceptic, for (at least) two reasons. Firstly, even if one removes the idea of *interpreting* basic concepts, that still provides no guarantee that the relevant primitive concept is *counting* as opposed to *quonting*. There is nothing logically amiss in the idea that a person might primitively 'count' in what *we* would consider to be an unusual manner. A person may just respond by *counting* as

[138] Note that we considered this thought in §3.3. I am myself unsure as to whether *addition* is basic or non-basic. One *might* define it in terms of *counting*, but one has something of a chicken-and-egg scenario here, since what is *counting* if not *adding one*? In what follows, therefore, I tend to regard *addition* as itself a basic case of rule-following.

we would ordinarily expect, but they might also just respond by *quounting*. At best, one would only 'rule out' *quonting* as an empirical hypothesis, not a logically possibility.

Secondly, we still have given no answer to the 'auxiliary problem' (Sprevak 2008): even if we grant that basic concepts have a universal, determinate pattern of application, how do we ensure that non-basic concepts, specified using rules involving basic concepts, will utilise the basic concepts in the way that we expect? Thus, granting that a person is primitively set up to just *count* without interpreting 'count', and that the pattern of application (barring performance errors) is exactly what we'd expect, one can still construct ersatz alternative interpretations of '+' using the primitive notion of 'counting' such that one is unable to uniquely identify the correct interpretation.

The point of introducing this basic/non-basic distinction is not to set up a semantic primitivist response to Kripke's sceptic, but rather that, like the underdetermination strand of Kripke's sceptic's argument, to illustrate the inherent limitation, i.e. one that remains even granted unlimited epistemic access, to our justificatory powers. If, once the question is raised as to whether I mean *addition* or *quaddition* by '+', I am unable to cite any consideration that would serve to show, in a definitive manner, that it is *addition* that I mean. That 'S means *addition*', and that *addition* supplies the pattern of correct application for '+' are matters that are not rationally demonstrable. Wright's point, brought into relief by the basic/non-basic distinction, is that we ought not be surprised by this, since the form of explanation we seek is on the *modus ponens* model. We always look for an underlying rational structure, a rule, that would serve to isolate *p* as the meaning of 'p'. At some point, however, such *modus ponens*-type explanations must bottom out in terms of *basic* concepts, such as *red*, where the most that can be said is that *this* is what we mean by 'p'. The way in which we would seek to uniquely identify some particular pattern of application as correct is via the *modus ponens* model, and we know in advance that this model cannot be appropriate in all cases.

4.4 – Wright's Wittgenstein and *Prima Facie* Legitimacy

The basic/non-basic distinction, then, allows us to make sense of why there is an inherent limit to our ability to justify one particular metalinguistic claim over another, or to explain what it is that makes it correct to ascribe *addition* to S in their usage of '+'. When it comes to forming a constructive philosophical explanation of what it might mean to have a commitment to following a rule, to applying a concept in accordance with its meaning, such that it makes sense to regard the application of a concept as correct or incorrect, the "only extant shot" (Wright 2007, p.497) is this *modus ponens* model, and we know in advance that this model must break down in certain cases of rule-following.

At this point, it is helpful to consider the following sort of response to the inherent limitation to our justificatory powers Wright offers on behalf of Wittgenstein,[139] in which Wittgenstein eschews any attempt at philosophical *explanation* of what makes *addition* the correct understanding of '+' in favour of a programme of *description*. Rather than regarding rules (thus fitting the *modus ponens* model) as *underlying* our linguistic competence (such that we can differentiate between correct and incorrect usage), Wright has it that Wittgenstein turns things on their head. Instead of regarding the constancy of our linguistic performances as *explainable* via the mechanism of rule-following (i.e. using the *modus ponens* model), it is rather that *because* our applications of various concepts as exhibited in our practice demonstrate universal (or near-universal) agreement that it is appropriate to think of our activity as describable using rules:

> It is, for epistemological purposes, a *basic* fact about us that ordinary forms of explanation and training do succeed in perpetuating practices of various kinds – that there is a shared uptake, a disposition to concur in novel

[139] It is unclear how far Wright would himself endorse what follows. Given that he maintains that "It is no good searching Wittgenstein's texts for a more concrete positive suggestion about the constitutive question [as to what constitutes the requirements of a rule]" (Wright 2007 p.488), it is likely that he regards what he takes to be Wittgenstein's view here as unduly quietist, where such 'quietism' amounts to the rejection of constructive philosophical explanation *tout court* (c.f. Wright 1994, p.202).

Whilst Wittgenstein might regard the question as to what entitles us to regard *red* as providing a standard against which applications of 'red' may be evaluated as a bad question, awaiting philosophical therapy as opposed to constructive philosophical response, Wright appears to demur from this reaction. Wright thinks that some alternative model of constructive explanation to that of the *modus ponens* model for why it is that *red* supplies the pattern of correct application of 'red' is available.

> judgements involving the concepts in question. The mythology of 'rules as rails' attempts an explanation of this fact. But the truth is the other way round: it is the basic disposition to agreement which sustains all rules and rule-governed institutions. The requirements which our rules impose upon us would not be violated if there were not this basic agreement; they would not so much as *exist*. (Wright 2007, p.487)

When it comes to the application of basic concepts, the primitive fact is that we *just do* agree in our judgements as to how to apply, say, '+' on any given occasion, novel or otherwise. An attempt at explanation here would involve the 'only extant shot' at explanation, applying the *modus ponens* model, and the problem here is that we know that this model is inappropriate for all cases of applying some concept. At this point, the key for Wright's Wittgenstein is to recognise certain responses to our concepts as primitive, where regarding a particular pattern of application as correct is a matter that does not stand in need of justification or explanation. To use a favoured metaphor of Wittgenstein, this involves recognising the ground as the ground. As Wittgenstein puts it in *Remarks on the Foundations of Mathematics*:

> The difficult thing here is not, to dig down to the ground; no, it is to recognize the ground that lies before us as the ground.
>
> For the ground keeps on giving us the illusory image of a greater depth, and when we seek to reach this, we keep on finding ourselves on the old level.
>
> Our disease is one of wanting to explain. (Wittgenstein 1989, VI §31)

The suggestion echoes a recurrent Wittgensteinian theme: that reasons come to an end.[140] Our tendency to regard some particular pattern of application of some concept as correct always outstrips, no matter how one improves the epistemological situation, our ability to give some philosophical explanation of or justification for why that pattern of application should be regarded as correct. Reasons come to an end, not for merely contingent reasons (we'd get bored, die of thirst, have a failure of imagination etc.), but because whatever reasons we come up with, they cannot

[140] The theme is no doubt more prominent in Wittgenstein's collection of notes *On Certainty*, but it can also be found in *Philosophical Investigations*:

> "To use a word without a justification does not mean to use it without right." (Wittgenstein 2001, §289)
>
> "If I have exhausted the justifications I have reached bedrock, and my spade is turned. Then I am inclined to say: "This is simply what I do." (Wittgenstein 2001, §217)
>
> "To be sure there is justification; but justification comes to an end." (Wittgenstein 2001a, §192)
>
> "Giving grounds,…justifying the evidence, comes to an end;--but the end is not certain propositions' striking us immediately as true, i.e. it is not a kind of *seeing* on our part; it is our *acting*, which lies at the bottom of the language-game." (Wittgenstein 2001, §204)

provide the sort of logical determination we seek, i.e. *rational demonstrability*.[141] In the end, our spade is turned, as we simply *act* in a particular way.

Nonetheless, it is a simple fact that we *do* agree over questions as to whether someone has 'added' or not. The principle lesson to be learnt here is that this agreement is not a *reasoned* one – it does not reflect some commonality of underlying rational structure, but is rather a matter of our happening to respond in a near-uniform manner. Our practice is such that we *just do* converge on one particular pattern of application when it comes to applying some concept, and thus converge in our judgements as to whether some person has used their terms correctly or made some linguistic mistake. We *just do* regard 1000, 1002, 1004 (say) as the correct application of "+2 to the last number, starting with 1000", but we would not be able to give the sort of definitive grounds required by *rational demonstrability*. Attempts at philosophical explanation of this constancy in our responses *over-rationalise* the matter.[142] We *just do* display this constancy in our responses, and it is because of this constancy that it makes sense to think of our activity as rule-governed.

Wright's Wittgenstein's suggestion, then, is that the reaction to the 'over-rationalisation' whereby we wish our linguistic activities to be explainable through invoking some underlying rational structure, is to remind think of cases where it is transparent that seeking such structure is out of place, and that *we just do* respond to concepts in a universal (or near-universal) patterned way. Here, what we should do is

---

[141] When Wright refers to Wittgenstein's position as 'rule-following without reasons', then, the sort of 'reasons' he has in mind are the sorts of things that fit into the *modus ponens* model, tied to *rational demonstrability*, whereby one uniquely identifies some metalinguistic claim as correct by invoking some underlying rational structure. Wright's point is that *reasons* in this sense will always give out, since there must be basic cases of rule-following whereby no further guidance can be given as to why one should apply 'red' (say) in *this* determinate way, as opposed to some other. Here there is no hope of *rationally demonstrating* that 'red' is only correctly applied if one applies it in that particular way, since any further attempt at rational demonstration will presuppose the very conceptual mastery we wished to have as the outcome of this underlying rational structure.

None of this precludes the possibility, of course, that one might enlist 'reasons' in the form of *persuasion*, such that one brings one's interlocutor around to seeing things from a particular perspective or ideology. This is to alter a person's usage in the first place, rather than citing some consideration that allows them to see that they are using their terms incorrectly.

[142] "*That* is why Wittgenstein's own response to his well-argued rejection of platonism is quietist. A non-quietist response would be called for only if platonism had given a bad answer to a good question. Then one would have to try to give a better answer. But the question was bad too. The real error in platonism is not the unsustainability of its sublimated conception of rule-facts, or the vulnerable epistemology that attends the sublimation. Rather the whole conception of rule-following to which it was a response was already an over-rationalisation – an implicit attempt to impose on rule-following everywhere a rational structure which can only engage the non-basic case." (Wright 2007, p.498)

describe how things go, without giving any further constructive philosophical explanation.[143]

We note that we *just do* have universal (or near-universal) responses to basic cases of rule-following, and that it is well that we do, since we presuppose this very constancy in our practice of giving definitions in cases of non-basic rule-following. For instance, it is because we are inclined to apply the notion of '/' in a near-universal manner that we may then invoke the notion in defining what we mean by a 'prime number', which then enables us to distinguish between correct and incorrect applications of this notion.

If one were to insist, however, that the fact that *prime* supplies the pattern of correct application for 'prime' is a matter that must be rationally demonstrable, we would need to consider at some point whether *division* supplied the correct interpretation of '/', and we search for some *modus ponens*-type explanation. The difficulty here is that we are faced with a basic case, and no explanation is forthcoming. It is only once we have *already accepted* that *divide* supplies the correct pattern of application for '/' that the business of using '/' in the definition of 'prime' gets off the ground. The process of giving grounds for regarding some pattern of application as correct comes to an end.

---

[143] These remarks connect to some of Wittgenstein's infamous and elusive comments on philosophical methodology:

> "Our disease is one of wanting to explain." (Wittgenstein 1989, VI §31)
>
> "Philosophy may in no way interfere with the actual use of language; it can in the end only describe it.
>
> For it cannot give it any foundation either.
>
> It leaves everything as it is." (Wittgenstein 2001, §124)
>
> "Philosophy simply puts everything before us, and neither explains nor deduces anything." (Wittgenstein 2001 §126)
>
> "Our mistake is to look for an explanation where we ought to look at what happens as a 'proto-phenomenon'. That is, where we ought to have said: this language-game is played." (ibid §654)

At various points, Wittgenstein insists that it is not the business of philosophy to 'explain' things, but of course this is not terribly transparent in the absence of some clarification as to what sort of 'explanation' Wittgenstein has in mind. Cases of basic rule-following may help in this regard, for here the relevant sort of 'explanation' is an explanation of why it is that we are *justified* in regarding a person's use of 'same' as correct. That is, we seek some theoretical description of what makes it the case that we are justified in regarding this application of 'same' as correct. However, when we come to basic cases of rule-following, we arrive at a point where attempted 'explanations' are not forthcoming. Wittgenstein's injunction, at this point, for whatever its merits, appears to be that one should attempt to overcome the "disease of wanting to explain".

If we are to continue thinking of our general practice of applying concepts in a way such that notions of correct and incorrect application apply, therefore, we need to accustom ourselves to thinking of the basic case of rule-following as *blind*, i.e. where we *just do* accept that the correct application of '/' is *divide*. Wright's Wittgenstein enjoins us to accept the fact that not only do we not have some explanation of why we are correct in taking '/' mean *divide*, but that it is inappropriate to think that some sort of explanation would be forthcoming, or is even appropriate.

To forestall confusion and emphasise the descriptive, non-explanatory nature of Wright's Wittgenstein's approach, it may help if I explicitly distinguish this understanding of Wittgenstein's purposes from a communitarian account. A communitarian would have it that community agreement over the application of 'p' *constitutes* the pattern of correct application of 'p'.[144] The correctness of S's usage of 'p' is assessed by comparing it with community usage. Consequently, there is no logical space for the supposed possibility that an entire linguistic community might come to mistakenly apply some particular concept, and no logical space for the idea that a 'language user', considered in isolation, can give any substantive content to the idea of their following a rule correctly.

On Wright's reading, Wittgenstein is not committed to a communitarian explanation of why we are justified in taking a person to task for using their words incorrectly because he does not believe that *any* sort of constructive explanation is appropriate. Take, for instance, Wittgenstein's following remark:

> A language game: to bring something *else*; to bring the *same*. Now, we can imagine how it is played. – But how can I explain it to anyone? I can give him this training. – But then how does he know what he is to bring next time as 'the same' – with what justice can I say he has brought the right thing or the wrong? – Of course, I know very well that in certain cases people would turn on me with signs of opposition.
>
> And does this mean e.g. that the definition of 'same' would be this: same is what all or most human beings with one voice take for the same? – Of course not.
>
> For of course I don't make use of the agreement of human beings to affirm identity. What criterion do you use, then? None at all.

[144] Depending on how one understands 'correctness' here, one can either have an account in which community agreement sets the conditions for warranted assertibility, which is the standard understanding of Kripke's sceptical paradox, or that community agreement sets correctness-conditions.

> To use the word without a justification does not mean to use it wrongfully. (Wittgenstein 1989, VII §40)

Wittgenstein here explicitly repudiates the suggestion that the what counts as setting the pattern of correct application for 'same' is "what all or most human beings with one voice take for the same." Wittgenstein rejects *any* form of explanation of why it is that we are correct in claiming that 'S means *same*': it is significant that Wittgenstein's response to the question as to what criterion he would use, i.e. that explains why S is correct in their use of same, i.e. as regarding one thing as the 'same' as another, is "none at all". A communitarian account would have it, then, that S is correct in their use of 'same' if they use the term in accordance with the linguistic community. Wittgenstein's response is markedly different: he avers that there is no relevant criterion here. He *blindly* takes that that S means *same* by 'same', without giving, or even potentially being able to give, any definitive grounds for that response.

Wright's Wittgenstein here is one that rejects the demand to supply some further rational consideration that underlies the basic case of rule-following. Wittgenstein rejects any sort of criteria-based explanation of why it is that we are entitled to make particular judgements regarding whether someone has applied their terms correctly. In basic cases, it is obvious that no explanation is forthcoming, since the 'only extant shot' at explanation is to apply the *modus ponens* model, and here any *modus ponens* explanation would presuppose the very conceptual mastery it was intended to explain. All we are left with are set of primitive responses, that we *just do* apply certain concepts in a universal (or near-universal) way, and no further rational considerations may be adduced that uniquely identify this particular pattern of application as correct. As a matter of *description*, Wittgenstein's approach should not be confused with criteria-based attempts at *explanation*, such as communitarianism.

Bearing in mind the descriptive, non-explanatory nature of Wright's Wittgenstein's approach, we can introduce a term of art, '*prima facie* legitimacy', in order to capture Wittgenstein's thought that "To use a word without a justification does not meant to use it without right." (Wittgenstein 2001, §289) To say that a claim is '*prima facie* legitimate' means that, as a matter of what we do, we accept, subject to future revision, the claim in the absence of rational demonstrability. That is, we accept it

without justificatory grounds, and without such grounds being available no matter how far we idealise our epistemological circumstances. Wittgenstein's point is that, at least as far things go as a matter of course, the question of *justifying* our taking S to mean *addition* by '+' does not arise, but that nonetheless it is not the case that we use our words 'without right'.

Clearly, it would be inconsistent with the interpretation here canvassed to ascribe to Wittgenstein the idea that some technical point is being made here, i.e. that 'using a word with right' points to some consideration that explains why it is correct to regard S as meaning *addition* by '+'. The rule-following considerations have already led us to see that no such explanation is forthcoming, no matter how far we idealise our epistemological circumstances. Regarding some metalinguistic claim as *prima facie* legitimate, then, does not amount to a constructive philosophical claim, because it involves our already accepting that there are no rational considerations available that could explain why some meaning attribution, as opposed to another, is in order.[145] We note the way that S uses '+', say, and take them to mean *addition* until we have reason to think otherwise. In thinking of our own universal (or near-universal) use of '+', we take it to mean *addition* until we have reason to think otherwise. The question of *justifying* these things does not, as a matter of course, arise.[146]

To better appreciate the connection between this point about taking metalinguistic claims as *prima facie* legitimate and the rule-following considerations, consider

---

[145] As such, the notion of '*prima facie* legitimacy' contrasts with Kripke's sceptical solution (as standardly understood). In the sceptical solution, some constructive explanation is given for what constitutes the correctness (warranted assertibility) of some claim of the form 'S means that *p*'. To say that a claim of the form 'S means that *p*' is *prima facie* legitimate, however, just means that it is a claim that we just do accept (subject to future revision) without *justificatory grounds,* and without such grounds being in principle available.

[146] As is implied by the locution '*prima facie*', it is not the case that some metalinguistic claim is immune to rational reconsideration. It may well emerge, for instance, that a person we take to be genuinely *adding* may, on further examination, turn out to be systematically mistaken in their application of '+', whilst insisting that they are indeed *adding*. That is, it may emerge that their responses diverge systematically from our own, which again we take to have *prima facie* legitimacy regarding the meaning of '+'.

It may even be that considerations emerge that entail that our collective application of some concept was in some way confused, vague, or in error in some way: as I argue in §4.6, this gives us a way of making sense of changes in our understanding of 'parallel', for instance.

The claim is, as it were, that our 'default position' is to treat our activity of making claims about meaning, and thus regarding some pattern of application of a concept as correct, as in order. But this should not be seen as either a *substantive* claim, such that one can cite considerations that *constitute* its legitimacy, or a claim that is invulnerable to revision upon further consideration.

Wittgenstein's remarks that are generally held to signal the end of the rule-following discussion:

> "So you are saying that human agreement decides what is true and what is false?" – It is what human beings say that is true and false; and they agree in the *language* they use. That is not agreement in opinions but in form of life.
>
> If language is to be a means of communication there must be agreement not only in definitions but also (queer as this may sound) in judgments. This seems to abolish logic, but does not do so. – It is one thing to describe methods of measurement, and another to obtain and state results of measurement. But what we call "measuring" is partly determined by a certain constancy in results of measurement. (Wittgenstein 2001, §§241-2)

Here we have an insistence on the primitive inclinations to respond in a universal (or near-universal) manner: the 'form of life'. We *just do* accept that *this* is what counts as measuring; it has *prima facie* legitimacy. It is not rationally demonstrable that it is *this* pattern of application that should count as the correct way to 'measure' as opposed to some ersatz alternative. We cannot point to some underlying rational structure to our 'measuring' that would allow us to uniquely identify *measuring* as opposed to some alternative. To say that it is *prima facie* legitimate to regard S as *measuring*, or that *measuring* supplies the pattern of correct use for 'measuring', is not to say that to claim such things is *justified*, but rather to insist that, as a description of how we proceed, the question of justification does not, at least *prima facie,* arise. We treat such claims as in order until some consideration comes to light that requires of us to question things.

The deeper, philosophical point, however, is that it is well that we take this inclination to measure in *this* particular way as *prima facie* legitimate, since we would not be able to make sense of areas of activity that relied upon this inclination to measure in this universal (or near-universal) way. If we didn't have these universal (or near-universal) responses to the requirements of 'measuring', and if we didn't just take these inclinations as *prima facie* legitimate, it would be impossible to justify or maintain as factual the claim, for instance, that stick A is *longer* than stick B. We will simply take it that someone who 'measures' in an ersatz way, e.g. as we would unless measuring green objects over 700m long (in which case they conclude that it is 5m long), is not applying the notion of 'measuring' correctly, although it is not rationally demonstrable that they are mistaken. These judgements, although not displaying an

underlying rational structure, nonetheless enjoy *prima facie* legitimacy within our practice.

Wright's Wittgenstein's point, encapsulated in Wright's phrase that "it is the basic disposition to agreement which sustains all rules and rule-governed institutions."(Wright 2007, p.487), is that our linguistic practice relies upon our accepting the *prima facie* legitimacy of just taking it that *this* counts as 'measuring', and that *this* counts as 'the same', etc. If we were to insist that such claims had to be rationally demonstrable, i.e. that they had to be grounded by some underlying structure, some considerations that could (in principle) serve to uniquely identify the relevant claim as correct, we would face a requirement that could never be satisfied under any circumstances. Our response to the failure of rational demonstrability should not be that metalinguistic claims, or distinctions between correct/mistaken usage of 'p', are illegitimate since they cannot be definitively grounded, but rather should be to recognise that it is well that we treat our inclinations in regarding *this* pattern of application of a given rule or concept as correct as *prima facie* legitimate. From here, we can then utilise such concepts in further linguistic practices such as definition, or appealing to such concepts in order to demonstrate that someone has misunderstood, etc.

4.5 – Rational Demonstrability and Explication

It remains to connect the observations of this and the last chapter with the task of providing an (ideally exact, sufficiently similar) explicatum. In this section, I shall outline the connection: namely, that the underdetermination strand of Kripke's argument creates a potential difficulty, but one that often turns out to be in practice benign, for matters of explication.

One relatively obvious way in which we might connect Kripke's sceptical challenge and matters of explication is by regarding the supplying of an ideally exact explicatum as providing a straight solution to Kripke's sceptical paradox. That is, where an explicatum, in the form of an algorithm that effectively decides its own pattern of correct application, would allow us to uniquely identify the relevant meaning-fact, thus answering Kripke's sceptic. We would be able to ascertain, for instance, if 'S means *p*' by S's indicating that some explicatum P is what they mean, and P effectively decides what counts as correct usage, rendering P determinate and incompatible with any ersatz alternatives. In effect, it would be a form of the 'algorithm response'[147] to Kripke's sceptical challenge.

The sceptic challenges us to cite, having been granted unlimited epistemic access, some consideration that uniquely identifies 'S means that *p*', as opposed to any number of ersatz metalinguistic claims that are equally consistent with our previous usage yet would diverge extensionally in new cases. The algorithm response here runs that the relevant meaning-fact may be identified through S stating that they regard what counts as the correct application of 'p' as identical to the output of some mechanical procedure. Say that I set up an *2-input-addition* machine that takes any two integers and effectively decides their sum (i.e. after a finite number of steps), then that machine has a determinate output given any appropriate input. The rule is, as Kripke puts it, "*embodied* in a machine that computes the relevant function." (Kripke 1982, p.33) The machine is set up such that, barring mechanical breakdown, it will always *2-input-add* the two numbers. The output of the machine, given inputs appropriate to its domain, is thus clearly determinate, since it is effectively decided by

[147] The notion of an 'algorithm response' is taken from (Kusch 2006, p.129).

the mechanism. Under such circumstances, then, I may simply treat the output of the machine as constituting the correct application of '+' for 2 integers.

Various arguments are employed by Kripke's sceptic at this point (Kripke 1982, pp.34-5) to discount the algorithm response: firstly, providing the inputs and outputs to the machine is something that may itself be variously carried out (one might, for instance, understand what it means to perform '+' on two numbers along the lines of, "If either of the numbers involved are more than 57, enter the numbers '2' and '3' into the machine; otherwise, enter the numbers you have"). Secondly, one cannot simply take it that the output of the machine *constitutes* correct application of the relevant rule, since we need to allow logical space for the possibility of mechanical breakdown. Thirdly, he takes it that, like dispositions, the actual performance of the machine is finite, whereas the extension of the correct application of '+' is infinite in nature, and thus the output of the machine cannot constitute correct usage. One may or may not regard such arguments as successful. On the basis of considerations over the last chapter, it seems unlikely that the third response, insofar as it relies on the supposed 'finiteness' of the machine's potential responses, will carry much weight, but it may well be that the other two responses have some force.

In any case, it seems relatively clear that the algorithm response, like many other responses, merely relocates the problem. The sceptic has no reason to, as it were, take S's word for it that the *2-input-addition* machine really does capture what S means by '+' (for two integers). Even if one grants that the machine is a genuine *2-input-addition* machine, ignoring the possibility of mechanical breakdown, and taking it that it is unlimited in its potential application, there is no reason available why we should regard *this* machine, rather than its ersatz counterpart, the *2-input-quaddtion* machine, as genuinely embodying what S means by '+'. Unless one can give some consideration that allows us to uniquely identify one particular machine as embodying the relevant meaning, having a physical instantiation of a particular determinate algorithm is useless in answering Kripke's sceptic. Furthermore, if S could already guarantee that it was the *2-input-addition* machine that genuinely embodied what S meant by '+', then we would have already uniquely identified the meaning-fact, and one would thus have already answered Kripke's sceptic. One needs to have already solved the sceptical paradox in order to get the algorithm reply

off the ground. One needs to already know that this machine, set up as it is to give a particular patterned causal response, *already* supplies the correct understanding of the relevant concept, before the machine can function as constitutive of the correct application of the relevant concept. This, however, is precisely what the sceptic calls into question.

To connect this to matters of explication, we might think of explication as a form of S stating that some mechanical procedure embodies what they mean. Suppose the sceptic, for instance, challenges S to cite some consideration that would enable us to uniquely identify *prime number* as providing the correct pattern of application for 'prime number'. At this point, S might insist that 'prime number' is explicated through our purported *explicatum* for 'prime number', that of PRIME:[148]

1. Let $i = 2$
2. Let $k = n / i$
3. If $k \in \mathbb{N}$, STOP, $n \notin$ PRIME
4. If $i > n/2$, STOP, $n \in$ PRIME
5. $i = i + 1$
6. GOTO 2

As such, S will insist that the mechanical procedure PRIME renders the notion of 'prime' such that the correctness of any particular application of 'prime' is effectively decided by the procedure. As such, it would be incorrect to insist, say, that 893710352 is prime, since it diverges from the value returned by the procedure.

However, Kripke's sceptic may simply enquire at this point why we ought to regard this procedure PRIME as explicating what S mean by 'prime'. After all, why is it not the case that the mechanical procedure QUIME explicates 'prime'?:

1. Let $i = 2$
2. Let $k = n / i$
3. If $n = 44^{666}$, STOP, $n \in$ QUIME
4. If $k \in \mathbb{N}$, STOP, $n \notin$ QUIME
5. If $i > n / 2$, then $n \in$ QUIME
6. $i = i + 1$
7. GOTO 2

[148] See §2.4.

Here, the procedure will return an identical extension to that of PRIME, until it 'bends' with numbers greater or equal to 44^{666}, at which point it decides that all such numbers are QUIME. As such, it will be (or could be suitably altered to be) entirely compatible with S's historical usage of 'prime number'. The question is, then: why accept PRIME as uniquely identifying the relevant explicatum, when QUIME is equally compatible? The basic underdetermination problem is simply relocated. All the available evidence is compatible with both PRIME and QUIME being used to satisfactorily explicate 'prime'. Both are ideally exact, and sufficiently similar to the explicandum. Thus, putting forward the explicatum PRIME does not solve the sceptical paradox: one cannot point to a relevant fact of the matter that guarantees that PRIME, as opposed to some ersatz alternative QUIME, gives us what counts as the correct extension of 'prime'. It is only within the context of an explicatum *having already been adopted* that it may serve as the supplying the correct pattern of application for the explicandum.

Seeing how the algorithm reply fails, however, enables us to better see the connection between explication and the basic underdetermination component of Kripke's sceptic's argument, and then appreciate the potential difficulty for any explicative project. The connection is this: matters of explication suffer from the same basic underdetermination problem as claims of the form 'S means that *p*'. Just as the claim that 'S means that *p*' is underdetermined by the available facts, even under ideal epistemological circumstances, the claim that PRIME better explicates 'prime number' than QUIME, say, is likewise underdetermined. Any given proposed explicatum is such that it there is an inherent limit, a limit that remains even granted unlimited epistemic access, to justify our adopting it to the exclusion of logically-distinct alternatives. It is always possible that some ersatz, equally compatible, alternative explicatum is available. In terms of rational demonstrability, it is not rationally demonstrable that PRIME explicates 'prime number'. We'd need to be able to uniquely identify PRIME as explicating 'prime', and this would require ruling out QUIME and other ersatz explicata. However, we are not able to do so; that any given explicatum supplies the pattern of correct application for the relevant explicandum is a matter that is not rationally demonstrable.

If we recall from chapter 2, we assigned a commitment to an explicative project to the standard reflective equilibrium theorist seeking a coherence theory of justification in ethics. This commitment was that it must be in principle possible to design some ideally exact explicatum, REFLECTIVE EQUILIBRIUM, that would effectively decide any given case in a manner that is sufficiently similar to our intuitive judgements about 'being justified'. If we are able (i.e. given enough time, resources, and theoretical wherewithal) to do so, we could use this ideally exact REFLECTIVE EQUILIBRIUM procedure to effectively decide whether it is correct that the agent holds their moral conception in REFLECTIVE EQUILIBRIUM. From there, assuming that the explicatum is satisfactory (most prominently, that it is sufficiently similar to our intuitive judgements about 'being justified' in holding some moral conception), we could utilise this explicatum (to the exclusion of alternatives) such that we would be able to effectively decide whether the agent *is justified* in holding their moral conception.

Motivating this explicative project of supplying of an (ideally exact, sufficiently similar) explicatum, is the idea that it renders an area of discourse transparently disagreement-entailing-error objective. The aim with an ideally exact PRIME explicatum, for instance, is to have some effective procedure such that we know in advance that there is a standard of correctness for PRIME for any appropriate input. Not only is it *a priori* that disagreements about whether some number is PRIME must involve some sort of error on the part of at least of one the parties to the disagreement, but *that* it is disagreement-entailing-error objective is transparent to us. Now, if we accept that PRIME explicates 'prime number' (to the exclusion of alternatives), we then have a method which makes the application of '… is a prime number' transparently disagreement-entailing-error objective. We would be in a position to, in principle at least, identify which of the parties to a disagreement about whether *x* is a prime number are in error.

The failure of rational demonstrability, however, creates a potential difficulty. Clearly, we cannot accept *both* PRIME and QUIME as explications of 'prime number', i.e. where we *replace* the question of 'is it a prime number?' with 'is it PRIME/QUIME?' If we regarded either alternative as on a par, we'd immediately lose the advantage of having an (ideally exact) explicatum, namely that it rendered the question transparently disagreement-entailing-error objective. We would not be able to decide, for instance,

whether 44^{666} is a prime number or not, since it would not be clear which explicatum we ought to be using. Accepting PRIME as explicating 'prime number' reveals that 44^{666} is prime; accepting QUIME as explicating 'prime number' reveals that 44^{666} is not prime. However, we have no means, no matter how far we idealise our epistemological circumstances, of demonstrating that it is PRIME that explicates the notion of a 'prime number'.

The problem here is that, *having already accepted* PRIME (to the exclusion of alternatives) as explicating the explicandum 'prime number', the explicatum serves to transparently secure the disagreement-entailing-error objectivity of the usage of 'prime number'. However, the acceptance of PRIME as explicating 'prime number' itself is not itself a transparently disagreement-entailing-error objective matter. If there were a disagreement about whether one should regard PRIME or QUIME as explicating 'prime number', there is not any consideration available that would allow us to rule out QUIME (or myriad other potential explicata) as explicating 'prime number'. If there is an error on the part of either (or both) of the parties to the disagreement, it is not one that we can, even in principle, locate. We could not cite, no matter how much we idealise the epistemological circumstances, some consideration that uniquely identifies PRIME as the best way to explicate 'prime number'. The upshot is that explication can only play a role in rendering an area of discourse transparently disagreement-entailing-error objective within the context of it having already been accepted to the exclusion of other logical alternatives.

It is important to note here that it is not an acceptable response to this problematic to say that some explication may *in fact* be correct[149] (and thus that the relevant area of discourse *is* disagreement-entailing-error objective because some party to the disagreement has made some sort of error), but not in a rationally demonstrable manner. This sort of response would be to mirror the reaction canvassed in §4.2 to the Unlimited Epistemic Access Yields Unique Identification Thesis, in which one insists upon a distinction between the relevant epistemological and constitutive concerns. Here one would maintain that, whilst there may be an epistemological

[149] As we've already seen (see ft.76), for Carnap, this suggestion would represent a category mistake, since he holds, for reasons connected to the principle of tolerance, that explicata are *proposals*, not *claims*, and can thus only be regarded as 'satisfactory' or not. Nonetheless, it is worth considering why insisting on the correctness of a particular explicatum is not an acceptable response in any case.

lacuna in the sense that it is not rationally demonstrable that PRIME explicates the notion of 'prime number', it is nonetheless the case that it does so, unknowably.

The problem here is that this sort of claim is available to any party to a disagreement. At any point in this hypothetical disagreement concerning how to explicate 'prime number', a party may simply insist that their favoured explication *just is* correct, but that its correctness is not rationally demonstrable. They may well be right for all we know, but the difficulty is that making this move does not help settle the disagreement in any way. The proponent of *any* explicatum may insist on the correctness of the explicatum at any point, but it does not furnish the disputant with a consideration that allows them to see their supposed error. Explicative projects aim to improve the objectivity of an area of discourse by furnishing us with procedures that allow us, at least in principle, to see what counts as the correct pattern of application for the explicatum (and thereby the explicandum). In order to do so, it needs to be possible that we might be in a position to actually ascertain that some particular explicatum is the one that we ought to adopt (to the exclusion of others). Merely insisting on the correctness of some explicatum is a move any disputant might make, and does not advance matters. It represents, to borrow a metaphor favoured by Wittgenstein (Wittgenstein 2001, §271), an idle wheel.

We have a general problem for explication, one that mirrors the problem identified by the underdetermination strand of Kripke's sceptic's argument. In order to render areas of discourse transparently disagreement-entailing-error objective, we need to be able to justify the adoption of some explicatum to the exclusion of alternatives. For an explicatum to make an area of discourse transparently disagreement-entailing-error objective, the selection of that explicatum would seemingly need to be itself a transparently disagreement-entailing-error objective matter. The underdetermination strand of Kripke's sceptic's argument, however, shows that this cannot be the case. We cannot rationally demonstrate that PRIME explicates 'prime number' to the exclusion of QUIME or myriad other ersatz alternatives.

4.6 – *Prima Facie* Legitimate Explication

It is not my intention here to pursue a sceptical line with regards to explication. Intuitively, this would be a bizarre conclusion, given the obvious appropriateness and usefulness of providing explicata for everyday mathematical concepts, for instance. Nobody, for instance, would question the thought that we can provide an ideally exact, sufficiently similar explication of 'prime number' to the exclusion of other alternatives, such that whether *x* is a prime number is a transparently disagreement-entailing-error objective matter. In what follows, therefore, my intention is to make room for the idea that an ideally exact explication may be satisfactorily adopted (to the exclusion of alternatives) in spite of the absence of rational demonstration.

The dialectical situation is thus as follows: I wish to preserve the thought that some explications are perfectly in order, i.e. they are satisfactory. There is nothing amiss in the thought that PRIME explicates 'prime number', and that it is appropriate (in some sense) to adopt PRIME (and not one of the myriad logically-distinct alternatives) as the criterion of correctness against which ascriptions of 'prime' may be compared, thus rendering the area of discourse transparently disagreement-entailing-error objective. However, the upshot of the underdetermination strand of Kripke's sceptical challenge is that a level of underdetermination is inevitable no matter how one improves one's epistemic situation, and from this conclusion it follows that there is an inherent limit to our ability to justifying adopting one particular explicatum to the exclusion of others. The aim, then, is to give some indication of why it is appropriate to regard PRIME as explicating 'prime number' in spite of the fact we cannot rule out innumerable ersatz alternatives.

It is here that I believe Wright's Wittgenstein and the notion of '*prima facie* legitimacy' are helpful. Let's take an example of a case which is even more basic than that of explicating 'prime number'. Imagine explicating '+ (for 2 integers)' using a 2-INPUT-ADDITION algorithm. Suppose we produce a mechanical device, say, a Turing Machine, that takes 2 integers as inputs, and returns their sum, and stipulate that the correct application of 2-INPUT-ADDITION *just is* whatever value the device returns. What makes it the case that it is appropriate to regard this explicatum as explicating '+ (for two integers)', and thereby treat the output of the computation as constitutive

of correct application for '+ (for two integers)' in the relevant domain? It cannot be that we have some consideration that uniquely identifies the output of 2-INPUT-ADDITION as supplying the correct pattern application of '+ (for two integers)', since a mechanical device performing 2-INPUT-QUADDITION will be equally compatible with all the available facts regarding the application of '+'.

Our basic responses to the requirements of '+ (for two integers)' are universal, or near-universal here: faced with the inputs {[0,0], [0,1], [0,2], [0,3], [1,0], [1,1]. [1,2], [1,3], [2,0], [2,1], [2,2], [2,3], [3,0], [3,1]. [3,2], [3,3]}, we return the answers {0, 1, 2, 3, 1, 2, 3, 4, 2, 3, 4, 5, 3, 4, 5, 6}, and these responses coincide extensionally with the outputs of the 2-INPUT-ADDITION procedure. The difficulty is, of course, that these responses also coincide with that of the 2-INPUT-QUADDITION procedure. At this point, we seem to be forced to conclude that taking 2-INPUT-ADDITION as explicating '+' is illegitimate, since we have failed to supply some consideration that rules out 2-INPUT-QUADDITION. This would entail that explicata would be illegitimate across the board. We could not allow, for instance, that PRIME (to the exclusion of competitors) explicates 'prime', since PRIME relies on our being able to supply unique explicata for 'divide', 'natural number' etc., and these can all be multiply interpreted even once all possible facts are in.

Alternatively, we can treat our response to regard 2-INPUT-ADDITION as explicating '+' in this restricted domain as *prima facie* legitimate, that is, where the matter is not rationally demonstrable, but nonetheless we *just do* accept it, without reasons, until there is reason to revise things. We do not have a constitutive explanation available for why the explication is superior to alternatives, but we may nonetheless accept it *with right*. The machine coincides extensionally with our basic inclinations to respond to '+', and we treat these inclinations as having *prima facie* legitimacy; similarly, our judgement that the procedure explicates the informal concept is one that has *prima facie* legitimacy. We do not have reasons available that justify this reaction, but it is clear that, as a matter of what how we proceed, it is clear that *we just do* accept that some procedure explicates '+' here.

Evidently, we could say exactly the same thing about taking 2-INPUT-QUADDITION (or a suitable variant thereof) as explicating '+ (for two integers)'. The 2-INPUT-

QUADDITION procedure would have returned identical outputs on all given occasions thus far, and there would be no basis for ruling it out as explicating '+'. If we were to accept 2-INPUT-QUADDITION, then, this would likewise be on a *prima facie* legitimate basis. The difference between the two here is that cases may emerge on future occasions such that the outputs of 2-INPUT-ADDITION and 2-INPUT-QUADDITION diverge, and we know in advance that our basic inclination would be to adopt 2-INPUT-ADDITION, once again with *prima facie* legitimacy.[150] In the case of two mechanical procedures, we could simply investigate the mechanism and, upon discovering that the constitution of the two machines is subtly different, we will have cause to investigate as to if and when their outputs will diverge. Then, discovering that their outputs diverge when one of the inputs is numbers over 57, we then need to ask ourselves which output accords with our inclinations for applying '+'. Thus, nothing precludes the possibility that what we here and now, with *prima facie* legitimacy, regard as explicating '+' may eventually need to be rejected.

Imagine, for instance, that we were attempting to explicate the notion of 'parallel' (in terms of 'parallel lines') before the advent of non-Euclidean geometry. Suppose part of the explicatum procedure involved testing whether lines ever intersected, and taking it that extended lines that did not intersect were PARALLEL. Clearly, such a procedure today would need to be restricted to the domain of Euclidean geometry: it would be illegitimate as an explication in non-Euclidean cases, since the extension of PARALLEL would radically diverge from that of 'parallel' in non-Euclidean cases. Thus, *considerations emerged* why the original explicatum would need to be rejected, or rather revised. Until such reasons emerged, however, mathematicians would have been within their rights to regard this hypothetical explication as satisfactory. It is in this sense that regarding explications as satisfactory is a matter of *prima facie* legitimacy: one notes that the explication coincides extensionally with one's inclinations to apply the explicatum until one has reasons to think otherwise, and one thereby regards, with right, that explication as satisfactory.

It is worth stressing, once more, that this notion of '*prima facie* legitimacy' does not amount to a constructive philosophical claim. It ought not be confused, for instance, with the claim that we are justified in adopting some explicatum (to the exclusion of

[150] This connects with my earlier point (see ft.146) of regarding the legitimacy of meaning attributions as *prima facie* legitimate.

others) if and only if that explicatum coincides extensionally with the sum total of community usage of the explicandum. The Wittgensteinian point is that it is essential to the entire business of supplying definitions (which is, in essence, what explication is about) that the question of *justification* does not initially arise, at least in some cases. Given the inherent limitation to our justificatory powers revealed via the underdetermination strand of Kripke's sceptic's argument, it is not possible to give that sort of constructive explanation of what justifies us in picking out one particular explicatum as the one we ought to adopt. Rather, we *just do* accept that, say, PRIME (to the exclusion of alternatives) explicates 'prime number', and it is only when difficulties arise that the question of choosing between explicata arises. Until then, we simply note that our inclinations with regards to the application of 'prime number' are universal (or near-universal), and we pick out a procedure that tallies with those inclinations, and then accept it as supplying the criterion of correctness for applications of '… is a prime number', rendering it a disagreement-entailing-error objective matter.

When it comes to transforming an area of discourse into a transparently disagreement-entailing-error objective matter, then, the Wittgensteinian approach (under the interpretation canvassed) is to recognise that whether an explicatum is accepted (to the exclusion of others) is a matter that we do not, and indeed cannot, definitively ground. Much like the acceptance of some metalinguistic claim, we cannot, no matter how far we idealise the epistemological circumstances, cite some consideration that would enable us to uniquely identify PRIME as explicating 'prime number'. Within the context of that explicatum having already been accepted, however, disagreements about whether some number is prime are rendered transparently disagreement-entailing-error objective. Our acceptance of the explicatum in the first place, however, is something that is undertaken with *prima facie* legitimacy: we are within our rights to do so until considerations emerge that require of us to rethink the matter. We have thus preserved the intuition that it is perfectly appropriate to think of 'prime number' as explicated by PRIME, but without giving some constructive suggestion of what it is that makes it appropriate.

All of which may give the impression that, as it were, providing satisfactory explicata is a cakewalk. However, this is a misleading impression created by the fact that we

have been dealing with paradigmatic, 'hard'[151] cases where it seems intuitively obvious that explication is perfectly appropriate. These cases are where we have inclinations to respond to '+' in ways that are unchangeable, culturally non-specific, etc. That is, we are dealing with universal (or near-universal) inclinations with respect to '+', which renders the business of supplying an effective procedure that explicates '+' none-too-difficult. We are dealing with cases where the extensional coincidence of explicatum and explicandum is almost perfect.[152] As Wittgenstein puts it, "Disputes do not break out (among mathematicians, say) over the question whether a rule has been obeyed or not. People don't come to blows over it, for example." (Wittgenstein 2001, §240) Because our responses in certain cases display this uniformity, it is thereby relatively easy (assuming we have the theoretical wherewithal) to design mechanical procedures that explicate these everyday concepts. It is because our initial inclinations in response to '+' display the requisite uniformity that it is very easy to identify, with *prima facie* legitimacy, some explication to take its stead.

Suppose, however, that we were faced with a situation where it emerged that our inclinations did not display such uniformity. Continuing with the adding example, suppose it genuinely was the case that some sizable percentage of the population were inclined to *quadd*: when asked to add the following pairs {[0,55], [0,56], [0,57], [0,58], [1,55], [1,56]. [1,57], [1,58], [2,55], [2,56], [2,57], [2,58], [3,55], [3,56]. [3,57], [3,58]}, this section of the linguistic community respond with {55, 56, 5, 5, 56, 57, 5, 5, 57, 58, 5, 5, 58, 59, 5, 5}, insisting that what they were doing was *adding*. As such, they would insist that we explicate '+' using some QUADDITION explicatum. We would be faced with a disagreement about which explicatum we ought to adopt (to the exclusion of others), stemming from a basic difference in our brute responses to a basic case of rule-following.

In this situation, I take it, it would not be *prima facie* legitimate to assume that what '+' requires of us is that we return the answers {55, 56, 57, 58, 56, 57, 58, 59, 57, 58, 59, 60, 58, 59, 60, 61}, since that is simply to point to one (partisan) set of responses.

---

[151] 'Hard' as opposed to soft, that is; not 'hard' as opposed to easy.

[152] Here I do not necessarily wish to claim that this extensional coincidence need be perfect. If one remembers the example of explicating our informal notion of 'life-form', for instance, it seems that the relevant formal explication may well have points where it diverges from our informal responses, such as in the case of whether viruses are genuine life-forms (say). However, it is probably safe to say that any explication needs to command an *overwhelming* level of extensional coincidence.

Here it would be a idle gesture to insist that the quadders were committing a linguistic error in their performances of '+', since we cannot give them some consideration that would demonstrate to them that they had made some sort of error. Clearly, if we are thinking of our own reactions *here and now*, it is *prima facie* legitimate for *us* to conclude that these quadders are mistaken. But that is simply because *we are not in this situation*. We operate under conditions where there are no quadders, and our additive practices are uniform or near-uniform (no 'bends' have yet emerged). If we were genuinely in the situation where people responded to basic cases of rule-following in divergent ways, it would not be *prima facie* legitimate to regard one particular pattern as thereby demonstrating the 'correct' application of the rule. It would thus not be *prima facie* legitimate to regard ADDITION as a satisfactory explication of '+'.

We may not be able, then, to give a constructive account of when it is *prima facie* legitimate to make some explicative claim, but we are in a position to recognise when, as a matter of how we proceed, we would not regard it as legitimate. Supplying some explicatum (to the exclusion of others) relies upon a bedrock of universal (or near-universal) inclinations with regards to applying the explicandum. We may explicate '/', taking it as *prima facie* legitimate to do so, because our inclinations when it comes to understanding what '/' requires of us are universal (or near-universal). Accepting some explicatum to the exclusion of others takes place against a background of agreement in our *blind* inclinations.

It is worth emphasising at this point that the genuinely *problematic* cases, the cases in which we would not take any given explicatum to be *prima facie* legitimate, are not where our inclinations to respond to some concept are *vague*,[153] but rather where they are subject to a significant level of *dispute*. There is, for instance, no difficulty in explicating the notion of a 'heap'. There will be cases, e.g. around 25 grains of sand, where any given language-user is unsure whether the notion of 'heap' is correctly applied or not, but there is not a *disagreement* as such. Any given language-user would tend to withhold judgement on such cases. We may explicate the notion of 'heap' here with some (ideally exact) explicatum HEAP taking the form of an effective

[153] Here I mean 'vague' in its epistemological sense (see §2.4), such that there are borderline cases where any given competent language-user would be unsure whether the concept is correctly applied or not. This may be because 'heap' is vague at a semantic level, or because we lack the 'fineness of discrimination' to decide between competing determinate interpretations of 'heap'.

procedure, because all we require is that the explicatum is *sufficiently similar* to the explicandum. Noting that there will be various explicata available ($\text{HEAP}_1 =_{\text{def}}$ any collection of small, discrete clumps numbering 25 or over; $\text{HEAP}_2 =_{\text{def}}$ any collection of small, discrete clumps numbering 26 or over; etc.), one may choose between the alternatives with impunity. That HEAP_1 ought to be adopted to the exclusion of alternatives is not rationally demonstrable, but this limitation to our justificatory powers is benign. Any of these alternatives would agree with our inclinations, would be regarded as equally similar to the explicandum, and we might adopt any of them (to the exclusion of alternatives) with *prima facie* legitimacy.

The genuinely *problematic* cases occur where we are faced with an explicandum that is not *vague* but *disputed*,[154] where the dispute either concerns how to apply some *basic* concept, or stems from a further dispute concerning the correct application of a component, *basic* concept.[155] In such cases, we are faced with a situation in which one person responds *blindly* to the basic concept in *this* way; another responds *blindly* in *that* way. In such situations where our *blind* inclinations to apply the explicandum are such that they differ from person to person, or from group to group, the extension of the explicandum will be disputed. However, since we then know that selecting from these disputed alternatives is a matter that is not rationally demonstrable owing to the inappropriateness of the *modus ponens* model, we have no means available, no matter how far we idealise our epistemological circumstances, of uniquely identifying one particular explicatum as explicating the relevant term.

In the example where a significant proportion of the population systematically and insistently applies '+' utilising *quaddition*, there are cases (e.g. 59 + 93) where quadders and adders have different *blind* inclinations, and dispute what counts as the correct usage of '+'. What we have here is a disagreement stemming from differences in inclination ('form of life') as to how one *blindly* applies '+'. As such, we have no rational considerations available, or even possibly available under better

[154] Again, we need to interpret this as an epistemological claim, such that there are disputed cases where different competent language-users disagree about whether the concept is correctly applied in such cases.

[155] The point here is to rule out cases of ambiguity, such as the question as to whether a 'prime number' applies to natural numbers or integers. A dispute here is fairly easily resolved by distinguishing between the two different senses of the same term. It does not pose any significant problems. Strictly speaking, we should not be faced with such a case, since the explicandum should have already been *clarified* such that it is clear what sense of the term we wish to explicate, but is pays to be explicit about what sort of explicative disputes are genuinely problematic.

epistemological circumstances, such that one could establish that one or more of the parties have made an error in their use of '+'. All one has here are differential inclinations to proceed in some determinate way. In such a scenario, it would not be *prima facie* legitimate to favour explicating '+' with ADDITION to the exclusion of QUADDITION.

The essential difference, then, between cases where we may explicate a notion with *prima facie* legitimacy, and those cases where it would not be regarded as legitimate, is the extent to which our blind inclinations to relevant basic concepts are universal (or near-universal). If there is already a sufficient degree of divergence, stemming from differences in *blind* inclination, in how the explicandum is applied, then any given proposed explicatum will not be *prima facie* legitimate. If we are tempted to think that some further rational consideration might be available to resolve any disagreement about which explicatum to adopt (to the exclusion of others), we need to note that there is an inherent limitation to our justificatory powers here, a limitation revealed by the underdetermination strand of Kripke's sceptical challenge. There must be *basic* cases where we accept, without underlying reasons but nonetheless with *prima facie* legitimacy, that *this* is how one applies the concept. It is only where our inclinations are universal (or near-universal) that we may supply some explicatum that will enjoy *prima facie* legitimacy. In those cases where our blind inclinations to respond to a basic concept diverge to a significant extent, no particular explicatum that utilises such a basic concept will enjoy *prima facie* legitimacy. In such a situation, therefore, it will not be possible to supply some explicatum to the exclusion of others such that it would serve as a basis for rendering an area of discourse transparently disagreement-entailing-error objective.

Chapter 5 – Explicating 'Being Justified' in Ethics

5.1 – Overview

My discussion of Kripke's sceptical paradox/the rule-following considerations has led us to a point at which we recognise the importance of accepting, with *prima facie* legitimacy, metalinguistic claims and, by extension, certain explications. From here, we recognise that there may well be genuinely problematic cases in which we would not accept some explicatum, to the exclusion of alternatives, with *prima facie* legitimacy. In this chapter, I shall argue that we are faced with exactly such a case when we attempt to explicate 'being justified' (in ethics) along the lines of reflective equilibrium *qua* coherence theory. That is, even once we follow the standard reflective equilibrium approach in thinking that the way to explicate 'being justified' is along coherence lines, we would not be able to select, with *prima facie* legitimacy, a REFLECTIVE EQUILIBRIUM explicatum to the exclusion of alternatives.

The difficulty, as we shall see, arises from problematic cases of disagreement, 'opaque disagreements' (a term explained in §5.3), being sufficiently prevalent within the moral sphere. The prevalence of 'opaque disagreements' in the moral sphere entails that, once one makes one's REFLECTIVE EQUILIBRIUM explicatum ideally exact (as required by the explicative project assigned to the standard approach), one will be faced with many, mutually-exclusive alternatives as to how to model the coherence of an agent's moral conception, each tracking different sets of inclinations agents have in cases of first-order moral disagreement. The prevalence of these problematic cases of first-order moral disagreement lead to problematic cases of second-order epistemological disagreement.

Evidently, much depends here on characterising why certain sorts of disagreement are problematic, and sufficiently prevalent within the moral sphere. To explain why they are problematic, I invoke considerations similar to McDowell's, and, more recently, Crary's, viewing of certain disagreements as stemming from differences in 'sensitivity' (a term I explain in §5.2) to relevant basic concepts, rather than from differences in the beliefs of the parties to the disagreement. Such cases are where different competent language-users have different, *blind* conceptual inclinations in

response to the same situation. That is, to use a moral example, one party to the disagreement will see, without having any underlying reasons, some action undertaken in a situation *as cruel*, whereas another party will not see, again without underlying reasons, the situation as one where the concept applies. The disputants have different sensitivities: they *blindly*, i.e. without invoking any further reasons, apply the relevant concept differentially.

I then discuss two examples, drawn from within contemporary Western culture, of moral disagreement that are persuasively analysed along such lines. In these examples, disagreements about the moral permissibility of eating meat, and the production/consumption of pornographic material, I argue that what is at stake, at least partially, is not disagreement about the facts, or disagreement about moral principles, but rather differences in how different people respond to the same situation by *blindly* invoking different concepts, having different 'sensitivities'. I then give some indication as to why we'd expect these problematic sorts of disagreements to be prevalent within the moral sphere.

I then spell out the unwelcome consequence for reflective equilibrium *qua* coherence theory, which is that, given the prevalence of such moral 'opaque disagreements', it is highly unlikely that there could be in principle, an ideally-exact REFLECTIVE EQUILIBRIUM explicatum available that we would accept with *prima facie* legitimacy. This then entails that, at least under circumstances where these cases of moral disagreement stemming from differences in 'sensitivity' are still prevalent, the standard approach to reflective equilibrium, regarding reflective equilibrium of primary importance *qua* coherence theory of justification in ethics, is misguided.

5.2 – Sensitivity

To begin with, then, we need to clarify the notion of a 'sensitivity', which I shall do by drawing on the work of McDowell and Crary, and relating it to the considerations of the previous two chapters. As we shall see, the way these authors utilise the notion allows us to connect the idea with Wright's use of *blind* rule-following, applying a *basic* concept, where any attempt to explain or justify what constitutes correct application (i.e. using *modus ponens*-style reasoning) would be inappropriate.

Starting with McDowell,[156] then, the notion of a 'sensitivity' is drawn from lessons garnered from Wittgenstein's rule-following considerations. McDowell wishes to argue that having some virtuous trait such as *kindness* requires an agent to develop a 'sensitivity' to situations such that one is able to discern that a particular sort of reaction is appropriate, and that this 'sensitivity' itself counts as a form of (moral) knowledge. However, he notes the difficulty that, for many philosophers, knowledge must be the sort of thing that can be *codified* in a propositional form, and here it is unclear how one might codify *kindness,* and what would constitute a *kind* reaction to a given scenario. McDowell here invokes the Aristotelian (Aristotle pp.64-5 [§1.3]) thought that what counts as a virtue, such as *kindness*, may be uncodifiable:

> To an unprejudiced eye it should seem quite implausible that any reasonably adult moral outlook admits of any such codification. As Aristotle consistently says, the best generalisations about how one should behave hold only for the most part. If one attempted to reduce one's conception of what virtue requires to a set of rules, then, however subtle and thoughtful one was in drawing up the code, cases would inevitably turn up in which a mechanical application of the rules would strike one as wrong – and not necessarily because one has changed one's mind; rather one's mind on the matter was not susceptible of capture in any universal formula. (McDowell 1979, pp.57-8)

The conjunction of the demand for codifiability and the Aristotelian thought that a mechanical formulation of a virtue such as *kindness* would result in the, for McDowell, perverse view that being *kind* does not amount to demonstrating a form of (moral) knowledge. In order to make room for the idea that it represents a form of knowledge that one understands what would count as exhibiting the virtue of

[156] The relevant papers here are 'Virtue and Reason' (McDowell 1979) and 'Non-Cognitivism and Rule-Following' (McDowell 1981). However, see also 'Two Sorts of Naturalism' (McDowell 1996).

kindness, then, McDowell rejects the idea that knowledge in general need be the sort of thing that is codifiable. For McDowell, it is merely a philosophical 'prejudice' that, in order for some moral judgement such as 'Eric was kind to Edwina in letting her borrow his car' to count as potentially an item of moral knowledge, we require that the moral judgement be potentially represented as the outcome of applying some set of underlying, codifiable principles (i.e. a moral theory or moral conception) to a situation.

Indeed, McDowell maintains that it is even misleading to think of cases where we explicitly follow some 'hard' mathematical rule as *fully* codifiable.[157] As we've seen, *no amount* of specifying rules in a *modus ponens* fashion can ever *uniquely* determine some particular pattern of application for some concept. Where we have codification, then, it still depends on our 'whirl of organism' (McDowell 1979, p.62, citing Cavell 2002, p.52) for the sort of consistency and near-universality we display in our particular judgements. Thus, even in the case of applying '+', say, there is still a dependence on our universal (or near-universal) practice of applying '+' in a certain way, our 'form of life', although the dependence is less obvious. At this point, we ought to "give up the idea that philosophical thought, about the sorts of practice in question, should be undertaken at some external standpoint, outside our immersion in our familiar forms of life." (McDowell 1979, p.63)

The notion of a 'sensitivity' is invoked by McDowell, then, in order to make sense of how it could be that a virtuous agent knows something about how to act virtuously that is not codifiable, i.e. that is not able to be represented as the outcome of some *modus ponens* reasoning involving some underlying moral principle. It is merely a prejudice, McDowell avers, to think such things could be understood from *outside* the particular 'form of life', the particular practices, within which this 'sensitivity' occurs. Even in the 'hardest' cases of mathematical rule-following, there is still some dependence on the 'whirl of organism' that is an underlying practice. One needs to

[157] That is, McDowell isn't denying that we might enlist *some* codifiable rules in order to help clarify the demands of a particular concept. He is, however, maintaining that the practice of supplying rules itself depends on an underlying constancy of our 'whirl of organism'. Invoking rules is only of assistance where there is a constancy of underlying practice in how we utilise such rules, e.g. in recognising cases as being the 'same', and constancy in our application of basic concepts where the invocation of further rules would be inappropriate.

be *inside* the 'whirl of organism' in order to be able to master the relevant concepts and understand how to *just apply* them correctly.

The relevant difference between knowing what is *kind* and knowing how to apply '+', for McDowell, is that the dependence on having mastered an underlying practice is *more obvious* in the former case. Here, virtuous action requires of us that we perceive situations in the right way, i.e. with the appropriate 'sensitivity'. One is unable to appreciate the 'sensitivity' here except from *within* the appropriate form of life: "Occasion by occasion, one knows what to do, if one does, not by applying universal principles but by being a certain kind of person: one who sees situations in a certain distinctive way." (McDowell 1979, p.73)

The notion of a 'sensitivity' is, for McDowell, a kind of *perception*, and is connected with the concept of *salience*: "A conception of how to live shows itself, when more than one concern might issue in action, in one's seeing, or being able to brought to see, one fact rather than another as salient." (McDowell 1979, pp.68-9) Having the respective sensitivity is the ability to perceive a particular situation as already requiring of us certain sorts of actions. Because I have the relevant sort of sensitivity, I can perceive in what someone says a particular sort of prejudice, say, and as requiring of me that I challenge their views. I perceive the implicit prejudice as a *salient* feature of the situation, and I react accordingly. That I reacted to this feature, and not the feature that what they said was in English (say), is on account of my having a sensitivity to prejudicial forms of language and behaviour. This, however, is not the sort of thing that it is at all well-represented through codifiable principles. Accepting that there may be aspects to moral thought that involve something worth describing as a 'sensitivity', then, ensures that there may be more to moral thinking than may be captured through delineating a particular moral conception using the familiar apparatus of moral theory.

In Crary's *Beyond Moral Judgment*, Crary likewise utilises the notion of a 'sensitivity' or a 'sensibility' to highlight areas of our moral thinking[158] that are not captured through

[158] I perhaps ought to note that it is part of Crary's position that *moral* thought is not some distinctive area of a person's thought. Her resultant position is highly holistic, such that nothing short of a person's entire conceptual repertoire is constitutive of their 'moral' outlook: "*Proper respect for challenges of moral conversation involves concern with nothing less than individuals' entire personalities, the whole complicated weaves of their lives.*" (Crary 2007a, p.45)

the delineation of a moral conception, i.e. making explicit an agent's moral commitments. The notion of a 'sensitivity' is understood by Crary through a pragmatic account of language, whereby conceptual mastery is itself regarded as a practical ability. Essentially, the thought is that responding to, and learning to apply, a particular concept should be regarded as a practical ability that is not amenable to being set out in an ideally abstract (i.e. codified) fashion.[159] Mastering a concept is then a matter of learning a particular *practice* of applying that concept in particular situations, rather than learning some set of instructions or rules that guide its application.[160]

Like McDowell, such 'sensitivities' are understood by Crary as a matter of being able to pick out salient features in a situation. One comes to a situation, and through having a particular complex of sensitivities, one understands that situation in terms of particular concepts as opposed to others. Through having mastered a set of practical abilities in one's sensitivities, one perceives a situation in terms of, say, *harassment*, as opposed to, say, good-natured fun. Crary stresses that being sensitive to cases of harassment is a matter that requires that we "apply concepts that... are such that considerations for applying them are accessible only in terms of particular evaluative perspectives." (Crary 2007a, p.176) Similarly, in reading Tolstoy's *The Death of Ivan Ilyich*, the kind of moral knowledge we garner from the work is an enhancement of our sensitivity to the *emptiness* of Ivan Ilyich's life. "We imbibe the moral message of the story to the extent that the ways it calls on us to respond enable us to make sense of certain features of Ivan's life that he has somehow forgotten or obscured." (Crary 2007, p.157) Although Crary herself does not stress perceptual metaphors here, it does not appear to be inappropriate to regard the notion of a 'sensitivity' as a kind of *seeing-as*: through having a complex set of sensitivities, we come to understand a situation through a particular evaluative perspective, utilising some set of concepts to the exclusion of others.

---

159 "There is no question of surveying the projection of a concept in an ideally abstract manner." (Crary 2007a, p.41)

160 Crary, like McDowell, connects this point to Wittgenstein's rule-following considerations: "We might say that here Wittgenstein is inviting us to see that our concepts, far from being instruments for picking out contents that are independently available (as the image of an ideally rigid rail suggests), are resources for thinking about aspects of the world to which our eyes are only open insofar as we develop certain practical sensitivities." (Crary 2007a, p.25)

There are obvious commonalities between Crary's and McDowell's usage of 'sensitivity', then. Both utilise the notion to point to a lacuna in what may be made explicit or codifiable. I may recognise, for instance, that one ought not harass people, or that certain occasions require kindness, but being able to recognise an action *as kind*, or *as harassment* requires a practical skill that is not codifiable, having a sensitivity. However, an area in which Crary's focus perhaps shifts from that of McDowell's[161] lies in her stressing the thought that sensitivities may differ both over one person's life, and between people:

> The practical ability that, according to the pragmatic account of language I am considering, conceptual mastery takes the form of a sensitivity that, in the case of any given concept, may vary not only within a single individual over time (i.e., because it may develop as she first arrives at mastery of the concept and as her master then deepens) but also among distinct individuals (i.e., because, in the case of each individual, it may develop in different ways and to different extents as the mastery of the concept deepens). (Crary 2007a, p.42)

The upshot of such differences in sensitivity between people is that we see distinctive and novel cases of moral disagreement that are not at all well-understood by contemporary moral theorising that seeks to represent moral disagreement explicitly through the delineation of the disputants' moral conception:

> Within contemporary moral philosophy, it is generally assumed that moral differences take the form either of disagreements about whether to apply a given moral concept or of disagreements about whether some moral concept (or set of such concepts) is one we ought to operate with in the first place. In contrast, within the context of this more expansive catalogue of forms of moral thought, moral differences may exist between people who inherit and develop different ways of thinking and talking about the world even where there is no question of a disagreement of either of these types... Now we can speak of such differences whenever we confront significant discrepancies between the sensibilities that inform individuals' ways of thinking and speaking or, in other words, significant discrepancies between the images,

[161] McDowell's ethical 'naturalism' (McDowell 1996) is such that it takes ethical situations as, in some sense, already 'enchanted' by considerations of value. (McDowell 1996, pp.181-2) Our normal ways of thinking about the world require that we have already adopted particular conceptual sensitivities to perceiving action as *kind, brave* etc. Such normative judgements are not inferred from value-neutral descriptions.

The difficulty, as McDowell is less-than-willing to advertise, is the evident variability in the kind of sensitivities we might have. It may be a salient consideration in cases of active euthanasia, for instance, whether such an act represents an act of *kindness* or not. It is opaque here whether it would be correct to apply the concept *kind* on such occasions.

> composed by these sensibilities, of what the world is like. (Crary 2007a, pp.44-5)

Just as McDowell seeks to maintain that maintaining some sensitivity still counts as *knowledge* in some sense, Crary attempts to incorporate the notion of a 'sensitivity' within a 'wider' conception of rationality and objectivity.[162] The obvious philosophical concern for someone sympathetic to the idea of a 'sensitivity' is that it renders *subjective* any moral judgements that are reliant on having a certain sensitivity. Consider disagreement-entailing-error objectivity for instance. Imagine person A, who through having a certain sensitivity, thinks Ivan Ilyich's life has been a good one because it was a happy, contented one; person B, on the other hand, has a different sensitivity such that they perceive the emptiness of Ivan Ilyich's 'comme il faut' existence. Here, it is not at all clear how one could establish whether either party had committed some sort of 'error' here. If it is a disagreement-entailing-error objective matter as to whether Ivan Ilyich's life has been empty, it would appear to be *beyond our possible grasp* in the absence of any means of deciding which person's sensitivity is correct.

To this worry concerning the objectivity of sensitivity-derived judgements, Crary often speaks of conceptual 'maturation' and 'deepening' where we open ourselves up to new uses of a particular concept. Obvious examples include a child being able to identify a dog when confronted with one, but then deepening this conceptual sensitivity when, for instance, they are able to differentiate between wolves and dogs, or when they are able to appreciate the metaphoric use of 'lapdog' in sentences like 'Tony Blair is George W. Bush's lapdog'. In a similar vein, one could regard one's becoming aware of the emptiness of Ivan Ilyich's life as a *deepening* of one's 'sensitivities'. In such a situation, it might well make sense for us lucky few with the deepened 'sensitivity' to regard one's former view, or the views of others, as

---

[162] Crary's wider notion of 'objectivity' is one in which one recognises that what she calls the 'abstraction requirement', the idea that 'objective' discourse is discourse that is wholly independent of our sensitivities and subjective responses, as inherently confused. (Crary 2007a, pp.25-9)

As I noted when introducing the notion of 'disagreement-entailing-error objectivity' (see ft.80), notions of 'objectivity' are various. As such, Crary's discussion of 'objectivity' seems to revolve around rejecting a view of objectivity as requiring 'stepping outside' one's practice, or adopting the 'view from nowhere', but there are many notions of objectivity that do not, at least on the face of it, appear to mandate such a view. It is not clear, for instance, that the notion of 'disagreement-entailing-error objectivity' upon which I have relied need rely on the 'abstraction requirement'. On the face of it, there does not appear to be any particular difficulty in thinking of, say, 'This book is blue' as conceptually dependent on our having particular sensitivities to the concepts *book, blue*, etc., and yet still reflecting a disagreement-entailing-error objective matter.

impoverished to some extent. One might then regard one's former self, and other people with the impoverished sensitivity, 'mistaken' in their understanding of Tolstoy's work, and matters of life and death more generally.

In order to make room, then, for the thought that a sensitivity-derived moral judgement may still count as knowledge, or be objective in the sense of disagreement-entailing-error, some room is made for the thought that sensitivities are not themselves beyond evaluation. Disagreements stemming from differences in sensitivity need not be *faultless* such that no party is in 'error' on the matter in question. One person's sensitivity may be more developed, more mature, deeper, than the other's, such that a person with an impoverished sensitivity will be in error should a disagreement arise.

In the following sections, I shall adopt this notion of a 'sensitivity', but I am less optimistic about the prospects of establishing that these cases of moral disagreement are genuinely disagreement-entailing-error objective. If one chooses one's examples carefully, one can select cases where a change in sensitivity is worth describing as 'deepening' of sensitivity. The child learning further uses for 'dog', cases of sexual harassment, the case of Ivan Ilyich, etc. all seem to be well-chosen in this respect. However, one can also select cases where it is completely unclear, at least to my mind, which complex of conceptual sensitivities ought to count as the 'deeper', more 'mature', response to a situation, and here it strikes me that it is far more plausible to maintain that these case reflect faultless disagreement. Here I shall utilise two examples, one from Diamond's influential paper, 'Eating Meat and Eating People', (Diamond 1979) and another drawn from variant attitudes from self-ascribed 'feminists' towards pornography. Both cases look to me to be persuasively analysed as disagreements stemming from differences in sensitivity, as opposed to disagreements stemming from differences in either relevant empirical beliefs or one's underlying moral conception. However, it seems to me to be unclear how one could possibly decide whose sensitivity is 'deeper' or more 'mature', certainly in such a way that anything worth calling an 'error' is involved.

As I shall use the term, a 'sensitivity' refers to the pattern of conceptual application that a particular person *blindly* employs when faced with different situations. Faced

with a certain situation, one's sensitivities will be such that one will *just respond* with some concept as opposed to another. So, one might have a particular sensitivity with regards to the concept of *cruelty*, say, such that one will respond to some situated action φ in taking it that the concept applies, to κ in taking it that it does not, λ in taking it that it applies, etc. Having a set of sensitivities in general involves being inclined to give a set of conceptually distinctive responses to situations. That is, whereby one is inclined such that some set of concepts are seen as (correctly) applicable, and where one is inclined such alternative concepts are seen as not (correctly) applicable. Having a sensitivity to a particular concept involves there being a distinctive set of scenarios in which one is inclined to apply that particular concept, and there being a set where one is not inclined to apply that concept.

The use of the term 'inclined' here should not be taken to imply that some dispositionalist account is on offer. I do not intend the notion of an 'inclination' to represent any attempt at *explaining* why it is correct, for instance, that 'φ is cruel', or that 'φ is good-natured fun' is mistaken. A relevant constructive explanation here would seek out some sort of underlying rational structure to the respective judgements such that it emerges as the conclusion from some *modus ponens*-style reasoning. Like McDowell and Crary, I take it that sensitivities cannot be usefully delineated utilising further *rules* or *principles*. Having a particular sensitivity with regards to some concept is not a matter of employing some underlying rational structure that may be revealed through rationally reconstructing some piece of *modus ponens* reasoning. It is rather something we need to take as *blind*, whereby no considerations are available, even in principle, such that one could isolate the respective judgement as correct to the exclusion of alternatives.

It is helpful to think of sensitivities in terms of *basic* concepts, where it is obvious that the *modus ponens* model breaks down, and we find ourselves *just responding* in patterned, universal (or near-universal) ways. Ultimately, we find ourselves at a point where any further minor premises involving a description of the situation such that a concept-rule can be invoked must necessarily utilise the very concept the concept-rule was intended to capture. At this point, all we can do is take the particular application we are inclined to make regarding some basic concept as reflecting our having a particular conceptual sensitivity. Doing so involves taking the set of judgements we

make regarding whether some concept applies in a particular situation at face value, as uninformed by any underlying rational structure.

In summary, like McDowell and Crary, I utilise the notion of a 'sensitivity' precisely to indicate an explicitly non-explanatory account, whereby no further considerations are considered appropriate on the question as to why some particular concept applies in a certain situation. Here, all we can say is that *we just do* respond in this particular way, and that having this particular pattern of responses amounts to having a sensitivity with regards to that concept. Invoking the notion of a 'sensitivity', then, points to cases where the *modus ponens* model breaks down, and where it seems hopeless to search for any further underlying rational structure. However, unlike McDowell and Crary, who wish to incorporate sensitivities within what they regard as a wider, less prejudicial view of objectivity and rationality, I shall continue to utilise the notion of 'disagreement-entailing-error objectivity'.

5.3 – Example 1: Eating Meat

In this section and the next, I shall outline two cases of moral disagreement that are persuasively viewed as deriving from differential sensitivities to situations, as opposed to differences in moral conception or relevant empirical claims. I have deliberately selected examples where the issue at stake is disputed strongly within contemporary Western culture. As such, I take it that they do not represent isolated or exotic examples, which emerges as a relevant consideration in §5.5 where I maintain that it is persuasive that cases such as these are sufficiently prevalent within the moral sphere.

The first case involves attempting to locate the source of moral disagreement between vegetarians and meat-eaters, and here I take my lead from Cora Diamond's influential paper 'Eating Meat and Eating People' (Diamond 1979). Diamond notes that typical philosophical attempts to clarify the issues at stake between ethical vegetarians[163] and meat-eaters take the form of clarifying the underlying rational structure of the parties to this moral disagreement.[164] We reconstruct a hypothetical chain of *modus ponens* reasoning whereby the conflicting moral judgements are revealed to be the outcome of differences in relevant empirical beliefs or underlying moral principles. In the case of disagreement concerning the moral permissibility of eating meat, this might well look something like the following:

[163] That is, vegetarians who regard eating meat as morally impermissible, as opposed to those who avoid eating meat for aesthetic or dietary reasons.
[164] The classic example is Peter Singer's 'All Animals are Equal' (Singer 1973), in which he maintains that it is pure 'speciesism' to count the suffering of animals as lesser to that of human suffering. (Singer 1973, pp.222ff.)

Person A:	Person B:
1. Animals bred for the production of meat are sentient beings capable of experiencing pain. 2. Animals bred for the production of meat experience more pain over the course of their existence than they would otherwise. (e.g. they live in cramped, stressful conditions, have underlying medical conditions as a direct result of their artificial selection, are subjected to painful courses of hormone therapy, etc.) 3. Eating meat entails that someone purchases the meat. 4. Purchasing meat entails that more animals will be bred for the production of meat than would otherwise be the case. 5. Eating meat is an easily avoidable piece of behaviour. 6. It is morally impermissible to cause easily-avoidable pain to sentient beings.	1. Animals bred for the production of meat are sentient beings capable of experiencing pain. 2. Animals bred for the production of meat experience more pain over the course of their existence than they would otherwise. (e.g. they live in cramped, stressful conditions, have underlying medical conditions as a direct result of their artificial selection, are subjected to painful courses of hormone therapy, etc.) 3. Eating meat entails that someone purchases the meat. 4. Purchasing meat entails that more animals will be bred for the production of meat than would otherwise be the case. 5. Eating meat is an avoidable piece of behaviour, but, for some people at least, involves giving up a significant source of pleasure, and is thus not *easily*-avoidable. 6. It is morally impermissible to cause easily-avoidable pain to sentient beings. Causing pain to sentient beings that is not easily-avoidable, however, is permissible.
C_1 Eating meat is morally impermissible.	C_2 Eating meat is morally permissible, for some people at least.

Obviously, there are various lacunae in this pattern of reasoning that one would obviously wish to extirpate, but one can see the general idea. The respective conclusions are seen as the outcome of a process of either actual or hypothetical reasoning, such that the respective moral judgements is justified and/or explained as the outcome of following through some piece of *modus ponens*-type reasoning. The moral judgement is conceived of as displaying an underlying rational structure. One need not actually claim that a person has literally thought it through in such terms; rather what is intended is that we explain or justify the judgement, and the person's overall moral conception, through moving to a higher level of abstraction. Faced with a disagreement in judgements, we move to a higher level of abstraction, such that the judgement is explained as resulting from some underlying rational structure. As Rawls puts it in *Political Liberalism*, "We should be prepared to find that the deeper

the conflict, the higher the level of abstraction to which we must ascend to get a clear and uncluttered view of its roots." (Rawls 1996, p.46) The task, then, is one of rendering the moral disagreement in a perspicuous form, such that its underlying 'roots' can be seen clearly. Let's call this thought, so clearly expressed by Rawls, the 'theoretical presumption'. The 'theoretical presumption' is to presume[165] that any given disagreement in moral judgement can be explained by revealing some difference in underlying rational structure, whereby either some empirical issue is at stake, or there is some difference in the set of moral principles held by the parties involved.

Diamond, however, is frustrated by this way of setting it out. It seems to misdiagnose the nature of the disagreement between ethical vegetarians and meat-eaters, which often does not stem from differences in moral principles or empirical information at all. To this end, she notes an oft-overlooked point: many, if not all, ethical vegetarians do not wish to eat dead wild animals when they've died accidentally.[166] At this point, the question of whether the 'meat industry' causes suffering is not the salient issue, although of course it is one that remains important.[167] Rather, the ethical vegetarian's view can be better appreciated, avers Diamond, by thinking about a comparison between eating animals and eating *people*. Just as meat-eaters would not regard the cooked muscles of dead people as a commodity we call 'meat',[168] many ethical vegetarians view the cooked muscles of dead non-humans in a similar vein. Just as eating one's recently deceased cousin would, at least in contemporary Western

165 We need not regard this presumption as reflecting some *metaphysical* commitment. One could, instead, regard it as a *methodological* presumption. Any given moral theorist's skills lie in recovering principled, rational disputes, so it makes sense for their efforts to be concentrated in such an area. I return to this issue in §5.5, where I discuss the 'theoretical presumption' more explicitly.

166 One might explain this as something like 'squeamishness', the merely contingent, psychological fact that eating meat does not provide any sort of pleasure for some people. However, for many vegetarians, the kind of reaction involved is of a different order: there is a genuine *repugnance* at the very idea of eating meat, one that indicates some deeper emotional reaction than mere distaste is involved.

167 Even here, relevant differences come out, such that it is not entirely clear that is *suffering per se* that is the issue, but what it says about our society and the people involved. For instance, one sometimes hears a comparison drawn between the activities of the meat industry and the Nazi holocaust. This comparison is highlighted by Cavell (Cavell 2007) as a particularly perspicuous example of the kind of gulf that can emerge between people with different sensitivities: "The extreme variation in human responses to this fact of civilised existence [the mass production of animals for food] is not a function of any difference in our access to information... The variation of attitudes that Diamond's discussion stresses between the horror of individuals and the indifference of most of society considers moments in which the variation of response seems one between visions of the world, between how its practices are regarded, or seen, or taken to heart, or not." (Cavell 2007, p.282)

168 Even in cases of cannibalism, for instance, eating another human being is highly ritualised. It would be highly unusual to regard the eating of dead human beings as an everyday, domestic affair such that it was just a commodity.

culture, be regarded as disrespectful (at best), the vegetarian feels that eating a dead animal would belie their regarding animals as *fellow creatures* or *companions*. In other words, many ethical vegetarians *extend* thoughts and attitudes nearly everyone regards as appropriate to human beings to animals more generally:

> The response to animals as our fellows in mortality, in life on this earth... depends on a conception of *human* life. It is an extension of a non-biological notion of what human life is... The extension to animals of modes of thinking characteristic of our responses to human beings is extremely complex, and includes a great variety of things. The idea of an animal as company is a striking kind of case; it brings it out that the notion of a fellow creature does not involve just the extension of moral concepts like charity or justice... The treatment of an animal as imply a stage (the self-moving stage) in the production of a meat product is not part of this mode of thinking. (Diamond 1979, pp.329-30)

We might follow Diamond's suggested analysis by utilising the notion of 'sensitivities.' We note that often the difference between the ethical vegetarian and the meat-eater is not so much one of underlying rational commitments, i.e. their moral conception or relevant empirical beliefs, but rather a difference in their respective sensitivities to animals. The vegetarian and meat-eater often agree perfectly upon what counts as an 'animal', they agree upon whether the animal feels pain, is sentient, etc. They differ, however, in their sensitivities when faced with animals. Diamond's ethical vegetarian is liable to see the animal as a *fellow creature* much in the same vein as they would with fellow humans, and for them it automatically follows that one does not *eat* such things. Diamond's typical meat-eater, by contrast, differs in their sensitivity towards animals. An animal may be a *pet*, in which case they may be regarded as a *fellow creature,* but they do not extend this sensitivity to all animals. The typical meat-eater, Diamond suggests, will learn in part what it means to be *human* by sitting at a table eating something that does not partake of this fellowship: animals, *them*. (Diamond 1979, p.324) Similarly, we sometimes learn to respond to animals as *vermin*, as requiring us to kill them. Our sensitivities to animals are indeed a complex mass of responses. Ethical vegetarians, however, tend to have a sensitivity where animals are seen as continuous with humans, where they are 'fellow creatures'. This, maintains Diamond, is more often the salient difference between meat-eaters and ethical vegetarians.

It may seem that this sort of example is helpful for the sort of case Crary wishes to make concerning the *deepening* of sensitivity. It may appear that the meat-eater *lacks* a sensitivity that a more mature, more sensitive person has, much like a more sensitive person can appreciate the emptiness of Ivan Ilyich's life, or appreciate metaphorical uses of the word 'dog'. Diamond herself appears in little doubt about which sensitivity is preferable: she takes it as an obvious problem that many are insensitive to the "awful and unshakeable callousness with which we most often confront the non-human world." (Diamond 1979, p.334) The difficulty here, however, is that these judgements are little more than manifestations of the fact that the respective authors share these sensitivities. Part of having the ethical vegetarian's sensitivity is to regard the sensitivity itself as *deeper*, more developed, etc. However, there is no reason to accept that the sensitivity *is* more developed or more mature independently of the fact one happens to share that sensitivity in the first place. This is hardly a consideration that would, even under conditions of unlimited epistemic access, allow someone with a different sensitivity to appreciate that they had committed some sort of error.

If we've reached a point where it is persuasive to regard a moral disagreement as stemming from underlying differences in sensitivity, we have already implicitly accepted that it is inappropriate to search for constructive explanation of why it would be correct to regard an animal as a 'fellow creature' continuous with human beings, i.e. that reasons have given way. Further attempts at persuasion are thus better understood as attempts at *conversion*. We will seek to persuade others to share our sensitivity to animals such that they are seen as 'fellow creatures', but we have no *modus ponens*-style patterns of reasoning available such that one could rationally demonstrate that one's interlocutor had committed some sort of error.

It is then, what I shall refer to as an 'opaque disagreement'. An opaque disagreement is one where it is not rationally demonstrable that a party, or both parties, to a disagreement have made some sort of error. In terms of the relationship between *opaque* disagreement and *faultless* disagreements, if something is a faultless disagreement, then it is an opaque disagreement. However, an opaque disagreement need not be a faultless disagreement, since it may still be the case that, although we lack a method that allows us to identify one (or more) party/parties to the

disagreement as having made an error, one (or more) party/parties to the disagreement is in error. In terms of disagreement-entailing-error objectivity, then, an opaque disagreement is one where it could not be made *transparent* that the relevant area of discourse is disagreement-entailing-error objective. Again, it may nonetheless be the case that the area of discourse *is* disagreement-entailing-error objective.

Diamond or Crary may insist, then, that her sensitivity to animals, her set of *blind* inclinations to respond to animals as *fellow creatures,* is such that a person who did not respond in this manner (and thus make the sort of moral judgements appropriate to having this sensitivity) would be making some sort of error.[169] However, we appear to have no means of rendering this claim transparent, such that we have a method for identifying when such an error has occurred. Which sensitivity is correct is not something that is rationally demonstrable. There is little value in asking which party to the disagreement has a 'mistaken' sensitivity with regards to animals since we have no means, no matter how we idealise the epistemological circumstances, of uniquely identifying the correct sensitivity. All we can really do is note that there *is* this subtle difference in how a person responds to animals, and note that it leads to an opaque disagreement with regards to the moral permissibility of eating meat.

It strikes me, then, that Diamond's diagnosis is good, and that the relevant disagreement between the meat-eater and the vegetarian concerns differences in sensitivity regarding animals. Attempts to rationally reconstruct the debate such that it turns on the capacities of living things, and then questions of how much they suffer, how avoidable that suffering is, etc. mischaracterise the nature of the disagreement. At bottom, what we have are disputes about whether certain *basic* concepts will be automatically invoked when faced with animals. Which animals, and under which situations, will invoke notions of 'fellow creature'? This does not appear to be a question that is at all well-understood by attempting to rationally reconstruct some hypothetical chain of *modus ponens* reasoning. Rather, it is better to think of

---

[169] Crary, for instance, argues that one can demonstrate the inferiority of a sensitivity that does not accord animals with *dignity* in "an objectively authoritative manner" and that "We might get our interlocutor to acknowledge more respectful modes of response to animals as correct." (Crary 2007, p.399)

By my lights, this would be an example of *persuasion* or *conversion*. Nothing seems to be added by regarding this conversion as undertaken in an 'objectively authoritative manner', since, as Crary herself grants, the only way to appreciate that these 'more respectful modes of response' are correct is for one to already *share* the relevant sensitivity such that one can perceive the *dignity* of animals.

these things as matters of sensitivity, such that different language-users have subtly different sets of *blind* responses to the same situations. However, my intention is not to incorporate such matters within the realm of objective discourse (perhaps under some revised understanding of 'objectivity'), but to note the epistemological difficulty that such disagreements pose. In such cases, if they concern genuinely objective matters, it is nonetheless not the case that we have, or even could have, a method such that one might be able to identify which party/parties to the disagreement have made an error. They are opaque disagreements.

5.4 – Example 2: Pornography

My second example concerns the judgements of self-ascribed 'feminists' to the moral (and often thereby legal) permissibility of pornography.[170] Here, there are many interrelated relevant arguments, broadly dividing into consequentialist concerns as to whether, and in what way, pornography is *harmful*,[171] or deontological concerns as to whether the acts of producing or consuming pornography represent moral transgressions of one sort or another. I am interested here in a deontological concern to do with whether pornography both objectifies women and endorses the objectification of women, such that its consumption and/or production constitute relevant moral transgressions.[172] It may perhaps lead to relevant harms, such as sexual violence, but the argument in question is that pornography, insofar as it objectifies women and endorses such objectification, itself violates some moral duty.

As one might expect, professional philosophers focus upon providing clear and precise definitions of the relevant terms involved in an attempt to exhibit the underlying rational structure in the judgement that pornography violates some moral duty. Here I shall concentrate on Martha Nussbaum's paper 'Objectification' (Nussbaum 1995). In this paper, she attempts to locate the particular underlying rational structure to the judgement that pornography *objectifies,*[173] and endorses the

---

[170] I shall make no attempt at defining 'pornography', but shall rather use the term 'pornography' as a general term picking out the sorts of things universally regarded as 'pornography'. Paradigm examples are 'softcore' pornographic magazines like *Playboy* or *Hustler*, and 'hardcore' pornographic films produced by companies such as Vivid Entertainment and Private Media Group.

[171] See, for instance, Diana Russell's *Dangerous Relationships: Pornography, Misogyny, and Rape* (Russell 1998), in which relevant harms, typically to do with sexual aggression, are identified, and experimental data is cited to establish correlations between exposure to pornography and the relevant harm. For a more recent meta-analysis of relevant data, see (Hald, Malamuth & Yuen 2010), which finds statistically significant correlations in non-experimental studies between pornography and attitudes of permissiveness towards sexual violence. Of course, this is only a *correlation*, and does not establish a *causal* link. To my knowledge, there is no evidence of a direct causal link. (c.f. Strossen 1995, p.251)

[172] A particularly prominent example is the lifelong work of legal theorist and self-ascribed 'feminist' Catherine McKinnon, who maintains that "All women live in sexual objectification the way fish live in water." (McKinnon 1989, p.149) and that "Sexual objectification is the primary process of the subjection of women." (MacKinnon 1989, p.124) Thus, for MacKinnon, pornography, understood as "The graphic sexually explicit subordination of women through pictures and words" (MacKinnon 1984, p.176), is a particularly perspicuous example of a more widespread phenomenon: the way women are systematically portrayed and understood primarily as objects for men. The acts of producing or consuming pornography are discriminatory in nature.

[173] To be clear, Nussbaum first attributes to MacKinnon and Andrea Dworkin the argument that pornography, insofar as it amounts to *objectification*, is morally objectionable. She then argues that, *pace* MacKinnon and Dworkin, it is not objectification *per se* that is objectionable, but rather *instrumentalisation*, a particular form of objectification, that is objectionable.

objectification of, women. Her first step is to offer the following list of things that we may mean when we speak of 'objectifying' a person:

1. *Instrumentality*: The objectifier treats the object as a tool of his or her purposes.
2. *Denial of autonomy*: The objectifier treats the object as lacking in autonomy and self-determination.
3. *Inertness*: The objectifier treats the object as lacking in agency, and perhaps also in activity.
4. *Fungibility*: The objectifier treats the object as interchangeable (a) with other objects of the same type, and/or (b) with objects of other types.
5. *Violability*: The objectifier treats the object as lacking in boundary-integrity, as something that it is permissible to break up, smash, break into.
6. *Ownership*: The objectifier treats the object as something that is owned by another, can be bought or sold, etc.
7. *Denial of subjectivity*: The objectifier treats the object as something whose experience and feelings (if any) need not be taken into account (Nussbaum 1995, p.257)

Armed with such distinctions, she sets about rationally reconstructing the argument as to why consumption or production of pornography constitutes a relevant moral transgression. (Nussbaum, p.268) She focuses on the first criterion: instrumentality. Pornography is then regarded a 'paradigm source' of the tendency to confuse sexual intimacy with hierarchy and domination, to view people as *tools* for the sexual gratification of another, typically of women by men. Nussbaum makes clear that the key point is that pornography exhibits a strong tendency to depict women solely as *instruments* for the gratification of men. Thus, Nussbaum's argument is that it is not necessarily that *objectification* is objectionable, but that a particular sort of objectification, *instrumentalisation*, is objectionable.[174] If it were the case that the depictions of sexual activity involved viewing the parties concerned as something more than mere instruments, then, the depictions would not be thus objectionable. For instance, Nussbaum argues that scenes in D.H. Lawrence's *Lady Chatterley* are not objectionable although they are cases of objectification in the sense of 'fungibility'. Lawrence describes sexual activity in a way such that either protagonist could be substituted for a different person with the same respective body parts, but Nussbaum regards this sort of objectification is benign since it does not involve

[174] Even then, things are not quite so clear-cut. Treating a person as a tool need not be impermissible provided it's not the only thing one does. It is not, presumably, morally impermissible to use a partner's hand as a paperweight whilst they're asleep; it would be, however, impermissible if that, or similar instrumentalising actions, were the *entirety* of one's contact with that person. In what follows, therefore, I take the argument to be that pornography *solely* instrumentalises people (particularly women).

instrumentalisation, particularly to the extent typical of pornography. (Nussbaum 1995, pp.274-5)

Nussbaum here provides some further underlying rational structure for the thought that treating a person solely as a tool for another's sexual gratification ought to be regarded as morally impermissible. The relevant underlying rational structure here is essentially a Kantian one: that instrumentalisation is incompatible with treatment of other people as loci of respect, duty and autonomy in the Kingdom of Ends.[175] Here, we have a moral duty to treat other human beings as autonomous agents, and pornography, insofar as it depicts the female human beings involved solely as instruments for the use of men, is itself a depiction of a dynamic between human beings that is morally objectionable.

Thus, the initial action, treating women solely as *instruments* for the sexual gratification of men, violates a respective moral duty, i.e. to treat one another not merely as *means* but also at the same time also as an *end*. However, pornography also serves to *endorse* such actions, either actively or by giving the impression that the actions and human relationships depicted are normal and/or permissible. The key point is that pornography, for many feminists at least, amounts to an *endorsement*, a *glamorisation* even, of instrumentalisation in the context of sexual activity. For such feminists, pornographic depictions of sexual activity endorse the view that women exist solely as instruments for the use of men. The routine impression created by pornographic depictions of sexual activity is that "The women, including whatever signs of humanity they display, are just there to be used as sex objects for men in whatever way suits them." (Nussbaum 1995, p.280) Even in 'polite' cases like that of *Playboy* magazine, where Nussbaum grants that women's "autonomy and subjectivity are given a nodding sort of recognition", the underlying message seems to her to be clear: "whatever else this woman is and does, for us [the vastly predominantly male readership] she is an object for sexual enjoyment." (Nussbaum, p.283)

---

[175] "The subject of ends, that is, the rational being itself, must be made the basis of all maxims of actions, never merely as a means but as the supreme limiting condition in the use of all means, that is, always at the same time as an end." (Kant 2005 p.87 [*Groundwork* 4:438])

In line with ft.174, *instrumentalising* (treating a person as a tool of his or her purposes) a person might not necessarily constitutes a violation of the Categorical Imperative since it need not involve treating a person merely as a means. However, *solely instrumentalising* a person clearly would violate the Categorical Imperative.

It seems to me that, at the heart of Nussbaum's argument, is a reliance on a certain sort of sensitivity to *instrumentalisation*. We are intended to be able to recognise, when faced with various representations of sexual activity, in which cases it is appropriate to apply the concept of *instrumentalisation*, i.e. where "The objectifier treats the object as a tool of his or her purposes." We also need to understand whether the case is one where all that is going on is that a person is being treated solely as a *tool*, and we are also intended to be able to recognise where such representations implicitly *endorse* such instrumentalisation. Her argument relies on our being able to apply correctly the relevant concepts, e.g. *tool, endorsement,* etc. when faced with pornographic depictions.

At this point, we have reached a point where no further *rules* or *principles* are in the offing. Here we have nothing approaching a *rule* that could be employed in the *modus ponens* fashion in order for us to be able to say whether any given sexually explicit depiction involves a person being used solely as a *tool* for sexual gratification. In Wright's terms, we have arrived at a *basic* case of rule-following, where no further reasons are even appropriate, let alone available. We are simply assumed to have the respective sensitivity, to be able to recognise the applicability of such concepts in particular situations. It is only given that we have the respective sensitivity to seeing that a person is being treated *solely* as a *tool* for another's *purposes*, and that the depiction *endorses* this treatment, that Nussbaum's Kantian argument can get off the ground. Regardless of what we might think of the argument itself, it relies on our having the appropriate sensitivities.

Doubtless, many feminists would regard having a sensitivity to see more cases as instrumentalisation as more developed, less callous, more mature etc. Crary, for instance, takes a very similar line on related feminist issues such as *harassment* and *domestic violence*,[176] whereby feminists offer "a representation of moral concepts as concerned with objective features of the world that are unavailable apart from particular evaluative perspectives... [and] as uncovering objective forms of gender bias that are likewise unavailable apart from particular evaluative perspectives –

[176] Again, I think Crary picks her examples with care. Knowing her audience, she can perhaps anticipate that there will be few readers who are likely to be insensitive to issues concerning sexual harassment and domestic violence. I choose the example of pornography precisely because we need to be aware that the *general* picture looks to be far messier. It seems to me to be undeniable that not all differences of sensitivity are such that we can be at all comfortable regarding one particular sensitivity as more developed, mature etc.

specifically, perspectives informed by an appreciation of the injustice of sexism." (Crary 2007a, p.191) As before concerning eating meat, and as Crary here explicitly acknowledges, one cannot give any reasons to those who lack the relevant sensitivity why they have committed some sort of mistake. It may well be the case that having a sensitivity such that one can see pornography as cases where people, particularly women, are treated solely as tools for the sexual gratification of others (i.e. men), and where this treatment is implicitly endorsed by the act of filming and distribution of this commodity, is superior to alternative sensitivities one might have. The problem is, however, that there does not appear to be any consideration available that could demonstrate this superiority in a way such one could legitimately regard having a different sensitivity as reflecting an error of sorts. Instead, we seem to be faced with cases where either commentator views another's sensitivity as impoverished or worse,[177] and an absence of any method whereby one could demonstrate who is correct.

To see this, first of all we note that many intelligent and informed commentators have a different complex of sensitivities to pornography. A prominent 'pro-sex feminist', Wendy McElroy, regards pornography's obsession with female sexual organs and sexual acts as a matter of *focus*: "Women are as much their bodies as they are their minds or souls. No one gets upset if you present women as 'brains' or as 'spiritual beings'. If I concentrated on a woman's sense of humour to the exclusion of her other characteristics, is this degrading? Why is it degrading to focus on her sexuality?" (McElroy 2008) Thus, whilst pornography depicts sexual activity between people in such a way as to make it plain that what is of interest about the women involved is purely her sexual characteristics and little else, this is not understood by McElroy as treating the women involved solely as *instruments*, but rather as a matter of *focussing* upon an aspect of their being that other people are interested in.

The situation, for her, is analogous to the producer's and viewer's attitude towards minor Hollywood actors: one might say that, for us, Seth Gabel, a supporting actor in the film rendition of Dan Brown's work of historical erudition, *The Da Vinci Code*, is of interest to us solely in terms of whether he can portray his role effectively, and

[177] Defenders of pornography, especially when they regard themselves as 'feminists', are often held by other self-ascribed 'feminists' as themselves corrupted. Nadine Strossen documents some of the more colourful terms of abuse that 'feminist' defenders of pornography receive in her *Defending Pornography*. (Strossen 1995, pp.23-35)

thus we, as an audience, are treating him purely a tool for our purposes. The Hollywood film industry, furthermore, implicitly *endorses* these acts of *instrumentalisation*. Presumably, however, nearly everyone's sensitivities are such that they would regard this matter along McElroy's line, i.e. where our interest in Seth Gabel as a matter of our *focussing* on an area of his being, rather than responding to the case as one where we are solely treating Seth Gabel as a tool for our purposes. The question is, however, why responding similarly in the case of pornography reflects some sort of error.

We are faced with a disagreement here that stems from differences in sensitivity. Nussbaum's sensitivity to pornographic depictions is such that she perceives the various actions involved as cases where a person is solely treated as a tool for another's purposes. McElroy's sensitivity to pornographic depictions is such that she perceives these same actions as where we focus on a particular aspect of that person. There does not appear to be any underlying rational structure to either's sensitivity, and there does not appear to be any considerations available that might allow us to demonstrate which set of sensitivities ought to be regarded as superior. It is more persuasive to regard the matter as one party to the disagreement having a different set of *blind* responses to the other party to the disagreement; of their having a different set of sensitivities.

As with the question concerning the moral permissibility of eating meat, it is not my purpose here to adjudicate this debate. The point of this example to make persuasive the claim that there are moral disagreements that do not reflect an underlying rational structure such that one could reveal what counts as the correct application of some concept in a *modus ponens* fashion. In these examples, it does not seem plausible to view the application of the relevant concepts, *animals*, *fellow beings*, *tool, endorsement, focus*, as themselves the outcome of some underlying piece of *modus ponens* reasoning. Rather, it seems to be more plausible to regard such cases of disagreement as stemming from having different sensitivities with regards to the application of the respective concepts, sensitivities that do not reflect some underlying rational structure but are rather a matter of differences in our *blind* responses to the same situation. If this analysis is correct, we need to accept there is no way to adjudicate the dispute such that one party to the disagreement may be revealed to be in error.

Both examples are opaque disagreements, such that we cannot rationally demonstrate that one or more parties to the disagreement are in error. We could not have, even in principle, some method that could establish which party/parties to the disagreement is/are in error.

5.5 – Against the 'Theoretical Presumption'

We shall see how these cases of opaque moral disagreement, if they are sufficiently prevalent, create an insuperable problem for the explicative project we've assigned to the would-be standard reflective equilibrium theorist in the next section. Evidently, however, much turns on whether the sort of analysis put forward over the last two sections is persuasive: if the reader finds the sensitivity-style analysis of the two examples above unpersuasive, it is unlikely that they will be sympathetic to the thought that such cases abound within the moral realm. In this section, then, I consider a reaction to the effect that the above analysis is unpersuasive since it underestimates the power of theoretical considerations to illuminate moral disagreement. Rather than serving to illuminate moral disagreement, the kind of analysis on offer amounts to a 'quietism' of sorts, whereby one leaps from the evident truth that one has been unable to give a persuasive rational explanation for the appropriateness of applying concepts such as *fellow creature* or *instrumentalising*, to the unwarranted conclusion that such an explanation is not available in principle.

In the eating-meat example, for instance, we saw that much turned on whether one's sensitivities to animals were such that one automatically invoked the concept of *fellow creature* or something similar. One might here reply that, rather than giving up the theoretical game, what is revealed here is that we are insufficiently clear as to what we *mean* by 'fellow creature', and that we should investigate exactly what sort of moral obligations we owe to 'fellow creatures', and in virtue of what. That is, we should attempt to uncover further underlying rational structure to the application of the concept of *fellow creature* such that one could cite considerations that establish to the meat-eater, say, that, they have made some sort of error in not recognising any given animal as a *fellow creature*.

Similarly in the pornography case. What seems to be the trouble here is that we need to better understand exactly when an action is correctly regarded as (solely) treating another person as a *tool* for one's purposes (i.e. *instrumentalising* someone), and when depictions of such instrumentalisation count as a relevant endorsement. What is needed here, we might think, is further theoretical examination such that an underlying rational structure may be revealed such that it can be revealed to McElroy,

say, that she has made some sort of mistake in not recognising pornographic depictions as endorsement of the treatment of another person solely as a tool for another's purposes.

We can generalise the point. We saw in §5.2 (c.f. ft.165) what I there called the 'theoretical presumption', so well expressed by Rawls in *Political Liberalism*, that "We should be prepared to find that the deeper the conflict, the higher the level of abstraction to which we must ascend to get a clear and uncluttered view of its roots." (Rawls 1996, p.46) The theoretical presumption is, when faced with any sort of dispute as to whether some concept applies in a particular situation, what is needed is *further theoretical work* such that we can clarify whether that concept is correctly applied in that situation or not. A sensitivity-style approach just seems to give up too soon; it may even be fair to regard it as a case of intellectual laziness.

In response, one can here agree with the general point that, faced with any particular example, one cannot, as it were, *prove* that further theoretical work could not be helpful. The examples given were intended as an appeal to the reader to regard it as *more plausible* to think that further theoretical work will not be helpful, and thus that it is more plausible to think of these cases as reflecting variability in sensitivity. Once we accept the analysis, and regard them as disagreements stemming from *blind* inclinations to apply this or that concept to a situation, it follows from the analysis that the disagreements need to be recognised as opaque disagreements where we have no means available of revealing which party/parties to the disagreement is/are in error. The claim, then, is not that we have established on any given occasion that further theoretical work *could not* be illuminating, but rather that, in these cases, a sensitivity-style analysis looks to be more persuasive. It is more illuminating to analyse the relevant moral disagreement in terms of differential sensitivities, competing sets of inclinations to situations where certain concepts are invoked *blindly*.

However, we can bolster this appeal to the persuasiveness of the analyses of these particular examples with an *a priori* consideration: the failure of rational demonstrability for metalinguistic claims. The upshot of the previous two chapters' argumentation is that we cannot sustain the thought that any given disagreement

about whether it is correct to apply some concept, moral or otherwise, *must* ultimately bottom out in some difference in underlying rational structure. That is, we have good reason to think that the *a priori* expectation that any case of moral disagreement must reflect some underlying difference in the rational commitments of the parties to the disagreement *must* be incorrect.

To see why, we need only consider the thought that any level of abstraction to the point where one sees the underlying moral principles at work in a rationally reconstructed argument requires that the concepts utilised in that rationally reconstruction conform to the pattern of application for that concept that one would expect, operating with the expected pattern of application for the relevant concepts. If, for instance, it is a salient point in a moral disagreement that something experiences pain, we take it that by 'pain' we mean *pain,* as opposed to *quain* (*quain* $=_{\text{def}}$ *pain* if they are a human; *simulation of pain* if they are non-human). In supplying the rational reconstruction of an argument in which the notion of 'pain' plays a key role, we rely on the stability of our inclinations with regards to what we regard as the correct application of 'pain', i.e. that we mean *pain* and not some ersatz alternative.

One might protest at this point that all we need do is further specify the relevant concepts at work in the argument, such that our ersatz understanding of 'pain' was, as it were, ruled out. However, the underdetermination strand at work in Kripke's sceptical paradox ought to remind us that such a strategy is not one that can be pursued indefinitely. At some point, all the available facts, all *potentially-available* facts, still underdetermine the correct application of pain. That 'pain' means *pain* is not rationally demonstrable. No one contests that we might be able to reach a point where all parties involved are reasonably satisfied that they are utilising their terms in the 'same' way, just as no one contests the thought that we all, as a matter of our actual practice, apply the 'same' concept of *addition* in cases of arithmetic. The point is that these matters are not rationally demonstrable. At some point, we take it, with *prima facie* legitimacy, that what we mean by 'pain' is *pain*, and we rely upon this stability in our further practices of supplying definitions and forming arguments.

To put things in terms of the basic/non-basic distinction, we know that the theoretical presumption cannot apply across the board because there must be *basic*

cases of rule-following. At some point, our responses to a concept have to be *blind*, where attempting to explain what counts as correct application via *modus ponens*-style reasoning is inappropriate. In basic cases, there is no non-circular way of specifying the relevant concept-rule such that its minor premise, a description of the situation such that the concept may be seen to apply or not, does not presuppose the very conceptual mastery that the concept-rule was intended to outline. Instead, we just take it, without reasons, and yet with *prima facie* legitimacy, that *this* is what counts as 'pain', and *this* gives the correct pattern of application for 'pain', and allows us to identify mistaken application of 'pain'.

The theoretical presumption is, when faced with differential moral judgements, to seek to rationally reconstruct a chain of *modus ponens* reasoning resulting in those differential judgements. The upshot of the failure of rational demonstrability, however, is that this theoretical presumption cannot be vindicated in all cases. We ought to *expect* that there will be cases where we have reached 'bedrock' and where "our spade is turned" (Wittgenstein 2001, §217), where it is inappropriate to look for underlying rational structure to illuminate the distinction between correct/mistaken applications of a concept. The question is not *if* there is a point at which it would be inappropriate to seek to illuminate the application of some *concept* on a particular occasion, but *when* it would be inappropriate. We cannot justify the *a priori* expectation that any case of moral disagreement must reflect some underlying difference in moral or non-moral beliefs without any variation in the disputants' conceptual sensitivities. It *may*, but that conclusion should be borne out from a careful examination of the particular disagreement, rather than from an *a priori* expectation that it any case of moral disagreement must reflect differences in underlying rational considerations.

In the two examples canvassed above, I have laid out what I take to be the most persuasive analysis of the nature of the moral disagreement. Faced with identical situations, agents have differential sets of sensitivities that lead them to form different moral judgements. The differential sensitivities incline them to regard an identical situation differently through their *blindly* invoking certain concepts, regarding certain concepts as (correctly) applicable, as opposed to others. The disagreement stems from differences in sensitivity as opposed to differences in

relevant non-moral beliefs or their moral conception. Further theoretical work will not serve to illuminate the nature of these disagreements, but will rather *obscure* them by over-rationalising the nature of the disagreement.

If the analyses of these examples are at all persuasive, as I believe they are, it ought to be relatively clear that they do not represent isolated or exotic cases. It strikes me that cases of moral disagreements due to differences in underlying sensitivities are present in disputes concerning abortion (do people respond to the unborn by invoking the concept of a *baby* or the concept of a *foetus*?), disputes over parenting style (is leaving one's child to resolve their own problems *neglect* or helping them achieve *independence*?), terrorism (the hackneyed phrase 'One person's 'terrorist' is another person's 'freedom fighter' has more than a grain of truth to it), is wearing the hijab an act of *submission* or *piety*? etc. It would labour the point to go through such examples one-by-one; if the reader is unpersuaded by the above examples, it is unlikely they will find further analyses any more persuasive. However, if the above analyses are persuasive, similar cases of opaque disagreements that stem from differential sensitivities look to be relatively common within the moral sphere.

This does not, however, entail that there is something especially peculiar about moral cases. As we've seen, even in 'hard' mathematical cases, we still had failure of rational demonstrability with regards to showing that 'S means *addition*'. There is likewise an inherent limit to our justificatory or explanatory powers such that further theoretical guidance was inappropriate in allowing us to see that *addition* is what we mean by '+'. Fortunately, due to the universal (or near-universal) nature of our inclinations with '+', the lack of rational demonstrability turned out to be a benign feature of our epistemology. In the absence of disagreement about how to apply '+', we have no particular need of some consideration that would allow us to uniquely identify *addition* as what we mean by '+'. Rather, we take our responses to '+' queries as having *prima facie* legitimacy, and the stability of our usage means we can then go on to presuppose that '+' means *addition* in explicating '+', constructing further functions using '+', etc.

The reason things tend to be different in *moral* cases is that there are many cases in which, where we reach the point at which further theoretical guidance is

inappropriate, we are not faced with universal (or near-universal) inclinations. We have exhausted our justificatory abilities, and yet we are still faced with extant moral disagreement. Such disagreements are better understood, on my approach, through regarding them as stemming from differential sensitivities. As such, they are more persuasively regarded as opaque disagreements, such that we have no means available (even in principle) such that an error can be identified.

There is no reason, peculiar to moral cases, why *moral* disagreement is especially problematic. The relevant distinction is not between cases of moral disagreement and non-moral disagreement, but rather between disagreements that bottom out in concepts where our inclinations are universal (or near-universal), and disagreements that do not. The claim here is that it looks to be a persuasive analysis of the moral realm in general that you get many cases of the latter. Nonetheless, one still needs to take each disagreement on a case-by-case basis. If it *turns out* that the relevant moral concepts are disputed, with no prospect of ascertaining which pattern of application ought to count as correct, then we have a case of opaque disagreement. But that, to echo Wittgenstein's phrase, (Wittgenstein 2001, §107) ought to be the *result* of our investigation, not its *requirement*.

My response to the theoretical presumption, and the attendant charge that it is unpersuasive to regard opaque disagreements as sufficiently prevalent in the moral sphere, is, to summarise, a two-stage one. The first stage is to acknowledge the *a priori* point that, given the failure of rational demonstrability, we should *expect* that our ability to show that some concept is (correctly) applicable to a situation is inherently limited, and that there must be some cases where we just take it, with *prima facie* legitimacy, that some relevant concept *just is* applied like *this*. A corollary of this is that we should not have the *a priori* expectation that every case of moral disagreement *must* ultimately bottom out in some dispute over a matter of fact or in some codifiable moral principle. The second stage of the argument is that, through a detailed examination of extant moral disagreements found within contemporary Western culture, one can see that it is more persuasive to analyse such moral disagreements as stemming from differential *blind* inclinations, different conceptual sensitivities to the situation. Whilst one *could* continue to search for further underlying principles at work, further differences in underlying rational structure, it

looks to be unlikely that this approach will, in these examples, be illuminating. It is philosophically more satisfactory to 'recognise the ground as the ground'. This would entail recognising that there are subtle variances at work in our inclinations to respond to a particular concept, and that these variances are not themselves reflections of underlying rational differences but are rather differences in how we *blindly* respond to the concept itself, i.e. differences in sensitivity.

5.6 – Reflective Equilibrium: Explicative Failure

We've been making room for the idea that certain cases of first-order moral disagreement reflect underlying variances in our sensitivities to situations. As such, it is possible that different people, even with identical moral and non-moral commitments, may well end up coming to different judgements when faced with the same situation. If A's sensitivities are such that they perceive some situated action φ such that 'cruel' is (correctly) applicable, yet B's sensitivities are such that they do not, they may well come to different moral judgements regarding φ. I've recommended that, for some cases of moral disagreement, it is more plausible to analyse them as reflecting differences in conceptual sensitivities rather than differences in underlying rational structure. As such, I've maintained that it is more plausible to think of them as opaque disagreements.

I shall now indicate exactly how such issues bear on matters of reflective equilibrium and explicating 'being justified' in holding some moral conception. We've seen that the standard approach to reflective equilibrium, an approach in which reflective equilibrium functions primarily *qua* coherence theory of justification in ethics, is committed to an explicative project in which it must be in principle possible (i.e. given enough time, resources, and theoretical wherewithal) to design some particular ideally exact REFLECTIVE EQUILIBRIUM explicatum sufficiently similar to an agent's 'being justified' in holding some moral conception. By accepting this REFLECTIVE EQUILIBRIUM explicatum (as opposed to myriad alternatives) as explicating 'being justified', we may then render the relevant area of discourse transparently disagreement-entailing-error objective. We'd know in advance that faced with any disagreement about whether some agent is justified in holding their moral conception, it is always possible to establish which party/parties to the disagreement is/are in error.

Regarding the acceptance of one particular explicatum to the exclusion of alternatives, we saw that there was an inherent limitation to our ability to *justify* or *explain* such an acceptance. As in matters of metalinguistic claims, there is a failure of rational demonstrability. Then, my 'Wittgensteinian' response, modelled on Wright's Wittgenstein, was to insist that just as it is well that we treat certain metalinguistic

claims as *prima facie* legitimate, it is well that we accept some particular explicatum (to the exclusion of alternatives) with *prima facie* legitimacy. The failure of rational demonstrability is, of itself, a benign concern. Whilst it is true that an explicatum only renders an area of discourse transparently disagreement-entailing-error objective *having already been accepted* (to the exclusion of alternatives), our practice is such that we accept particular explicata with *prima facie* legitimacy. The genuinely *problematic* cases are where we are attempting to explicate a notion that is already disputed to a significant extent, where this dispute arises from differential sensitivities. The problem in such cases is that the disagreement itself is opaque. We do not have, even in principle, some method whereby we could identify which party or parties to the disagreement is in error.

The argument now is, given the prevalence of the sort of moral disagreements canvassed above, it is likely we will be faced with a significant level of dispute as to how to explicate 'being justified' in ethics along reflective equilibrium lines. Recall that however exactly one develops the notion of 'reflective equilibrium', it always involves incorporating Rawlsian narrow reflective equilibrium, i.e. where there is a 'fit', i.e. a coherence, between one's initial considered moral judgements and the judgements that would result from correct application of one's moral principles to particular cases. Any 'bells and whistles' approach, such as Daniels' wide reflective equilibrium approach, incorporates *further* considerations that need to be incorporated within this coherent state. Our ideally exact REFLECTIVE EQUILIBRIUM explicatum would minimally require deciding the judgements that stem from the application of the moral principles contained in the agent's moral conception cohered with the agent's initial considered moral judgements.

The problem here is that whether a moral judgement *fits* with a set of moral principles looks to be a matter that is itself vulnerable to disagreements stemming from divergences in sensitivity. Two evaluators may share exactly the *same* moral principles, and yet, in virtue of having different sensitivities, will be inclined to apply those principles in different ways, resulting in their forming different judgements as to whether their moral judgements that result from the application of the moral theory 'fit' or cohere with their initial considered moral judgements. This, as we shall see, means that it is highly problematic to design an ideally exact REFLECTIVE

EQUILIBRIUM explicatum that we might accept (to the exclusion of alternatives) as explicating 'being justified' in ethics.

To see this, let's return to our examples. Firstly, with regards to the moral status of eating meat, imagine the following case: A and B share identical rational commitments, such that they both agree in their initial considered moral judgements that eating meat is impermissible, and further agree upon a certain moral conception, which includes the principle that eating one's *fellow creatures* is morally impermissible. However, A is disposed to respond to animals generally by automatically invoking the concept of *fellow creature*, whereas B's sensitivities when it comes to the application of this concept are somewhat narrower: they typically only respond to animals with the concept of *fellow creature* if the animal is considered a *pet*. In this scenario, upon asking themselves what their moral conception requires of them with regards to the moral permissibility of eating meat, A finds that it follows from their moral conception that one issue the judgement that eating meat is impermissible, since eating meat is eating one's fellow creatures, and A's sensitivity to animals is such that they *blindly* regard all animals from which 'meat' is derived as *fellow creatures*. However, B finds that the same moral conception does not, for them, mandate this judgement, since they have a different sensitivity to animals such that some are considered *livestock*, not *fellow creatures*. B is unwilling to revise their initial considered moral judgement that eating meat is impermissible, but cannot locate any particular reason for this judgement.

Given that attaining the state of reflective equilibrium requires of us that we achieve some sort of 'fit' or coherence between judgements resulting from the application of moral principles and considered moral judgements, A and B will form different judgements as to whether they are in reflective equilibrium with regards to eating meat. A will judge that they are in reflective equilibrium, and B will judge that they are not. B is unwilling to revise their considered moral judgement concerning the moral impermissibility of eating meat, and is then likely to seek to revise the set of moral principles they hold such that it returns judgements that align with this considered moral judgement. A and B, despite having identical moral and non-moral commitments, will *dispute* whether the particular moral conception they both initially held can be upheld in reflective equilibrium. If we make room for the idea that A and

B may have different sensitivities when it comes to animals, we need also to make room for the plausible scenario where A and B diverge on the question as to whether their (identical) moral conception gives rise to the sort of judgement that tallies or 'fits' with their initial considered moral judgements. Whether they come to see the moral conception as one that may be upheld in reflective equilibrium or not will depend upon their respective sensitivities, and not merely their rational commitments.

Now, consider our second example concerning the moral status of pornography. Let's imagine that A and B, in their considered moral judgements, are both firmly convinced that pornography is morally impermissible. They uphold exactly the same moral conception, a broad Kantian moral theory requiring of us that we treat one another as autonomous beings, such that treating someone, or endorsing the treatment of someone, solely as a tool for another's purposes is morally impermissible. However, upon thinking through the implications of their moral conception, they come to different moral judgements regarding the moral permissibility of pornography due to differences in their underlying sensitivities to instrumentalisation. A regards pornography as *endorsement* of situations where a person is *solely* being treated as a *tool*; B regards pornography merely as a matter of *focussing* on a particular aspect of a person.

Again, given that attaining the state of reflective equilibrium requires of us that we achieve some sort of 'fit' or coherence between judgements resulting from the application of moral principles and considered moral judgements, A and B will form different judgements as to whether they are in reflective equilibrium with regards to pornography. To A, the judgements issuing from the application of their moral principles to the situation will 'fit' or cohere with their initial considered moral judgements, and they will judge that they are in reflective equilibrium on the matter. To B, however, their moral conception does not generate moral judgements that tally with their considered moral judgements in this area, and will seek to revise their moral conception such that it does. In spite of having identical rational commitments, A judges that they are in reflective equilibrium on the matter, and B judges that they are not. Again, whether they come to see the moral conception as

one that may be upheld in reflective equilibrium or not will depend upon their respective sensitivities, and not merely their rational commitments.

If the respective analyses of the two examples of moral disagreement as stemming from differential sensitivities are correct or, at least, it is correct that there are such examples of disagreements where we have, in principle, no means of establishing error, it follows that there will be occasions where different evaluators will come to different conclusions as to whether an agent is or is not in reflective equilibrium. In the above scenarios, for instance, B will not only judge that they are themselves not in reflective equilibrium on the matter, but that neither is A. Similarly, A will not only judge that they are themselves in reflective equilibrium on the matter, but also that B is equally in reflective equilibrium. Their respective sensitivities inform their judgements as to whether the moral judgements issuing from application of the relevant moral principles contained in the agent's moral conception cohere with the agent's initial set of considered moral judgements.

Consider now the prospects for the supplying of an ideally exact REFLECTIVE EQUILIBRIUM explicatum as explicating (to the exclusion of alternatives) the notion of 'being justified' in ethics. Even if we already accept that something like Daniels' wide reflective equilibrium approach is the way to understand the notion of 'being justified', we are faced with an insuperable difficulty. Suppose that A designs a reflective equilibrium explicatum REFLECTIVE EQUILIBRIUM-A that, as ideally exact, is a procedure that takes any appropriate input, i.e. the set of considered moral judgements and moral principles the agent happens to hold (and, depending on the nature of the reflective equilibrium account on offer, relevant non-moral beliefs such as responses to relevant arguments and background theory), and effectively decides whether the agent is in REFLECTIVE EQUILIBRIUM-A or not. As it happens, REFLECTIVE EQUILIBRIUM-A precisely mirrors their evaluations concerning whether some agent holds their moral conception in reflective equilibrium. However, B also designs a reflective equilibrium explicatum REFLECTIVE EQUILIBRIUM-B, that also effectively decides any appropriate input in a way that precisely mirrors their evaluations concerning whether some agent holds their moral conception in reflective equilibrium.

Here we have two proposed explicata that are logically distinct. The question arises, which explicatum ought we adopt to the exclusion of alternatives? Since, *ex hypothesi*, both are ideally exact, we cannot discriminate between them on grounds of exactness. Let us suppose also that there are not any particular differences between the two proposed explicata on grounds of fruitfulness or simplicity. The problem here is that A and B will have quite different estimations of the *similarity* between the explicatum and the explicatum. Both will have already granted, noting that the notion of 'being justified' is epistemologically vague (hence needing explicating in the first place), that any proposed explicatum need only be *sufficiently similar*, but they will nonetheless disagree as to how similar the relevant explicata are to their initial judgements concerning the coherence of an agent's moral conception. Under such conditions, it would not be the case that both A and B regard it as a *prima facie* legitimate to adopt one particular explicatum, rendering the question of whether an agent is justified in holding their moral conception transparently disagreement-entailing-error objective. The adoption of one would be regarded by either A or B as essentially question-begging. It would consequently not be regarded as a good reason to regard as mistaken one's judgement that some agent is not justified in their moral conception to be shown that that agent is in REFLECTIVE EQUILIBRIUM-A if one does not already accept that REFLECTIVE EQUILIBRIUM-A (to the exclusion of REFLECTIVE EQUILIBRIUM-B) explicates 'being justified' (in ethics).

The underlying problem here is that an evaluator's judgements as to what follows from a set of moral principles, and thus judgements of coherence between the judgements that follow from the principles and initial considered moral judgements, depend on the evaluator's set of sensitivities. Agents with different sensitivities make different moral judgements given identical moral conceptions, which gives rise to certain sorts of first-order moral disagreement in which agents diverge in their *blind* inclinations as to what concepts are (correctly) applicable. Such disagreements are opaque disagreements (such that we are unable to cite, under ideal epistemological circumstances, some consideration that would allow us to identify which party or parties to the disagreement are in error). At a second-order evaluative level, we get similar disagreements: evaluators with different sensitivities form different evaluative judgements as to whether the judgements that stem from applying the agent's moral

conception cohere with the agent's set of initial considered moral judgements. Again, such disagreements are cases of opaque disagreement.

An explicatum can only render an area of discourse transparently disagreement-entailing-error objective within the context of it having already been accepted (to the exclusion of alternatives). There are clearly cases, such as accepting PRIME, where it is *prima facie* legitimate to adopt an explicatum to the exclusion of alternatives, i.e. where it matches our universal or near-universal inclinations with regards to applying the explicandum. However, here we are faced with cases where our inclinations with regards to applying the explicandum do not display such universality or near-universality, even having already accepted that some reflective equilibrium approach is a good way to proceed. Under such circumstances, it is not *prima facie* legitimate to adopt one ideally exact REFLECTIVE EQUILIBRIUM explicatum to the exclusion of others. There is not a sufficient level of extant agreement concerning the application of concepts of *fitting* or *coherence* when it comes to moral cases.

The standard approach to reflective equilibrium contains a commitment to the theoretical possibility that some ideally exact explicatum resembling 'reflective equilibrium' exists that would transform questions of whether an agent is justified in holding their moral conception transparently disagreement-entailing-error objective. We can now conclude that, if there are many cases of moral disagreement stemming from differences in the respective sensitivities of the agent are sufficiently prevalent, such an explicatum is not available, no matter how much time or resources are expended on the matter. As such, we are forced to conclude that the standard approach, in which reflective equilibrium functions primarily *qua* coherence theory in ethics is fundamentally misguided. No such ideally exact explicatum could, at least given the current predicament, be designed, even in principle.

Chapter 6 – The Alternative Approach

6.1 – The Distinctiveness of the Argument

We saw from the outset that the notion of 'reflective equilibrium' is a term that is variously understood, from Rawlsian narrow reflective equilibrium, where moral judgements and principles are held in a coherent manner, to Daniels' wide reflective equilibrium, where (subject to the independence constraint) a state of reflective equilibrium obtains where one has a coherent triad of moral judgements, moral principles, and background beliefs (accompanied by relevant philosophical arguments). Furthermore, the term 'reflective equilibrium' is ambiguous between its use in referring to the epistemological theory that an agent is justified in holding some moral conception if and only if the various components of that moral conception are in a distinctive epistemic state of coherence, and its use in referring to a methodological recommendation in cases of moral reasoning.

These claims, as I have emphasised, are distinct. One need not think that 'reflective equilibrium' describes an epistemic state of coherence (in ethics) in order to think it worthwhile to proceed using the method advocated by Rawls (or some variant thereof), i.e. by collecting one's intuitions, filtering them into considered moral judgements, then attempting to delineate moral principles that capture those considered moral judgements, before evaluating, and revising either one's considered moral judgements or one's moral principles. Similarly, one need not think that advocating reflective equilibrium *qua* coherence theory entails that one ought to adopt the particular methodological recommendations falling under the aegis of 'reflective equilibrium', since it may be that another method more reliably leads to this epistemic state of coherence.

However, these concerns are far from universally recognised as distinct. An overwhelming consensus has emerged in which it is taken for granted that reflective equilibrium *qua* coherence theory is of primary concern, and that any methodological recommendations one might make merely serve to increase the likelihood of attaining the distinctive epistemic state characteristic of coherence in one's moral conception. I have called this the 'standard' interpretation, and stressed that an

alternative approach is available, such that one justifies recommending reflective equilibrium *qua* method in ethics for reasons unconnected to the coherence theory.

Given the popularity and influence of the standard conception of reflective equilibrium, it is clearly of importance if reflective equilibrium *qua* coherence theory of justification represents a misguided approach within moral epistemology. It is this that I take to be the principle achievement of the foregoing. Through three subsidiary claims (discussed below), (a) Clarifying the commitments of the standard approach to reflective equilibrium using Carnap's notion of 'explication' and Kölbel's notion of 'disagreement-entailing-error objectivity', (b) Tracing the implications of the rule-following considerations for explication as a tool for rendering areas of discourse transparently disagreement-entailing-error objective, and (c) Arguing that it is persuasive that certain sorts of problematic moral disagreements, opaque disagreements, are sufficiently prevalent in contemporary Western culture such that estimations of whether any given agent holds their moral conception in a state of reflective equilibrium will also be subject to opaque disagreements, I have established that, under such conditions, even if one accepts that 'being justified' in ethics is a matter of being in an epistemic state of coherence, there is no possibility of supplying an ideally exact REFLECTIVE EQUILIBRIUM explicatum (to the exclusion of alternative explicata along reflective equilibrium lines) that would be accepted with *prima facie* legitimacy.

Furthermore, each of these subsidiary claims represent advancements in terms of philosophical clarification. Concerning (a) then, there were three areas of potential confusion concerning the nature of reflective equilibrium *qua* coherence theory of justification in ethics that required attention. Firstly, we saw that we needed to clarify the intended relationship, hinted at in my vague expression 'tells us something about', between the idea of 'reflective equilibrium' and the idea of an agent's 'being justified' in their moral conception. One could construe the connection as one of traditional conceptual analysis in which the notion of 'reflective equilibrium' spells out the meaning of 'being justified', for instance, or one could construe 'reflective equilibrium' as purely a stipulative proposal, unconnected to the notion of 'being justified'. Here I followed Stich in regarding the relationship as one of explication, in which we replace the question 'Is agent *x* justified in holding their moral

conception?' with the question 'Does agent *x* hold their moral conception in reflective equilibrium?'

Secondly, it was not clear what it meant to regard reflective equilibrium as an epistemic state that 'obtains' in virtue of the logical and evidential relationships between the set of moral and non-moral beliefs that an agent happens to hold. Such 'obtaining' might be construed as a truth-apt matter (i.e. a fact), or it might be construed as a matter of it being warranted to assert that 'Agent *x* is in reflective equilibrium'. Here, I clarified what it meant for a state of reflective equilibrium to 'obtain' as a matter of it being *correct* to claim that 'Agent *x* is in reflective equilibrium' where such a claim reflected a disagreement-entailing-error objective matter. If 'Agent *x* is in reflective equilibrium' is correct, and the matter is disagreement-entailing-error objective, then it would be mistaken to claim the contrary.

Thirdly, it was unclear whether, and in what way, the coherence theory should be of practical *use* in questions of moral theory acceptance. Clearly, in considering matters of justification, one wants to make matters of theory acceptance *more tractable* than questions of truth, but need it be of *immediate* use? Here, I argued that, whilst it was premature to demand that we provide an explication of immediate use, it is nonetheless legitimate to demand that any working explicata approximate to an ideal explicatum that is in principle available. It ought to be in principle possible, given that 'reflective equilibrium' is intended to pick out an epistemic state that 'obtains' in virtue of the logical and evidential relationships between the relevant moral and non-moral beliefs, to supply an exact REFLECTIVE EQUILIBRIUM explicatum that would, through being an effective procedure, effectively decide whether it would be correct to regard any appropriate case as one that is in REFLECTIVE EQUILIBRIUM or not. If we can be confident that such an ideal is in principle available, and eludes us due to contingent limitations of time, resources, or theoretical wherewithal, it continues to make sense of the attribution of 'reflective equilibrium' as a disagreement-entailing-error objective matter, and from here one can design working explicata that approximate to this ideal.

Concerning (b), I have, consistent with Wright's construal of Wittgenstein, detached from Kripke's sceptic's argument a strand of argument that reflects an *epistemological*

concern, distinct from Kripke's sceptic's overall constitutive concern in urging a sceptical paradox on the would-be meaning factualist. The importance of this strand of argument is, as Wright's distinction between basic and non-basic cases of rule-following brings out, that there is an inherent limit to our justificatory powers, a limit I've referred to as a failure of 'rational demonstrability'. As a result, we ought to *expect* that there will be cases where there are no considerations available, no matter how far we idealise the epistemological circumstances, where one metalinguistic claim, say that 'S means *addition*' can be uniquely identified as correct, thereby excluding 'S means *quaddition*' and countless other ersatz alternatives. Nonetheless, it would be absurd to maintain that it would be inappropriate to make such claims. At this point, I've urged the view, urged by Wright's Wittgenstein, that no constructive explanation of why it is justified to make such claims is available. Nonetheless, at least on some occasions, we accept such metalinguistic claims with *prima facie* legitimacy.

This reading of Wittgenstein is not especially ground-breaking, but what is of novel interest is that identical considerations carry over into questions of explication. On any given explicative occasion, there is an inherent limit to our ability to justify one explicatum over another, as is needed where we intend the explicatum to replace the explicandum. It is not rationally demonstrable *that* some particular explicatum should be adopted to the exclusion of others. As a result, an explicatum can only render an area of discourse transparently disagreement-entailing-error objective within the context of it having already been accepted. Nonetheless, it would be absurd to maintain that it would be inappropriate to utilise explicata in a manner such that they rendered areas of discourse transparently disagreement-entailing-error objective (think of cases like TEMPERATURE or PRIME). The proposal is, mirroring Wright's Wittgenstein's approach, that we accept explicata (to the exclusion of alternatives) with *prima facie* legitimacy. This allows us to draw a distinction between benign cases of explicating some concept (e.g. where the explicandum is vague or ambiguous), and where an explication would be regarded as problematic. The benign case is where we explicate a concept that is already applied in such a universal (or near-universal) manner that any given language-user will accept it with *prima facie* legitimacy, in spite of being unable to rule out logically distinct alternatives. The problematic case is where we are faced with explicanda that are disputed to an certain extent. Here, there will be extant disagreement concerning which explicatum to accept to the exclusion

of others, and we have reached the limit of our justificatory powers. That is, we are faced with an opaque disagreement over which explicatum to accept.

Concerning (c), then, I have adapted an understanding of certain sorts of moral disagreement that are persuasively viewed as stemming from differences in underlying sensitivities, differences in the set of *blind* inclinations to respond with certain concepts as opposed to others when faced with a situation. However, my understanding here is distinct in that I am sceptical concerning the extent to which, when disputes arise as to which sensitivity/sensitivities are superior, they are amenable to rational resolution. There are, at least, some cases of moral disagreements stemming from differences in sensitivity, such as my two examples, where there does not appear to be any consideration available such that it could be rendered transparent which party to the disagreement has made some sort of error (through having an impoverished sensitivity, say). One is forced to conclude that, whilst the Crary/Diamond/McDowell form of sensitivity-style analysis concerning particular examples of moral disagreement is persuasive, of what they've really persuaded us is that there is a prevalence of opaque disagreements within the moral sphere, where we've gone as far as we are able with theoretical considerations and yet there is nonetheless significant extant disagreement. Ultimately, as we've seen, it is this *de facto* prevalence of opaque disagreements in the moral sphere, as opposed to some feature unique to moral disagreement, that forces us to conclude that there cannot be an ideally exact REFLECTIVE EQUILIBRIUM that would be accepted with *prima facie* legitimacy.

Although these subsidiary matters represent useful clarifications and developments in themselves, it is their synthesis that yields the most salient conclusion: that reflective equilibrium *qua* coherence theory, at least as things stand currently in the moral sphere, represents a misguided approach. Whether or not an evaluator would be inclined to judge that an agent held their overall moral conception in reflective equilibrium depends on their sensitivities when faced with a whole host of particular scenarios, such as whether they are inclined to see eating meat as eating *fellow creatures*, or whether pornography represents *endorsement* of treating a person *solely* as a *tool* for another's purposes, and myriad other inclinations. As such, the ideally exact REFLECTIVE EQUILIBRIUM explicatum they would be inclined to accept (to the

exclusion of alternatives) would vary considerably depending on their sensitivities over these myriad cases. Under such a scenario, it would not be *prima facie* legitimate to adopt any particular REFLECTIVE EQUILIBRIUM explicatum. Even if one already accepted that the way to explicate 'being justified' in ethics would be through adopting a standard reflective equilibrium approach, i.e. a coherence theory of justification in ethics, one is nonetheless committed to the availability, in principle, of an explicatum that, given the prevalence of opaque disagreements in the moral sphere, is not available. One would be engaged in an explicative project where one attempts to design explicata that approximate to an ideal that is not even in principle available. It is in this sense that the standard approach to reflective equilibrium is misguided.

This yields a further, important, conclusion. We've already noted that one might adopt an alternative interpretation of reflective equilibrium in which one understands reflective equilibrium *qua* methodological recommendation without necessarily appealing to the coherence theory. My argument, however, establishes something stronger, namely that if one is at all interested in reflective equilibrium *qua* methodological recommendation, as, for instance, many are in the field of applied ethics and bioethics in particular, one would be well-advised to avoid making such a recommendation via an appeal to the coherence theory of justification in ethics. That is, if one wishes to advocate the use of reflective equilibrium methodology, one had better *not* do so through the invocation of a coherence theory of justification, since it is committed to the in-principle-availability of an ideal explicatum that we have good reason to think cannot be available.

In this final chapter, I wish to develop my alternative interpretation[178] of the method of reflective equilibrium, by setting out what I take to be the best way to advocate

[178] Strictly speaking, there is not *one* alternative interpretation, but rather *many* alternative interpretations in which the reflective equilibrium methodology is advocated for reasons unconnected to achieving an epistemic state of coherence in one's moral view.

One way, for instance, in which one might carry out the meta-justification of the use of reflective equilibrium methodology is to argue that reflective equilibrium methodology is *inescapable* given a commitment to rational thought. DePaul, in 'Why Bother With Reflective Equilibrium?' (DePaul 1998, c.f. DePaul 2006 pp.616-8), argues that any supposed alternative method would have to either advocate abandoning reflection, or advocate incomplete reflection, or not allow an enquirer's reflections to guide their moral beliefs, all of which are irrational proposals. (DePaul 1998, pp.301-7)

As we shall see in §6.4 where I discuss the 'vacuity worry', DePaul may well have overplayed his hand here. There appear to be methodological recommendations one might make that are not obviously irrational, especially concerning the recommendation that one avoids the use of moral

reflective equilibrium methodology independently of the coherence theory.[179] Although a developed account of the virtues of the method of reflective equilibrium, divorced from the coherence theory, is beyond my scope, I can at least indicate why I take it to be worth thinking about (wide) reflective equilibrium as a methodological doctrine independently of the coherence theory, as well as giving some indication as to why my own construal of the method of (wide) reflective equilibrium is distinctive.

I shall clarify my alternative interpretation of the method of (wide) reflective equilibrium by considering three potential worries one might have concerning my recommending the method. The first worry concerns, not so much the merits of the method itself, but rather whether it is consistent with my argument to recommend the method of (wide) reflective equilibrium.[180] One might worry that, *by my own lights,* the method of (wide) reflective equilibrium ought to be abandoned. I have employed, one might maintain, an argument for an 'anti-theory' position, and yet I wish to recommend a method in moral reasoning that is fundamentally theory-driven. I call this 'the anti-theory worry', and discuss it in relation to very similar concerns that have emerged in the burgeoning literature concerning moral particularism.

Secondly, we might worry about the *point* of engaging in such a method if we have implicitly abandoned the idea that being in reflective equilibrium (and thus 'being justified' in the area of moral theory) could reflect transparently disagreement-entailing-error objective matters. If one divorces the method of (wide) reflective equilibrium from the coherence theory of justification, it is unclear why one ought to engage in the method of (wide) reflective equilibrium. Here I shall argue that two of Daniels' suggestions in favour of the method of wide reflective equilibrium can be carried over into my alternative approach, i.e. that we can extricate his methodological recommendations from considering reflective equilibrium as a coherence theory. This shall make it clear that, although it is not the case that the method, on my alternative approach, aims at achieving a distinctive epistemic state of

---

intuitions. The typical reply here (as we shall see) is to insist that one *must* rely on intuitions at some point, but I do not take this to have been conclusively demonstrated.

[179] In what follows, where I refer to 'the alternative interpretation' or 'the alternative approach', I am referring to *my specific proposal*, as opposed to any approach to reflective equilibrium methodology that is not committed to the coherence theory.

[180] Although, as we shall see, responding to this worry requires certain adaptations to how we conceive of the method of (wide) reflective equilibrium, which is what I shall do in my own alternative approach.

coherence, the adoption of the method may nonetheless *improve* our moral epistemological predicament. I shall even maintain that, with an appropriate modification to Daniels' method, in line with adapting to considerations of sensitivity, the alternative method of (wide) reflective equilibrium here canvassed is useful in allowing a person to become more responsive to the sources of moral disagreement.

Thirdly, we might worry about whether the method of wide reflective equilibrium imposes any substantive constraints on moral reasoning, and thus whether it is merely an empty or vacuous recommendation, amounting merely to a claim along the lines of: If one is interested in thinking about what to think in moral matters, one should think about moral matters. Broadly in line with Scanlon's response to this 'vacuity worry', I argue that the method of (wide) reflective equilibrium is not vacuous, and that, furthermore, its distinctiveness *increases* once one adopts my alternative interpretation of that method.

6.2 – The Anti-Theory Worry

The argument of the last three chapters may well be regarded as an argument for an 'anti-theory'[181] position in ethics. More specifically, the argument shares affinities with certain strands of 'moral particularist'[182] argument, which may be regarded as falling under the umbrella of 'anti-theory'.[183] However, at the same time, I wish to recommend the method of (wide) reflective equilibrium, a method that is evidently a matter of systematising and extending one's moral views utilising moral principles. On the face of it, this would appear to be an inconsistency: how, given the argument of the last three chapters, can the method of reflective equilibrium, on any interpretation such that it is worth regarding as 'reflective equilibrium' methodology, be recommended?

The incompatibility of the method of reflective equilibrium (as standardly conceived) and the moral particularist approach is noted by Tersman:

> Coherence is a matter of certain evidential and explanatory relations holding between the agent's moral views, where some explain and others are explained by the rest (relative to the agent's nonmoral beliefs). This entails that a reflective equilibrium, and thus also justification, is achieved only if the agent has come to accept certain general normative views. There are views that deny that the justification of moral beliefs requires acceptance of such

---

[181] 'Anti-theory' in ethics is a very loose catch-all term and, as Robert Louden points out, (Louden 1992, pp.87ff.) very much depends on what one means by 'moral theory' in the first place. Louden's examples of those holding 'anti-theory' positions include Annette Baier in her aim "To attack the whole idea of a moral "theory" which systematizes and extends a body of judgements." (Baier 1985, p.232) and Bernard Williams, who argues that "Philosophy should not try to produce ethical theory" since "In ethics the reductive enterprise has no justification and should disappear." (Williams 1985, p.17) Broadly speaking, an 'anti-theory' position in ethics is the view that moral theories ought to play, at best, only a minimal part in our reasoning about moral matters.

[182] 'Moral particularism' refers to a family of interrelated arguments and positions regarding the use of moral principles in morality. The debate between particularists and generalists in morality is "About the extent to which morality can and should be understood in terms of moral principles." (McKeever & Ridge 2006, p.3) Much as being 'anti-theory' depends on what one means by 'theory', here being a particularist and rejecting the widespread use of moral principles depends on what we mean by a 'moral principle'. In turn, there are many corresponding arguments against the use of moral principles. Consequently, there are various strands of particularism that I shall briefly discuss below.

[183] Given the platitudinous assumption that moral theory requires the use of moral principles, it would seem to follow that, for moral theory to have any significant (i.e. *justificatory* and/or *explanatory*) role in moral reasoning, one needs moral principles to enjoy a significant (i.e. *justificatory* and/or *explanatory*) role in moral reasoning.

However, as I shall become clear below, not all strands of particularism disavow the use of moral principles, or even restrict them to playing a non-justificatory or non-explanatory role. Whilst *some* forms of particularism may be 'anti-theory' therefore, other forms, and in particular the form I shall endorse, are not 'anti-theory' positions.

> principles. For example, it is denied by the approach called 'moral particularism'. (Tersman 2008, p.400)

On the standard view of reflective equilibrium, being in a state of coherence with regards to one's moral conception requires that the judgements that would follow from one's set of moral principles cohere with the set of considered moral judgements one is inclined to make. As such, it requires that an agent's moral conception contains a set of general moral principles from which, in conjunction with facts about particular situations and actions, the agent has enough information to *explain* and/or *justify* the particular moral judgement (and the evaluator can thereby compare it with the agent's set of considered moral judgements to see if they are in a state of reflective equilibrium). The standard view of reflective equilibrium takes it that, because coherence is a matter of the logical and evidential relationships between the moral principles, moral judgements, and (depending on the exact conception of 'reflective equilibrium') relevant background beliefs, it will be possible, using only this information, to *explain* why a certain judgement *derives from* the moral theory that an agent holds. As such, the method of reflective equilibrium (as standardly conceived) requires a conception of 'moral principles' whereby it is assumed that the agent and evaluator *have enough information* to perform their respective tasks. Once you specify the relevant rational commitments of the agent (i.e. the set of moral and non-moral beliefs), and specify the facts about a situation and an action, the agent can see what moral judgement to make, and the evaluator, given a suitable specification of 'reflective equilibrium', can see if the agent is in reflective equilibrium.

The method of reflective equilibrium (as standardly conceived) requires firstly, a general commitment to moral principles having an *explanatory* and *justificatory* role, and secondly, a more specific commitment to the nature of moral principles. This first commitment I take to be a broad commitment to the worth of moral theory, and is incompatible with an 'anti-theory' position. The second commitment, however, I take to be incompatible only with a specific conception of 'moral principles'. What I shall argue here is that one can, indeed should,[184] disabuse oneself of one specific conception of 'moral principles' and retain the broad commitment to moral

[184] As I shall emphasise, my argument commits me to the thought that one ought not require of 'moral principles' that they are universally (or near-universally) followed by different agents in identical situations, whereas the standard approach implicitly assumes that moral principles will be universally (or near-universally) followed. I shall argue that this requires of us to *reconceive* the nature of 'moral principles', and not to abandon them.

principles playing an explanatory or justificatory role. I take this to be a position that is antithetical to a certain *conception* of moral principles and moral theory but, in the absence of some demonstration that this conception is compulsory, not an 'anti-theory' position.

Before proceeding further, we need a better characterisation of the family of arguments and resultant positions within 'moral particularism'. Here I wish to emphasise that not all strands of particularist thought entail the *rejection* of moral principles; some emphasise the need to *reconceive* the nature of 'moral principles' especially the idea that 'moral principles' need to be *exceptionless*. Following Mark Lance & Margaret Little's survey of particularist positions (Lance & Little 2006), we first of all need to distinguish between *epistemological* particularists and *metaphysical* (or, more accurately, constitutive) particularists. Epistemological particularists are interested in the role of moral principles in *moral reasoning*, and here there is a spectrum of stances one might adopt. Many, including Baier and Iris Murdoch, favour a distinctly anti-theoretical stance in which "Moral principles can serve pedagogic and heuristic roles—they can help us to develop mastery of moral concepts and discern their instances; but they do not mark epistemically or explanatorily rich inferential relationships between propositions." (Lance & Little 2006, p.578) Others, including Lance & Little themselves, take it that moral principles, suitably reconceived, "As both argument for moral conclusions and unifying explanation of moral phenomena." (Lance & Little 2006, p.591)

Constitutive/metaphysical particularists, notably Jonathan Dancy, are interested in the role of moral principles as *instantiating*[185] the moral value that an action carries. For any given action carrying a particular moral value (i.e. permissible, impermissible, supererogatory), the relevant moral-value-makers, what Dancy calls 'reasons', operate holistically. That is, there are no *invariant* reasons that will necessarily contribute to the moral value of an action in a uniform, 'univalent' way, such that useful explanatory generalisations, codifiable in the form of moral principles, are available.

[185] Dancy refers to this relationship as one of 'resultance': "Resultance is a relation between a property of an object and the features that 'give' it that property. Not all properties are resultant; that is, not all properties depend on others in the appropriate way. But everyone agrees that moral properties are resultant. A resultant property is one which 'depends' on other properties in a certain way. As we might say, nothing is just wrong; a wrong action is wrong because of other features that it has." (Dancy 2006, p.85)

Any 'explanation' of why some action φ is morally permissible (say), i.e. in virtue of what φ carries the moral value it carries, would need to include an indefinite number of considerations before one could regard the moral value of φ as determined. It would need to mention, for instance, the absence of potential *disablers* (Dancy 2006, pp.39-41) and these alone are indefinite in number. The list of reasons, the properties in virtue of which φ carries the moral value it carries, will thus not resemble anything worth calling 'moral principles'.

Since we're interested in the role of theory in moral reasoning, I shall restrict myself to the epistemological concern within particularist positions, in which we can distinguish between those who seek to *abandon* moral principles in any significant, i.e. justificatory and/or explanatory, role, and those who seek to *reconceive* the notion of 'moral principles' such that they can continue to play a significant role in moral reasoning. What is agreed upon, however, as Lance & Little argue, is that moral particularists all reject the 'classical' conception of 'moral principles', moral principles as universal, exceptionless, law-like generalisations that form part of a structure (i.e. a moral theory), and are crucial to justifying particular moral judgements. (Lance & Little 2006, pp.570-1) The dispute amongst epistemological particularists concerns how one ought to react to this rejection of the classical conception, and is between, as Lance & Little put it, 'moderates' and 'radicals'. (Lance & Little 2004, pp.436-7) As such, it is clearly possible to be a 'moral particularist' of the moderate or reformist stripe, allowing a significant role for moral principles *under some alternative conception* of 'moral principles', as opposed to the radical's abandoning the idea of moral principles playing any role beyond mere heuristics or pedagogical use.

I shall now sketch one sort of reformist moral particularism proposed by Lance & Little (Lance & Little 2004; Lance & Little 2007), before illustrating how my own argumentative strategy is very similar. Their strategy involves reconceiving the nature of 'moral principles' such that they are no longer conceived of as *exceptionless* generalisations, but rather along lines of being defeasible generalisations.[186] In doing so, they retain a justificatory and explanatory role for moral principles.

[186] In actuality, their account is not restricted to *moral* cases of defeasible generalisations. They regard it as something of a puzzle that the kind of reason holism emphasised by moral particularists should be so widely disputed, yet regarded as almost platitudinous in epistemology more generally. For instance, take the generalisation that, if something *appears* to be a cup, it is a cup. Normally, the fact that

Lance & Little maintain that such defeasible generalisations involve implicit 'privileged conditions' modifiers. Privileged conditions are understood via reference to paradigm examples in which the relevant moral principle would be invoked, and may reflect a *moral* privileging of certain conditions as being superior, and/or an 'explanatory' or conceptual privileging. An example of the former case of morally superior privileged conditions is that of the impermissibility of killing. (Lance & Little 2006 pp.589-90) Killing is always wrong under privileged conditions, where the privileged condition here is to be understood as 'in morally superior situations'. Exceptions to the moral principle exist, however, where conditions deviate from the morally superior situation: e.g. cases of self-defence, where obviously it would be a morally superior situation if no person needed to utilise self-defence in the first place. An example of the latter case of 'explanatory' privileged conditions is that of the impermissibility of lying. Lying is always wrong under privileged conditions, and here it is crucial to understanding cases that happen to be exceptions to this generalisation *that* one understands them *as exceptions*. Understanding the situations in which lying is acceptable is parasitic upon understanding the paradigm case of lying being impermissible. The default position is that lying is impermissible, and any deviations stand in need of explanation. Lying within a particular game (such as Diplomacy) is permissible, for instance, because we have agreed to play a certain game, and we explain the game in part by explaining that the normal expectation of honesty is inappropriate. That is, you need to lie to play the game and you need to expect other people to lie to play the game. We only understand the game where we recognise it as an exception to the paradigm case in which lying is understood as impermissible.

The proposal is to reconceive the notion of 'moral principles' such that we allow that moral principles are tolerant of exceptions just in case the exceptional case occurs under conditions that are not privileged, either through being a morally inferior situation to start with, or being a case that is only understood *as an exception* to the

---

something appears to be a cup is a reason for us to believe that it is a cup. Nothing here precludes the possibility that, in some contexts (such as if one were in a hall of holograms), the very same reason, something appearing to be a cup, would not yield this conclusion. However, no one would seriously claim that the relevant generalisation, in that it admits of exceptions, is of no value or use. If queried about why I believe something is a cup, it appears to be a genuine justificatory and explanatory move to reply that it looks like a cup. It is something of a puzzle to Lance & Little why *moral* generalisations are held (under the 'classical' conception) to require exceptionlessness, whereas in epistemology more generally this requirement is clearly overly onerous.

principle. On this approach, Lance & Little aver, one can assign a genuine justificatory/explanatory role for moral principles provided one did not operate with the 'classical' conception of moral principles. One could still, as it were, point to the fact that some action constituted a *lie* in order to explain or justify the moral judgement that that action was impermissible. One would need to check that the case was not an exceptional case itself, and fell within the scope of the privileged conditions modifier, but at first sight at least the moral principle looks to be performing some genuine explanatory and/or justificatory function.

Evidently, much more needs to be said here concerning the nature of Lance & Little's proposal, especially concerning the contrast with alternative proposals such as Pekka Väyrynen's 'hedged principles' approach.[187] What I should like to do is draw attention to a further feature of Lance & Little's proposal that, as we shall see, makes it a natural ally to my alternative approach to reflective equilibrium. Moral principles are, they maintain, to be understood via reference certain paradigm cases that represent privileged conditions. However, "Interpretation arises in understanding what counts as a paradigm example, what counts as an acceptable deviation from that paradigm, and what follows from the way that an acceptable deviation deviates." (Lance & Little 2004, p.452) It is entirely possible, then, that different evaluators will agree upon a certain moral principle but, in virtue of having different paradigm examples in mind, will nonetheless come to different conclusions when faced with particular examples. The application of moral principles *qua* defeasible generalisations will not only be exception-laden, but *non-universal* as well. We would not expect all agents to come to the same judgement in identical situations using the same moral principle.

[187] Somewhat confusingly, and perhaps more indicative of the limitations of applying terms with the suffix '~ism' in philosophy, Väyrynen regards Lance & Little's 'defeasible moral generalisations' approach as a variant of a 'hedged principle' approach, which is itself construed as a way to preserve moral generalism. (Väyrynen 2006, p.727n.51)

Väyrynen's own 'hedged principles' approach is developed in his 'A Theory of Hedged Moral Principles' (Väyrynen 2009), in which moral principles tolerate exceptions. This account is (very roughly) as follows: in cases where there a moral principle tolerates exceptions, it is because moral principles generally work in virtue of some underlying moral-value-instantiating property, the 'normative basis', for that principle. This means that the moral principle itself can tolerate exceptions. By way of example, take the principle that 'Lying is impermissible'. If there is a 'normative basis' for this principle, then the principle does explanatory/justificatory work in virtue of this underlying 'normative basis'. Exceptions, however, occur in those situations where an action is a genuine case of lying, and yet does not fit the criteria for when the underlying 'normative basis' kicks in.

This is exactly the sort of reformist proposal for the role of 'moral principles' that my alternative understanding of the method reflective equilibrium requires. On the standard approach to reflective equilibrium, whether it is correct to regard an agent as holding their moral conception in reflective equilibrium depends upon the 'fit' between their considered moral judgements and the specific judgements that would follow from the set of moral principles in their moral conception. However, this relies upon the implicit assumption that moral principles would, if applied correctly, be applied *universally* (or near-universally). I've argued that the following is a more persuasive analysis of what occurs in many cases (such that they are sufficiently prevalent) within the moral realm: agents, in virtue of having different sensitivities to situations, will tend to apply their moral principles in a *non-universal* manner such that moral disagreements arise. Furthermore, these disagreements are opaque, such that we have no means of identifying, even in principle, which party or parties to the disagreement have made some sort of error. Evaluators, in virtue of having different sensitivities to situations, will tend to diverge in their estimations of whether any given agent holds their moral conception in reflective equilibrium. These epistemological disagreements are also opaque disagreements.

Given my argument, then, I am committed to the rejection of the classical notion of 'moral principles' as universal, exceptionless, law-like generalisations. However, like reformist particularists, I believe that one may reconceive the notion of a 'moral principle' such that one can still retain a significant, i.e. justificatory and/or explanatory, role for moral principles and moral theory. The reformist proposal is actually fairly simple: in proposing some set of moral principles, we need to regard them as necessary, but not sufficient, for ensuring universality (or near-universality). Specifically, we need to bear in mind the impact that variable sensitivities may have, and not maintain the *expectation* that the pattern of application for a moral principle will be universal.

There will be cases, for instance, in which the impact of sensitivity considerations is minimal because the relevant sensitivities *are themselves* universal (or near-universal). Our sensitivities with respect to the concept of *pain,* for instance, may enjoy such a status. On such occasions, a moral principle, such as 'One ought not inflict easily-

avoidable pain on others' plays a justificatory and explanatory role when it comes to specific moral judgements about the moral value of φ.

However, there will also be cases, such as in my two examples, where the relevant sensitivities diverge, and opaque moral disagreements arise as a result. Here, it is part of our explanation for the disagreement itself that moral principles are not necessarily applied in a universal way. It is part of the justification and explanation of A's judgement that φ, say, eating meat, is impermissible because of the presence of the moral principle that 'One ought not eat one's fellow creatures'". However, the relevant moral principle is not seen as applicable by B, and plays no justificatory or explanatory role. Nonetheless, it would be impossible to understand the nature of this disagreement without recognising the justificatory and explanatory role of moral principles. What this suggests is that moral principles, on my approach, play a genuine justificatory and/or explanatory role. The relevant difference between my approach and the standard approach to reflective equilibrium is that, for any given moral judgement to be justified or explained by a moral principle, my approach requires that the agent needs to have a sensitivity such that the relevant concern picked out by the moral principle is seen by them as applicable in the specific situation. Moral principles, *in conjunction with sensitivities*, play a genuine justificatory and explanatory role.

Here, then, is how one ought to understand the method of (wide) reflective equilibrium on my alternative view: one can take it as a *methodological presumption* that, for any given moral disagreement, that disagreement will reflect some differences in the underlying rational structure of the parties to the disagreement.[188] However, we need to be open to the possibility that we might nonetheless reach a point where that structure has been laid bare (using appropriate moral principles and theory), and what we are faced with are differences in our sensitivities to key concepts involved in the relevant moral principles. The activity of theorising in ethics needs to be balanced against an awareness of the potential confusion created through different evaluators having different sensitivities, operating with subtly variant understandings of how

[188] This *methodological presumption* in favour of theory contrasts with the *theoretical presumption* of §5.5, in which it is assumed that moral disagreement *must* bottom out in differences between the rational commitments (i.e. the moral conception and relevant non-moral beliefs) of the parties to a disagreement.

one would *blindly* apply a given concept. As we shall see in the next section, this improves the method of reflective equilibrium by making those who adopt the method more responsive to potential sources of moral disagreement.

The answer to the anti-theory worry, then, is that my alternative interpretation of the method of (wide) reflective equilibrium does not relegate moral principles (and thus moral theory) to solely performing a pedagogical or heuristic function, but does nonetheless require us to reform the notion of 'moral principles'. One may still regard moral principles as performing justificatory or explanatory roles, but one ought not expect that specifying moral principles in the absence of sensitivity considerations alone will guarantee universal patterns of application. As such, it is incompatible with the universalist aspirations of the 'classical' conception of moral principles, and in that sense the argument shares affinities with reformist particularist arguments. In particular, it may be seen as a natural ally to the privileged conditions account of moral principles as defeasible generalisations put forward by Lance & Little. However, the point to stress is that my own argument does not preclude the use of moral theorising, and the use of moral principles. Without presupposing the classical conception of 'moral principles', there is no inconsistency in recognising the importance of sensitivities whilst recommending a method of moral reasoning that presumes in favour of theoretical considerations.

6.3 – The Objectivity Worry

My argumentative position is not, as we've seen, internally inconsistent. The next worry, however, concerns the merits of the *method* of (wide) reflective equilibrium (on my alternative approach) itself. One can see, then, that stressing the importance of sensitivities, *blind* inclinations to regard an action as a case of *cruelty, instrumentalisation, eating one's fellow creatures,* etc. requires, on my alternative approach to the method of (wide) reflective equilibrium, that we engage in a process of seeing how far moral theory can take us, but being prepared to recognise where we have reached 'bedrock', and are faced with a disagreement stemming from differences in sensitivity. The worry here is that such a method is *impotent* in the face of moral disagreement.

Here's how one might formulate the objection: On the standard interpretation of the method of (wide) reflective equilibrium, the method offered us some philosophical hope in the face of persistent moral disagreement. It offered us the hope of giving the lie to, for instance, Alasdair MacIntyre's infamous allegation that the modern moral epistemological predicament is that of "pure assertion and counter-assertion." (MacIntyre 2000, p.8) By explicating the notion of 'being justified' using some ideally exact REFLECTIVE EQUILIBRIUM explicatum (to the exclusion of others), we would have a method that would transparently render an important question in moral theory acceptance disagreement-entailing-error objective. A person's 'daffy' moral conception could be transparently revealed, using this explicatum, to not be in REFLECTIVE EQUILIBRIUM, and we would thereby have some transparently objective consideration that would justify our either attempting to convince them otherwise, or quietly ignoring them, much as we would an Intelligent Design theorist. If we are to give up on this explicative project whereby 'reflective equilibrium' could function as a transparently disagreement-entailing-error objective matter and settle questions of moral theory acceptance in a definitive manner, what function could engaging in the method of reflective equilibrium play? It does not even *aim at* producing justifiably-held moral theories, where 'justifiably-held' reflects a transparently objective matter.

This worry is, I believe, premature. It's worth noting here that Daniels offers some remarks in favour of the method of wide reflective equilibrium that can be extricated

from the standard view that 'wide reflective equilibrium' functions primarily as picking out an epistemic state of coherence:

> I have suggested that seeking wide reflective equilibrium may render problems of theory acceptance in ethics more tractable and may thus produce greater moral agreement. Specifically, it may lead us to understand better the sources of moral agreement and disagreement and the constraints on what we count as relevant and important to the revision of moral judgements. It may allow us to reduce moral disagreements (about principles or judgements) to more resoluble disagreements in the relevant background theories. None of these possibilities guarantees increased agreement. How much convergence results remains an empirical question. But I think I have made it at least plausible that wide equilibrium could increase agreement and do so in a *nonarbitrary* way. (Daniels 1979, p.34)

There are two distinct claims here. Firstly, widespread adoption of the method of wide reflective equilibrium would plausibly lead to *convergence*. Although it is impossible to say in advance of such an eventuality just how much convergence would result, it is plausible that recognising the need to make one's moral conception 'fit' with the best available background theories could well lead to the rejection of certain moral conceptions. Secondly, Daniels maintains that the widespread adoption of the method of wide reflective equilibrium would be a good *diagnostic tool*. It would allow us to better understand sources of moral disagreement, by revealing the underlying rational structure to the specific moral judgements made by the parties to the disagreement.

I believe we can adapt both claims in response to the objectivity worry. The first claim has some plausibility, even once one takes into account sensitivity considerations. There is more than one way in which a person's moral conception can be 'daffy', and not all of them can be persuasively rescued by an appeal to sensitivity considerations. Various moral conceptions, for instance, implicitly rely upon empirical assumptions that are *demonstrably* false. It is worth reminding ourselves, for instance, that we object to inherently sexist or racist ideological commitments, not merely on *moral* grounds, but often also because they are simply not borne out by the empirical evidence.[189] Whilst, as Daniels himself argues,[190] it

[189] It is plausible that a whole host of empirical assumptions, such as the common empirical assumption that women are more talkative than men (see ft.55), underlie moral conceptions in which highly significant differences in terms of the assignment of duties and what is regarded as permissible

would be naïve to suggest that such commitments will inevitably fade once either falsifying empirical information (or the realisation that the commitment is ill-evidenced) is available, it is not likewise naïve to assume that the long-run implication will be *some* convergence away from moral conceptions that have difficulty accommodating available evidence.

The point here is that acknowledging the possibility of opaque disagreements stemming from differential sensitivities within the moral sphere in no way commits us to the thought that the method of (wide) reflective equilibrium, on my alternative view, is useless in the face of *any* moral disagreement. It is an advantage of Daniels' exposition that he stresses the importance of background theoretical considerations, and there appears to be no reason why one cannot simply adopt this methodological recommendation, shorn of any commitment to achieving a distinctive epistemic state of coherence.

It is the second claim, however, that it is particularly interesting: the thought that the method of wide reflective equilibrium, once widely adopted, becomes of use as a *diagnostic tool*.[191] The method is of use in allowing one to be responsive to the *sources* of moral disagreement. Daniels canvasses the following possible sources, which he takes to be exhaustive, of moral disagreement that may emerge following the uptake of the method of wide reflective equilibrium:

1. Differential background theoretical information – agents may have informational asymmetries. One person, for instance, may be completely unaware of relevant empirical information that would constrain, or even force them to abandon, their initial moral conception.

---

behaviour are built upon empirical assumptions without an empirical basis, or sometimes even *in spite of* the empirical evidence.

Another example is any moral conception that traded on the notion of 'race' as a biological category, in spite of the flagrant implausibility of such a notion being explicable in a naturalistic enquiry. (Appiah & Guttman 1998, pp.71-4) Such moral conceptions are not merely *morally objectionable*, but are objectionable on empirical grounds.

[190] See ft.198.

[191] Note that regarding the method of reflective equilibrium as a diagnostic tool remains compatible with Scanlon's point (see §1.2) that "The process of seeking reflective equilibrium is something we each must carry out for ourselves." (Scanlon 2003, p.149) The method becomes diagnostic when it is widely adopted by various agents.

2. Differential 'starting-points' (largely culturally-inherited) – agents may vary either in their considered moral judgements, or in their moral principles that they are initially inclined to accept.
3. Errors in reasoning – agents may fail, for various reasons such as cognitive bias, partiality, lack of time, etc., to follow through the implications of their commitment to a certain moral principle, and not realise that it is incompatible with various considered moral judgements or other moral principles. They may also fail to realise the implications of various background theoretical considerations.

According to Daniels, then, the adoption of the method of wide reflective equilibrium would reveal which of these three sources of moral disagreement is present. Engagement in the method of wide reflective equilibrium, one would expect, would tend towards the removal of disagreement stemming from errors in reasoning, or differences in relevant background theoretical information, but it may well be the case that, due to having different starting-points, the most that could be hoped for would be convergence on a *range* of equally acceptable moral theories that could be held in a state of wide reflective equilibrium. However, the methodological presumption here is that moral disagreement is often structured. It is rarely the case that one will simply have a moral disagreement that really is just a matter of 'pure assertion and counter-assertion', and cannot be illuminated by the respective parties to the disagreement engaging in some moral theoretical work (as recommended by the method of wide reflective equilibrium). There may be a point at which relevant empirical considerations do not serve to arbitrate any dispute. The facts may be in and yet it is still possible to justifiably hold various moral conceptions that are mutually incompatible. However, one would not be able to say in advance of the respective parties to the disagreement carrying out the method of wide reflective equilibrium. As such, the method of wide reflective equilibrium may serve as a diagnostic tool, and allow us to better understand the sources of moral disagreement.

On my alternative understanding of the method of wide reflective equilibrium, however, to these three possible sources of moral disagreement, we need to add the following:

4. Differential sensitivities – agents may vary in their *blind* responses to relevant concepts utilised in moral principles. As such, they may accept the same moral principle and yet come to divergent judgements in particular cases.

Such cases emerge where agents, engaging in the method of (wide) reflective equilibrium go as far as they can go with moral theoretical concerns, and arrive at moral principles where the relevant concepts are *blindly* applied[192] (or not) given specific situations and actions. The idea here is that, the method of (wide) reflective equilibrium, by being modified in such a manner, allows us to be sensitive to a different source of moral disagreement that we would have previously overlooked. *Without* anticipating that such cases might arise, the method of (wide) reflective equilibrium would tend us towards misdiagnosing these cases of moral disagreement as reflecting some difference in the rational commitments of the parties to the disagreement. Faced with a case where another agent, apparently sharing my set of moral principles and yet deriving different moral judgements from them, I would first of all check whether some sort of error in reasoning had been committed or whether there was a relevant background belief that explained the disagreement. Failing this, on Daniels' method of wide reflective equilibrium, I would be forced to conclude (given the theoretical presumption) that, in spite of appearances, they did not *share* my set of moral principles, and that further theoretical work would be needed to reveal the difference between our moral conceptions. As such, I may well be in the situation of 'over-rationalising' the case: failing to recognise that, in virtue of our having different sensitivities, we are inclined to apply the *same* moral principle in subtly different ways.

On the alternative understanding of the method, however, it is only a methodological presumption, when faced with moral disagreement, to search for underlying differences in our rational commitments, but nonetheless recognise that there may well be cases of moral disagreement where it is more persuasive to regard them as stemming from variable sensitivities. As such, the method of wide reflective equilibrium becomes a way to more accurately depict the sources of moral disagreement. Just as with Daniels' understanding of the method of wide reflective

[192] To remind the reader (see §4.3), to apply a concept *blindly* is to apply it non-inferentially, such that it would be inappropriate to attempt to seek out some underlying rational structure to the application of the concept.

equilibrium, one is not in a position to know in advance of carrying out the method whether we are in such a position or not. I have indicated that it is plausible to think that there will be such cases, and that they will be common enough to give the lie to the idea that 'reflective equilibrium' could pick out a distinctive epistemic state. However, that is not to say that any given moral disagreement will bottom out in differences in sensitivity.

We can, then, adapt for my alternative view Daniels' two interrelated claims that the widespread adoption of the method of wide reflective equilibrium would plausibly lead to *convergence* and *diagnostic utility*. With regards to the first, we've seen that the possibility that a moral disagreement might be an opaque disagreement stemming from variability in sensitivity does not preclude the possibility that it may be more simple than this, and that one party to the disagreement may be reliant on ill-evidenced or demonstrably false empirical assumptions. It is thus plausible that the method of wide reflective equilibrium on my alternative understanding would likewise lead to convergence. With regards to the second, the point here is that the alternative understanding of the method of (wide) reflective equilibrium is an *improvement* on Daniels' method in terms of *diagnostic utility*, since the adoption of the alternative method opens up the possibility that moral disagreement stemming from variability in sensitivity will be recognised as such, rather than leading to over-rationalised moral theories.

How may we then answer the objectivity worry with which we started this section? Well, the worry stems from a confusion between, if you like, the reality on the ground in the moral realm, and methodological recommendations for moral reasoning. My argument is that there is no reason, peculiar to morality, why we cannot design a REFLECTIVE EQUILIBRIUM explicatum that could render the question of whether some agent is justified in holding their moral conception a transparently disagreement-entailing-error objective matter. It is simply that, in terms of the reality on the ground, there is a prevalence of opaque disagreements stemming from variability in sensitivity. This in turn creates opaque disagreements as to whether any given agent holds their moral conception in a state of coherence or not. Given the reality on the ground, then, it is at best *misleading* to think of a reflective equilibrium

approach as able to render the question of whether some agent is justified in holding their moral conception a transparently disagreement-entailing-error objective matter.

The point is, then, that the standard approach to the method of reflective equilibrium, in that it thinks of reflective equilibrium as an epistemic state of coherence, would lead us astray under such conditions. This, however, is not the case if we augment the method such that there is only a methodological presumption in terms of seeking underlying moral principles in the face of moral disagreement where no obvious error or difference in empirical information emerges. On my alternative approach, the method of (wide) reflective equilibrium is better suited to the reality on the ground.

Here, the objectivity worry really only amounts to the protest that we do not like the reality on the ground, and that we'd prefer it if it wasn't the case that there was variability in sensitivity.[193] It is not appealing that we'd end up with a method of moral reasoning such that there are situations in which very little, short of methods of conversion or persuasion, can be done to arbitrate the dispute. The reply here is simply that these areas of disagreements areas never were such that we would ever be in a situation to identify which party/parties to the disagreement is/are in error. That is the reality on the ground (or, at least, that is a more persuasive analysis of the reality on the ground). The only question then is whether one adopts a methodology that, once widely adopted, would allow people to recognise such sources of moral disagreement.

One might even make an 'opportunity cost' point here. If the reality on the ground is as I've claimed such that there is a sufficient prevalence of opaque disagreements stemming from variability in sensitivity, then it represents a significant opportunity cost to expend considerable efforts in attempting to clarify areas of moral disagreement by invoking further underlying rational structure to the disagreement. That effort could be better spent understanding the nature of the relevant parties'

[193] Obviously, one could claim that my analysis of the reality on the ground is unpersuasive, but one would need to be careful to do so in such a way that one avoided an *a priori* commitment to moral concepts (and thus moral principles) being applied universally, since that position is vulnerable to the argument (in §5.5) stemming from the rule-following considerations that there must be cases in which we proceed *blindly*.

conceptual sensitivities, such that the nature of the disagreement could be better understood.

All the while, it is plausible to think that adopting the method of (wide) reflective equilibrium, even once we incorporate sensitivity-style considerations, will improve this reality on the ground. It remains the case that adoption of the method would allow an agent to revise their 'daffy' moral conception if that moral conception relied upon ill-evidenced or demonstrably false empirical assumptions. It is plausible to think that, even on my alternative construal of the method of (wide) reflective equilibrium, significant *convergence* could well take place.

6.4 – The Vacuity Worry

We've seen, then, that my approach does not involve the rejection of the use of moral principles and retains a (more modest) role for moral theorising. At this point, however, a quite different worry crops up. On my alternative approach to the method of (wide) reflective equilibrium, I'm recommending a certain method in moral matters independently of the coherence theory. However, an objection that occurs within the secondary literature is that the method of wide reflective equilibrium, whether tied to this theory or not, fails to pick out a distinctive methodology at all, but is rather an empty or vacuous proposal.[194] As such, recommending the 'method' of wide reflective equilibrium, whether Daniels' version or my alternative version,[195] is little more than recommending that a person thinks though their moral commitments, check them for consistency, especially with relevant empirical evidence, which is hardly a distinctive methodological proposal.

Let's elaborate this vacuity worry in more detail by running through the steps leading to the worry. One can see that the method of *narrow* reflective equilibrium is distinctive, since it calls upon a person to reflect upon their considered moral judgements, attempt to systematise them using some set of moral principles, then, by moving back and forth, achieve a satisfactory fit between one's principles and judgements. The standard objection here is that such a method does not much resemble a process of justifying one's moral conception.[196] Justifying one's moral views often takes the form of invoking relevant empirical evidence, or arguing against specific alternative conceptions, and the method of narrow reflective equilibrium has no role for such activity. As such, the method of narrow reflective equilibrium is inherently conservative. The method of *wide* reflective equilibrium, however, arguably avoids such inherent conservatism as the process incorporates

---

194 Singer (Singer 2005, p.347; p.349) puts forward this criticism of *wide* reflective equilibrium methodology, and replies are canvassed by Scanlon (Scanlon 2003, pp.150-1) and Tersman (Tersman 2008, pp.398-400). DePaul (see ft.178), by contrast, by-and-large accepts the criticism.

195 In this section, I shall, for the main part, focus on the vacuity worry as it concerns Daniels' method of wide reflective equilibrium, before indicating how things stand on my own alternative interpretation of the method.

196 This criticism is essentially a *process-focussed* version of the garbage-in/garbage-out objection. This objection ran that the resultant epistemic *state* of being in reflective equilibrium did not sufficiently resemble the epistemic state of being justified in one's moral conception. Here, the objection is that the methodology of reflective equilibrium does not sufficiently resemble a method of justifying one's moral conception.

relevant philosophical arguments, alternative moral conceptions, and background theoretical considerations.

The vacuity worry surfaces at this point. Once one incorporates such considerations into the reflective equilibrium process, it is difficult to see what the method of wide reflective equilibrium *rules out*. One could argue that 'foundationalist' approaches to moral theory construction become a 'limiting case' (Singer 2005, p.347) of reflective equilibrium. On one end of the spectrum, one has a judgement-heavy methodology, where one's reflections are dominated by the attempt to systematise one's considered moral judgements; on the other end of the spectrum, one has a principle-heavy methodology, where one's reflections are heavily dominated by principles and background theories. 'Foundationalist' approaches are simply the latter limiting case. The overall problem here is that, even supposing one accepts that the method of wide reflective equilibrium is appropriate to moral reasoning, it is unclear that there are any particular ways in which one could possibly proceed that would be incompatible with that method. As such, it is 'empty' or 'vacuous' as a methodological doctrine.

Since we find an attempt to answer this charge in Scanlon's 'Rawls on Justification', his reply seems an obvious place to start. Scanlon regards this objection as 'largely correct', but offers the following considerations to mitigate how damaging the objection is:

> What the method of [wide] reflective equilibrium prescribes is, so to speak, a level playing field of intuitive justification on which principles and judgments of all levels of generality must compete for our allegiance. It thus allows all possible sources of justificatory force to be considered. But the method is not vacuous because it is incompatible with some views about these sources. It is incompatible, first, with the idea that any particular class of judgments or principles can be singled out in advance of this process as justified on some other basis and, second, with the idea that any class of *considered* judgments should be left out of this process (for example that "intuitions" about what is just or unjust in particular cases should not be given any weight in justifying general principles but must be derived from them.) (Scanlon 2003, p.151)

If one were to adopt the method of wide reflective equilibrium, Scanlon argues, one could *not* proceed in either of the following ways. Firstly, one could not go into the wide reflective equilibrium process with, as it were, any sacred cows. There are no

moral principles or judgements that are not in principle revisable, or that retain justificatory force independently from carrying out the method of wide reflective equilibrium. Scanlon here maintains that a wide reflective equilibrium methodology can incorporate 'foundationalist' appeals to 'self-evidence' or 'incorrigibility' within a wide reflective equilibrium methodology, but only if such 'self-evidence' or 'incorrigibility' is established *within* that reflective equilibrium process.[197] (Scanlon 2003, p.151) As such, it is clearly incompatible with the method of wide reflective equilibrium to maintain, for instance, that one's moral conception *must* incorporate any particular moral principle or judgement, be they regarded as divine commands, or essentialist claims about human flourishing, or the result of some special capacity for intuiting moral truths, etc. It is a level playing field.

Secondly, Scanlon argues that one could not 'leave out of the process' the use of moral intuitions (in the form of considered moral judgements), as Brandt and Singer recommend. Here, however, Scanlon's point misfires somewhat. It is, at least on the face of it, completely compatible with the method of wide reflective equilibrium to avoid relying on considered moral judgements in one's own reflections. It *is* incompatible with wide reflective equilibrium to, as it were, *ban* anyone from using considered moral judgements in their reflections, but it is not incompatible to *avoid* using them oneself. By analogy, it is compatible with taking a permissive attitude to the consumption of alcohol to oneself remain abstinent, but it is incompatible with that permissive attitude to take the view that everyone (morally) ought to be likewise abstinent.

Here, I think, a stronger point is needed, one along the lines of claiming that one *cannot avoid* reliance on considered moral judgements (or intuitions more generally) in one's moral reflections. The claim here would be, as Singer puts it, that "without intuitions, we can go nowhere." (Singer 2005, p.349) This is, notably, exactly how Daniels responds to Brandt's injunction to eschew ethical intuitions. Brandt's intention in *A Theory of the Good and the Right* is to follow a method of theory

[197] To my mind, Scanlon's remark here seems odd. This may be something of a verbal quibble rather than a substantive disagreement, but I find it clearer to maintain that any appeals to 'self-evidence' or 'incorrigibility' are incompatible with the method of wide reflective equilibrium, since such appeals are more naturally regarded as *ruling out* the possibility of revision, which is itself integral to the method of wide reflective equilibrium. The point, however, is the same: it is incompatible with the method of wide reflective equilibrium to bring any sacred cows, anything that is in principle not able to be revised, into that method.

construction that would involve "stepping outside our own tradition somehow" (Brandt 1998, p.21), and relying on facts and logic alone. Brandt operates with a notion of a 'rational person' understood along the lines of someone "maximally influenced by evidence and logic." (Brandt 1998, p.11) From here, Brandt advocates a method of 'cognitive psychotherapy', (Brandt 1998, chs.5-6) whereby the intention is to delineate a set of desires that only a 'rational person', in Brandt's sense, would have (i.e. a theory of the good), and then one can design a moral theory or social moral code (Brandt 1998, p.195) that would be best for the satisfaction of such desires (i.e. a theory of the right).

Daniels claims in response that Brandt is operating under an illusion if he believes such a response genuinely avoids relying on prior moral intuitions. Daniels grants that Brandt does not *explicitly* appeal to moral intuitions or judgements in his method of delineating desires that only a rational person (i.e. someone maximally influenced by evidence and logic) would have, but some implicit appeal to moral intuitions is nonetheless present.[198] Daniels seems to regard this argument as illustrative of the larger lesson that one cannot avoid appeal to intuitions in moral reasoning or, at least, no one has as of yet demonstrated that such a thing is tenable:

> There is no real escape from prior moral "intuitions" after all. Our rationalized desires are not a bedrock of morally neutral facts. It is therefore better to lay all our moral cards on the table, where they can be assessed, as wide reflective equilibrium proposes, than to pretend that we can end up justifying moral beliefs without appealing in any way to other moral beliefs.
>
> If there were a better alternative to appealing to our moral judgments and then criticizing them as much as we can, then we should consider it. We seem to lack, however, a plausible alternative to appealing to some moral judgments. (Daniels 1996, p.5)

[198] Essentially, the argument is that, even after a process of 'cognitive psychotherapy', it is likely that the set of 'rationalised desires' one ends up with will still be shaped by the peculiarities of one's particular culture and social institutions. Someone growing up in a caste system, may nonetheless expect and desire, for instance, to be treated preferentially according to their caste, even after 'cognitive psychotherapy'. Brandt would of course argue that such a desire would not survive 'cognitive psychotherapy', as it would seemingly rely upon empirical assumptions of caste superiority or whatnot, but here Daniels notes that "ideologies are highly resistant to extinction merely through exposure to "all the relevant facts"." (Daniels 1985, p.93) Brandt's notion of what a 'rational person' would desire is either truly procedural, in which case it would appear difficult to sustain the thought that every ideology we find egregious would fall by the wayside and the desires will be shaped by one's cultural upbringing, or it contains some implicit moral intuitions, such that it is part of our understanding of a 'rational person' that they do not subscribe to egregious ideologies, in which case the method is itself moral theory-laden. Either way, the set of 'rationalised desires' will not be free from some implicit moral intuitions.

It seems reasonable, then, to attribute to the *method* of wide reflective equilibrium the commitment that one recognises that one cannot avoid reliance on some moral intuitions in moral theory construction. As such, it is incompatible with the method of wide reflective equilibrium to proceed in constructing a moral theory with the view that one's construction represents some sort of *neutrality*. The method of wide reflective equilibrium requires us, to use Daniels' phrase, to 'lay all our moral cards on the table' and make clear where one is appealing to one's considered moral judgements. One would not be engaging in the method of wide reflective equilibrium if one maintained that one was not appealing to any sort of moral intuition in one's reasoning.[199]

Although Scanlon does not explicitly mention this in connection with the vacuity charge, there is another aspect of the method of reflective equilibrium worth mentioning, since it increases the distinctiveness of the method: the requirement that one's 'intuitions' take the form of *considered* moral judgements. This is a substantive constraint that one might well question.[200] According to the method of wide reflective equilibrium, we are only to include those intuitions in the wide reflective equilibrium process that are relatively stable and confidently held, and that are made under good epistemological conditions (i.e. good information, impartiality, no temporally heightened emotional state, etc.). It would be incompatible with the method of wide reflective equilibrium to reason from intuitions that violated such conditions. One could not, for instance, argue for the moral permissibility of revenge on the basis of one's firmly-held (at the time, at least) moral intuition that a person who hurts someone close to you deserves to be hurt in return, since it is likely that this intuition would fail to display the requisite stability to be a considered moral judgement.

---

[199] I ought to make it clear that my point here is only that the method of wide reflective equilibrium is genuinely incompatible with certain methodological recommendations. As I noted earlier (see ft.178), I am not convinced by this form of response to Singer's or Brandt's attack on the use of moral intuitions *tout court*. Space precludes further investigation of the matter; this is only a *sketch* of how one might advocate reflective equilibrium methodology, not a fully-fledged defence.

[200] As does Raz, for instance, in 'The Claims of Reflective Equilibrium'. (Raz 1982) Raz objects that, if we take at face value Rawls' emphasis on moral theory as a kind of moral psychology, it does not appear sensible to exclude rash or ill-considered moral intuitions, as these surely constitute a part of a person's 'moral sense' as much as considered intuitions. The response here, as we saw in §1.2, is to note that the descriptive interpretation of the method of reflective equilibrium is merely a provisional description that in the end is supplanted by the deliberative interpretation, and then "the rationale for concentrating on *considered* judgments is that these are the most likely to be correct judgments about their subject matter (morality, or justice)." (Scanlon 2003, p.142)

Thus far, I have, broadly following Scanlon, mentioned three aspects of the method of wide reflective equilibrium that demonstrate that, purely considering the method of wide reflective equilibrium, the method is not vacuous. The method excludes anyone from having sacred cows and claiming 'self-evidence' or 'incorrigibility' for a particular moral principle or judgement, it requires a certain amount of intellectual transparency such that it excludes anyone purporting to avoid reliance on ethical intuitions *tout court*, and it excludes anyone who reasons from ill-considered or rash ethical intuitions. These reflect substantive constraints on how one proceeds, especially constraints at the epistemological level regarding how one ought to conceptualise the process of justifying a moral conception.

I have, in this section, referred to the method of wide reflective equilibrium as conceived of by Daniels, and here we can see that it is distinctive. As we've seen, however, I have been advocating an alternative view of the method of (wide) reflective equilibrium in which we drop the theoretical presumption that moral disagreement must always bottom out in some difference in the rational commitments held by the parties to the disagreement. On my alternative understanding, delineating the underlying rational structure behind a particular moral judgement only goes so far. There is, if you like, a residuum created by the fact that, at some point, the invocation of moral principles relies upon our taking one's sensitivities, one's *blind* inclinations to respond to situations by invoking certain concepts, as *prima facie* legitimate. As such, the *distinctiveness* of the method increases: it is incompatible with the theoretical presumption.

I conclude that my alternative approach to the method of (wide) reflective equilibrium is not vulnerable to the worries I've considered. There is no internal inconsistency involved in recommending the method of (wide) reflective equilibrium (alternatively construed) given my argument's trading on the non-universal application of moral principles, since, as Lance & Little have shown, we might adopt a reformist conception of 'moral principles' such that we relinquish the idea that moral principles need be universal, exceptionless, law-like generalisations. By moving to my alternative approach, in which there is a methodological presumption in terms of seeking moral principles that genuinely justify and explain moral judgements, but such that implicitly recognises that it will not always be the case that any given case

admits of further theoretical illumination, we can also see that the adoption of (wide) reflective equilibrium (alternatively construed) would plausibly still lead to convergence, whilst also being considerably better as a diagnostic tool for understanding the sources of moral disagreement. This methodological alteration also helps put pay to the idea that recommending the method of wide reflective equilibrium is an empty gesture, since the recommendation is incompatible with various alternative methods one might employ.

Glossary

Basic/*Blind* Cases of Rule Following – In Wright's terms, "Cases where rule-following is *uninformed by anterior reason-giving judgement.*" (Wright 2007, p.496) Basic/*blind* cases are ones in which it would be inappropriate to apply the *modus ponens* model. See §4.3

Basic Concepts – Basic concepts are concepts where the pattern of correct usage is not inferred from further concepts. They represent basic/*blind* cases of rule-following. See §4.3

Considered Moral Judgements – Moral intuitions that have passed through a filter to rule out rash or ill-considered moral intuitions. In Rawls' words, "Considered judgements are simply those rendered under conditions favourable to the exercise of the sense of justice, and therefore in circumstances where the more common excuses and explanations for making a mistake do not obtain." (Rawls 1999, p.42) See §1.2

Correctness – may be used either to refer to *truth* or to *warranted assertibility* (or some variant thereof). I use the term in a neutral manner, as a placeholder term for either of these two approaches. See ft.82

Decidability, Effective – In Peter Smith's words, "A property/relation is *effectively decidable* iff there is an algorithmic procedure that a suitably programmed computer could use to decide, in a finite number of steps, whether the property/relation applies in any given case." (Smith 2007, p.9) See §2.4

Disagreement-Entailing-Error Objectivity – A matter is disagreement-entailing-error objective if and only if, in Kölbel's terms, "For all thinkers *A* and *B*: it is a priori that if *A* believes that *p* and *B* believes that not-*p* then either A has made a mistake or *B* has made a mistake." (Kölbel 2002, p.31) See ft.81

Explication – The replacement of an inexact concept (the *explicandum*) used in everyday life with a constructed alternative (the *explicatum*). See §2.3

Faultless Disagreement – A disagreement concerning a matter that is not disagreement-entailing-error objective. See ft.83

Intuitions, Moral – Moral intuitions are, in Brandt's words "Beliefs in, and dispositions on occasion to utter, certain normative statements." (Brandt 1998, p.17) However, they tend to be believed or uttered in a manner such that they are not obviously the outcome of some ratiocination. See §1.4.

Metalinguistic Claim – A claim about the meaning of a sentence in the object-language that is made in the meta-language. In 'By 'p', S means that *p',* 'p' refers to the sentence in the object-language, and the rest of the claim is made in the meta-language. See ft.103

Metalinguistic Fact – A true metalinguistic claim. See ft.93

Opaque Disagreement – A disagreement is opaque if and only if it is not rationally demonstrable that a party, or both parties, to a disagreement have made some sort of error. See §5.3

***Prima Facie* Legitimacy** – A claim is *prima facie* legitimate where, as a matter of how we proceed, we initially accept the claim, subject to future revision should it become necessary, without any call for justification or explanation.

Intended to capture Wittgenstein's thought that "To use a word without a justification does not meant to use it without right." (Wittgenstein 2001, §289) See §4.4

Rational Demonstrability – To say that some claim is *rationally demonstrable* means that, given ideal epistemological circumstances, we would be able to uniquely identify that claim as correct. See §4.2

Sensitivity – May refer to either the overall pattern of conceptual application that a person *blindly* employs when faced with different situations, or, if one is referring to one's sensitivities with regards to a particular concept, it refers to the distinctive pattern of responses in which a person would *blindly* employ that concept. See §5.2

Theoretical Presumption, the – An *a priori* commitment such that, when faced with any sort of dispute as to whether some concept (and, by extension, moral principle) applies in a particular situation, we need to reconstruct a chain of *modus ponens* reasoning such that we can clarify whether that concept (or moral principle) is correctly applied in that situation or not. See §5.5

Vagueness – May refer to semantic/constitutive or epistemological forms of vagueness. Semantically/constitutively vague concepts admits of borderline cases in their extension such that there is no standard of correctness given an appropriate inputs (e.g. 25 grains of sand is neither a heap, nor not a heap). Epistemologically vague concepts are where we lack the means, or a method, by which we could ascertain whether a concept is correctly applied given an appropriate input. See §2.4

Bibliography

APPIAH, K.A. & GUTMANN, A. (1998) *Color Conscious: The Political Morality of Race* (Chichester, Princeton University Press)

ARRAS, J.D. (2007) 'The Way We Reason Now: Reflective Equilibrium in Bioethics' in B. STEINBOCK [ed.] *The Oxford Handbook of Bioethics* pp.46-71

ARISTOTLE (1976) *Ethics* trans. J.A.K THOMSON (London, Penguin)

AYER, A.J. (1986) *Language, Truth and Logic* (Harmondsworth, Penguin)

BAIER, A. (1985) 'Doing Without Moral Theory' in *Postures of the Mind* (London, Methuen) pp.228-45

BEANEY, M. (2004) 'Carnap's Conception of Explication: From Frege to Husserl?' in S. AWODEY & C. KLEIN [eds.] *Carnap Brought Home* (Chicago, Open Court) pp.117-50

BENJAMIN, M. (2003) *Philosophy & This Actual World: An Introduction to Practical Philosophical Inquiry* (Oxford, Rowman & Littlefield)

BLACKBURN, S. (1984) 'The Individual Strikes Back' in A. MILLER & C. WRIGHT [eds.] (2002) *Rule-Following and Meaning* (Durham, Acumen)

BOGHOSSIAN, P.A. (1989) 'The Rule-Following Considerations' in A. MILLER & C. WRIGHT [eds.] (2002) *Rule-Following and Meaning* (Durham, Acumen)

BOGHOSSIAN, P.A. (2005) 'Is Meaning Normative?' in A. BECKERMANN and C. NIMTZ [eds.] *Philosophy – Science – Scientific Philosophy* (Paderborn, Mentis)

BONJOUR, L. (1976) 'The Coherence Theory of Empirical Knowledge' in *Philosophical Studies 30*, pp.281-312

BONJOUR, L. (1985) *The Structure of Empirical Knowledge* (Cambridge, Harvard University Press)

BRANDT, R.B. (1990) 'The Sciences of Man and Wide Reflective Equilibrium' in *Ethics 100,* pp.259-278

BRANDT, R.B. (1998) *A Theory of the Good and the Right* (New York, Prometheus)

BRINK, D.O. (1987) 'Rawlsian Constructivism in Moral Theory' in *Canadian Journal of Philosophy 17:1,* pp.71-90

BRINK, D.O. (1989) *Moral Realism and the Foundations of Ethics* (Cambridge, CUP)

BYRNE, A. (1996) 'On Misinterpreting Kripke's Wittgenstein' in *Philosophy and Phenomenological Research 56:2*, pp.339-43

CARNAP, R. (1945) 'The Two Concepts of Probability: The Problem of Probability' in *Philosophy and Phenomenological Research 5:4*, pp.513-32

CARNAP, R. (1970) *Meaning and Necessity* (Chicago, University of Chicago Press)

CARNAP, R. (1962) *Logical Foundations of Probability* [2nd Ed.] (Chicago, University of Chicago Press)

CARUS, A.W. (2007) 'Carnap's Intellectual Development' in M. FRIEDMAN & R. CREATH [eds.] *The Cambridge Companion to Carnap* (Cambridge, CUP) pp.19-42

CARUS, A.W. (2008) *Carnap and Twentieth-Century Thought* (Cambridge, CUP)

CAVELL, S. (2002) *Must we Mean What we Say?* (Cambridge, CUP)

CAVELL, S. (2007) 'Companionable Thinking' in A. CRARY, [ed.] (2007) *Wittgenstein and the Moral Life* (London, MIT Press) pp.281-98

CHERNIAK, C. (1986) *Minimal Rationality* (London, MIT Press)

CHOMSKY, N. (1965) *Aspects of the Theory of Syntax* (Cambridge [Massc.], MIT Press)

COHEN, L.J. (1991) 'Stephen P. Stich, *The Fragmentation of Reason*' in *Philosophy and Phenomenological Research 51:1* pp.185-88

COHEN, S. (2006) *The Nature of Moral Reasoning* (Oxford, OUP)

COPP, D. (1996) 'Review: *Balance and Refinement* by Michael R. DePaul' in *Philosophy and Phenomenological Research 56:4* pp.959-62

CRARY, A. (2007) 'Humans, Animals, Right and Wrong' in A. CRARY, [ed.] (2007) *Wittgenstein and the Moral Life* (London, MIT Press) pp.381-404

CRARY, A. (2007a) *Beyond Moral Judgment* (London, Harvard)

CRISP, R. (2000) 'Particularizing Particularism' in B. HOOKER, B. & M.O. LITTLE [eds.] (2003) *Moral Particularism* (Oxford, Clarendon) pp.23-47

DANCY, J. (2006) *Ethics Without Principles* (Oxford, Clarendon)

DANIELS, N. (1979) 'Wide Reflective Equilibrium and Theory Acceptance in Ethics' in N. DANIELS (1996) *Justice and Justification: Reflective Equilibrium in Theory and Practice* (Cambridge, CUP) pp.21-46

DANIELS, N. (1980a) 'Reflective Equilibrium and Archimedean Points' in N. DANIELS (1996) *Justice and Justification: Reflective Equilibrium in Theory and Practice* (Cambridge, CUP) pp.47-65

DANIELS, N. (1980b) 'On Some Methods of Ethics and Linguistics' in N. DANIELS (1996) *Justice and Justification: Reflective Equilibrium in Theory and Practice* (Cambridge, CUP) pp.66-80

DANIELS, N. (1985) 'Two Approaches to Theory Acceptance in Ethics' in N. DANIELS (1996) *Justice and Justification: Reflective Equilibrium in Theory and Practice* (Cambridge, CUP) pp.81-102

DANIELS, N. (1996) *Justice and Justification: Reflective Equilibrium in Theory and Practice* (Cambridge, CUP)

DANIELS, N. (1996a) 'Wide Reflective Equilibrium in Practice' in N. DANIELS (1996) *Justice and Justification: Reflective Equilibrium in Theory and Practice* (Cambridge, CUP) pp.333-52

DANIELS, N. (2009) 'Reflective Equilibrium' in E.N. ZALTA [Ed.] *The Stanford Encyclopedia of Philosophy* [Summer 2009 Edition] [http://plato.stanford.edu/archives/sum2009/entries/reflective-equilibrium/]

DEPAUL, M.R. (1987) 'Two Conceptions of Coherence Methods in Ethics' in *Mind 96,* pp.463-81

DEPAUL, M.R. (1993) *Balance and Refinement: Beyond Coherence Methods of Moral Inquiry* (London, Routledge)

DEPAUL, M.R. (1998) 'Why Bother with Reflective Equilibrium?' in M.R. DEPAUL & W.RAMSEY [eds.] *Rethinking Intuition: The Psychology of Intuitions and Their Role in Philosophical Inquiry* (Oxford, Rowman & Littlefield) pp.293-309

DEPAUL, M.R. (2006) 'Intuitions in Moral Inquiry' in D. COPP [ed.] (2006) *The Oxford Handbook of Ethical Theory* (Oxford, OUP) pp.595-623

DIAMOND, C. (1979) 'Eating Meat and Eating People' in C. DIAMOND (2001) *The Realistic Spirit* (London, MIT Press)

DREBEN, B. (2003) 'On Rawls and Political Liberalism' in S. FREEMAN [ed.] *The Cambridge Companion to Rawls* (Cambridge, CUP) pp.316-346

DRETSKE, F. (1981) 'The Pragmatic Dimension of Knowledge' in M. HUEMER [ed.] (2002) *Epistemology: Contemporary Readings* (London, Routledge) pp.539-51

DUMMETT, M. (1991) *The Logical Bases of Metaphysics* (Cambridge [Mass.], Harvard)

DWORKIN, R. (1973) 'The Original Position' in N. DANIELS [ed.] (1989) *Reading Rawls* (Stanford, Stanford University Press) pp.16-52

EDWARDS, J. (1992) 'Best Opinion and Intentional States' in *The Philosophical Quarterly 42*, pp.21-33

EBERTZ, R. (1993) 'Is Reflective Equilibrium a Coherentist Model?' in *Canadian Journal of Philosophy 23:2* pp.193-214

FREEMAN, S. (2007) *Rawls* (London, Routledge)

GLÜER, K. & WIKFORSS, Å. (2009) 'Against Content Normativity' in *Mind 118:1,* pp.31–70

GOODMAN, N. (1983) *Fact, Fiction and Forecast* [4th edition] (Cambridge [Mass.], Harvard University Press)

GOLDFARB, W. (1985) 'Kripke on Wittgenstein on Rules' in A. MILLER. & C. WRIGHT [eds.] (2002) *Rule-Following and Meaning* (Durham, Acumen)

GREENE, J. D. & HAIDT, J. (2002) 'How (and Where) Does Moral Judgment Work?' in *Trends in Cognitive Sciences 6*, pp.517–23.

HALD, G.M. MALAMUTH, N.M. & YUEN, C. (2010) 'Pornography and Attitudes Supporting Violence Against Women: Revisiting the Relationship in Nonexperimental Studies' in *Aggressive Behavior 36*, pp.14-20

HALE, B. (1998) 'Realism and its Oppositions' in B. HALE & C. WRIGHT [eds.] *A Companion to the Philosophy of Language* (Oxford, Blackwell)

HARE, R.M. (1973) 'Rawls' Theory of Justice' in N. DANIELS [ed.] (1989) *Reading Rawls* (Stanford, Stanford University Press) pp.81-108

HARMAN, G. (1974) *Thought* (Princeton, Princeton University Press)

HASLETT, D.W. (1987) 'What is Wrong With Reflective Equilibria?' in *The Philosophical Quarterly 37*, pp.305-11

HAUSER, M.D. (2006) *Moral Minds* (New York, HarperCollins)

HATTIANGADI, A. (2006) 'Is Meaning Normative?' in *Mind & Language 21:2,* pp.220-40

HOLMGREN, M. (1989) 'The Wide and Narrow of Reflective Equilibrium' in *Canadian Journal of Philosophy 19:1,* pp.43-60

JONES, K. (2007) 'Moral Epistemology' in F. JACKSON & M. SMITH [eds.] *The Oxford Handbook of Contemporary Philosophy* (Oxford, OUP) pp.63-85

KAPPEL, K. (2006) 'The Meta-Justification of Reflective Equilibrium' in *Ethical Theory and Moral Practice 9* pp.131-47

KANT, I. (2005) 'Groundwork of the Metaphysics of Morals' in I. KANT (2005) *Practical Philosophy* trans. M.J. GREGOR (Cambridge, CUP)

KÖLBEL, M. (2002) *Truth Without Objectivity* (London, Routledge)

KÖLBEL, M. (2003) 'Faultless Disagreement' in *Proceedings of the Aristotelian Society 104*, pp.53–73.

KRIPKE, S. (1998) *Wittgenstein on Rules and Private Language* (Oxford, Blackwell)

KUSCH, M. (2006) *A Sceptical Guide to Meaning and Rules: Defending Kripke's Wittgenstein* (Stocksfield, Acumen)

LANCE, M.N. & LITTLE, M.O. (2004) 'Defeasibility and the Normative Grasp of Context' in *Erkenntnis 61,* pp.435-55

LANCE, M.N. & LITTLE, M.O. (2006) 'Particularism and Antitheory' in D. COPP [ed.] (2006) *The Oxford Handbook of Ethical Theory* (Oxford, OUP) pp.567-94

LANCE, M.N. & LITTLE, M.O. (2007) 'Where the Laws Are' in R. SHAFER-LANDAU [ed.] *Oxford Studies in Metaethics Volume 2* (Oxford, OUP) pp.149-171

LARMORE, C.E. (2003) 'Public Reason' in S. FREEMAN [ed.] *The Cambridge Companion to Rawls* (Cambridge, CUP) pp.368-93

LEHRER, K. (1974) *Knowledge* (Oxford, OUP)

LOAR, B.F. (1986) *Mind and Meaning* (Cambridge, CUP)

LOUDEN, R.B. (1992) Morality and Moral Theory: A Reappraisal and Reaffirmation (Oxford, OUP)

LYONS, D. (1989) 'Nature and Soundness of the Contract and Coherence Arguments' in N. DANIELS [ed.] (1989) *Reading Rawls* (Stanford, Stanford University Press) pp.141-168

MACINTYRE, A. (2000) *After Virtue* (London, Duckworth)

MACKINNON, C.A. (1984) 'Francis Biddle's Sister: Pornography, Civil Rights, and Speech' in C.A. MACKINNON (1987) *Feminism Unmodified* (Harvard, Harvard University Press) pp.164-97

MACKINNON, C.A. (1989) *Toward a Feminist Theory of the State* (Cambridge [Mass.], Harvard University Press)

MCDOWELL, J. (1979) 'Virtue and Reason' in J. MCDOWELL (1998) *Mind, Value, & Reality* (London, Harvard) pp.50-73

MCDOWELL, J. (1981) 'Non-Cognitivism and Rule-Following' in J. MCDOWELL (1998) *Mind, Value, & Reality* (London, Harvard) pp.198-218

MCDOWELL, J. (1984) 'Meaning and Intentionality in Wittgenstein's Later Philosophy' in J. MCDOWELL (1998) *Mind, Value, & Reality* (London, Harvard) pp.263-78

MCDOWELL, J. (1993) 'Wittgenstein on Following a Rule' in J. MCDOWELL (1998) *Mind, Value, & Reality* (London, Harvard) pp.221-62

MCDOWELL, J. (1996) 'Two Sorts of Naturalism' in J. MCDOWELL (1998) *Mind, Value, & Reality* (London, Harvard) pp.167-97

MCELROY, W. (2008) 'A Feminist Overview of Pornography -- Ending in a Defense Thereof' [http://www.wendymcelroy.com/e107_plugins/content/content.php?content.31]

MCGINN, C. (1984) 'Wittgenstein, Kripke and Non-Reductionism About Meaning' taken from C. McGinn (1984) *Wittgenstein on Meaning* (Oxford, Blackwell) pp.150-64 in A. MILLER & C. WRIGHT [eds.] (2002) *Rule-Following and Meaning* (Durham, Acumen) pp.81-91

MCKEEVER, S. & RIDGE, M. (2006) *Principled Ethics: Generalism as a Regulative Ideal* (Oxford, OUP)

MEHL, M.R. et al. (2007) 'Are Women Really More Talkative Than Men?' in *Science 317,* p.82.

MILLGRAM, E. (2000) 'Coherence: The Price of the Ticket' in *The Journal of Philosophy 97:2*, pp.82-93

MILLGRAM, E. (2005) *Ethics Done Right* (Oxford, OUP)

MILLER, A. (2000) 'Horwich, Meaning and Kripke's Wittgenstein' in *The Philosophical Quarterly 50*, pp.161-74

MILLER, A. (2007) *Philosophy of Language* [2nd ed.] (London, Routledge)

MILLER, A. (2007a) 'Another Objection to Wright's Treatment of Intention' in An*alysis 67:3*, pp.257-63

NUSSBAUM, M. (1995) 'Objectification' in *Philosophy and Public Affairs 24:4*, pp.249-914

PETERSSON, B. (1998) 'Wide Reflective Equilibrium and the Justification of Moral Theory' in W. VAN DER BURG & T. VAN WILLIGENBURG [eds.] *Reflective Equilibrium: Essays in Honour of Robert Heeger* (Dordrecht, Kluwer) pp.127-134

QUINE, W.V.O. (1951) 'Two Dogmas of Empiricism' in W.V.O. QUINE (1980) *From a Logical Point of View* [2nd edition] (Cambridge [Mass.], Harvard University Press) pp.20-46

QUINE, W.V.O. (1976) *Word and Object* (Cambridge [Mass.], M.I.T. Press)

RAWLS, J. (1951) 'Outline of a Decision Procedure for Ethics' in S. FREEMAN [ed.] (1999) *John Rawls: Collected Papers* (Cambridge [Mass.], Harvard University Press) pp.1-19

RAWLS, J. (1975) 'The Independence of Moral Theory' in S. FREEMAN [ed.] (2001) *John Rawls: Collected Papers* (Cambridge [Mass.], Harvard University Press) pp.286-302

RAWLS, J. (1975a) 'Fairness to Goodness' in S. FREEMAN [ed.] (2001) *John Rawls: Collected Papers* (Cambridge [Mass.], Harvard University Press) pp.267-285

RAWLS, J. (1980) 'Kantian Constructivism in Moral Theory' in S. FREEMAN [ed.] (2001) *John Rawls: Collected Papers* (Cambridge [Mass.], Harvard University Press) pp.303-358

RAWLS, J. (1996) *Political Liberalism* [paperback edition] (New York, Columbia University Press)

RAWLS, J. (1997) 'The Idea of Public Reason Revisited' in S. FREEMAN [ed.] (2001) *John Rawls: Collected Papers* (Cambridge [Mass.], Harvard University Press) pp.573-615

RAWLS, J. (1999) *A Theory of Justice* [rev. ed.] (Oxford, OUP)

RAZ, J. (1982) 'The Claims of Reflective Equilibrium' in *Inquiry 25,* pp.331-52

RAZ, J. (1990) 'Facing Diversity: The Case of Epistemic Abstinence' in *Philosophy and Public Affairs 19,* pp.3-46

RUSSELL, D.E.H. (1998) *Dangerous Relationships: Pornography, Misogyny, and Rape* (London, SAGE Publications)

SAINSBURY, R.M. & WILLIAMSON, T. (1998) 'Sorites' in B. HALE & C. WRIGHT [eds.] *A Companion to the Philosophy of Language* (Oxford, Blackwell)

SAYRE-MCCORD, G. (1996) 'Coherentist Epistemology and Moral Theory' in W. SINNOTT-ARMSTRONG & M. TIMMONS [eds.] *Moral Knowledge? New Readings in Moral Epistemology* (Oxford, OUP) pp.137-89

SCANLON, T.M. (2003) 'Rawls on Justification' in S. FREEMAN [ed.] *The Cambridge Companion to Rawls* (Cambridge, CUP) pp.139-167

SHAPIRO, S. & TASCHEK, W.W. (1996) 'Institutionism, Pluralism, and Cognitive Command' in *The Journal of Philosophy 93*, pp.74-88

SIDGWICK, H. (1981) *The Methods of Ethics* [7th Ed.] (Indianapolis, Hackett)

SINGER, P. (1973) 'All Animals Are Equal' in P. SINGER [ed.] (1986) *Applied Ethics* (Oxford, OUP)

SINGER, P. (1974) 'Sidgwick and Reflective Equilibrium' in *The Monist 58,* pp.490-517

SINGER, P. (2005) 'Ethics and Intuitions' in *The Journal of Ethics 9,* pp.331–352

SINNOTT-ARMSTRONG, W. (1996) 'Moral Skepticism and Justification' in W. SINNOTT-ARMSTRONG & M. TIMMONS [eds.] *Moral Knowledge? New Readings in Moral Epistemology* (Oxford, OUP) pp.3-48

SMITH, P. (2007) *An Introduction to Gödel's Theorems* (Cambridge, CUP)

SORENSEN, R.A. (1988) *Blindspots* (Oxford, OUP)

SPREVAK, M.D. (2008) 'Kripke's Paradox and the Church-Turing Thesis' in *Synthese 160*, pp.285–295

STICH, S.P. (1990) *The Fragmentation of Reason* (Cambridge [Mass.], MIT Press)

STROSSEN, N. (1995) *Defending Pornography: Free Speech, Sex, and the Fight for Women's Rights* (London, Scribner)

TENNANT, N. (1997) *The Taming of the True* (Oxford, Clarendon)

TERSMAN, F. (1993) *Reflective Equilibrium: An Essay in Moral Epistemology* (Stockholm, Almqvist & Wiksell)

TERSMAN, F. (2006) *Moral Disagreement* (Cambridge, CUP)

TERSMAN, F. (2008) 'The Reliability of Moral Intuitions: A Challenge From Neuroscience' in *Australasian Journal of Philosophy 86:3*, pp.389-405

THOMSON, J.J. (1976) 'Killing, Letting Die, and the Trolley Problem' in J.J. THOMSON (1986) *Rights, Restitution & Risk* (Cambridge [Mass.], Harvard) pp.78-93

TIMMONS, M. (1999) *Morality Without Foundations: A Defense of Ethical Contextualism* (Oxford, OUP)

VÄYRYNEN, P. (2006) 'Moral Generalism: Enjoy in Moderation' in *Ethics 116:4*, pp.707-741

VÄYRYNEN, P. (2009) 'A Theory of Hedged Moral Principles' in R. SHAFER-LANDAU [ed.] *Oxford Studies in Metaethics Volume 4* (Oxford, OUP) pp.91-132

WHITING, D. (2007) 'The Normativity of Meaning Defended' in *Analysis 67:2*, pp.133-40

WILLIAMS, B. (1985) *Ethics and the Limits of Philosophy* (Cambridge [Mass.], Harvard)

WILLIAMSON, T. (2005) *Vagueness* (Oxford, Routledge)

WIKFORSS, Å. (2001) 'Semantic Normativity' in *Philosophical Studies 102,* pp.203-26

WILSON, G.M. (1994) 'Kripke on Wittgenstein on Normativity' in A. MILLER, & C. WRIGHT [eds.] (2002) *Rule-Following and Meaning* (Durham, Acumen) pp.234-59

WITTGENSTEIN, L. (1989) *Remarks on the Foundations of Mathematics* trans. G.E.M. ANSCOMBE [3rd Edition] (Oxford, Basil Blackwell)

WITTGENSTEIN, L. (2001) *Philosophical Investigations* trans. G.E.M. ANSCOMBE (Oxford, Blackwell)

WITTGENSTEIN, L. (2001a) *On Certainty* trans. G.E.M. ANSCOMBE (Oxford, Basil Blackwell

WRIGHT, C. (1984) 'Kripke's Account of the Argument Against Private Language' in *The Journal of Philosophy 81:12* pp.759-78

WRIGHT, C. (1987) *Realism, Meaning and Truth* (Oxford, Blackwell)

WRIGHT, C. (1987a) 'On Making Up One's Mind: Wittgenstein on Intention' in C. WRIGHT (2001) *Rails to Infinity* (London, Harvard) pp.116-42

WRIGHT, C. (1989) 'Critical Notice of Colin McGinn's *Wittgenstein on Meaning*' in A. MILLER & C. WRIGHT [eds.] (2002) *Rule-Following and Meaning* (Durham, Acumen)

WRIGHT, C. (1989a) 'Wittgenstein's Rule-Following Considerations and the Central Project of Theoretical Linguistics' in C. WRIGHT (2001) *Rails to Infinity* (London, Harvard) pp.170-213

WRIGHT, C. (1994) *Wittgenstein on the Foundations of Mathematics* (Aldershot, Gregg Revivals)

WRIGHT, C. (1994) *Truth and Objectivity* (London, Harvard)

WRIGHT, C. (2007) 'Rule-Following Without Reasons' in *Ratio 20*, pp.481-502

ZALABARDO, J. (1987) 'Rules, Communities, and Judgements' in *Critica 21*, pp.33-58

ZALABARDO, J. (1997) 'Kripke's Normativity Argument' in A. MILLER & C. WRIGHT [eds.] (2002) *Rule-Following and Meaning* (Durham, Acumen)