Judging by a Different Standard?

Examining the Role of Rationality in

Assessments of Mental Capacity

by

Natalie F. Banner

A thesis submitted in partial fulfilment for the requirements of the degree of PhD at the University of Central Lancashire

May 2010

Student Declaration

Concurrent registration for two or more academic awards

I declare that while registered as a candidate for the research degree, I have not been a registered candidate or enrolled student for another award of the University or other academic or professional institution.

Material submitted for another award

I declare that no material contained in the thesis has been used in any other submission for an academic award and is solely my own work.

Collaboration

This thesis is not part of a collaborative project.

Signature of Candidate

_____

Type of Award

PhD by Research

School

International School for Communities, Rights and Inclusion

Abstract

Decision-making capacity is an increasingly important medico-legal concept. The recent Mental Capacity Act employs a cognitive, process-based test of capacity, but in many psychiatric conditions pathological beliefs and values impair capacity even when the decision-making process is logically coherent. In such cases, capacity assessments implicitly rely on normative epistemic and evaluative standards. This raises a worry for the capacity test's reliability, objectivity and tolerance of differences in beliefs and values.

There is currently little conceptual research on capacity and the normative standards underpinning its assessment. This thesis makes an original contribution to research by employing a number of philosophical approaches to map out a conceptual terrain within which questions about the substantive standards of capacity assessment can be framed.

Focusing on the nature of epistemic standards and third-person judgements about decision-making, the thesis examines the normative constraints determining what counts as a recognisable reason for a decision. It employs the theoretical apparatus of Davidson's project of Radical Interpretation to explore the epistemology of interpretation, interrogating the conditions under which intentional attribution and the provision of reason explanations for behaviour are possible. It is contended that beliefs are intrinsically rational and intersubjective, and that judgements of irrationality are only possible against a background of shared belief between interpreter and observed agent. This view is defended against the objection that rationality is too stringent a constraint on belief.

A misconception giving rise to this objection is then diagnosed. Drawing an analogy with Wittgenstein's rule-following considerations, it is submitted that the constitutive normativity of belief need not be codified in order to exert a genuine constraint on intentional behaviour. Rather, the norms of belief ought to be construed as emerging from shared practice. This indicates that normative judgements are disciplined through expertise and experience, rather than adherence to abstract principles. Finally, the implications of these insights for conceptualising and assessing capacity are considered.

Table of Contents

Acknowledgements

First and foremost sincere thanks are due to my supervisor, Tim Thornton, for encouraging me to develop a life of the mind, providing careful guidance, thorough criticism and relentless sanguinity in the face of my anxieties. I am grateful for the role he has played not only in being an exceptional academic mentor but an intellectual role model too.

I would like to extend my thanks to Gloria Ayob and the King's College Alumni Philosophy of Psychiatry reading group for helpful discussions at various stages of thesis development. Thanks also to my sister Kathryn, for her boundless optimism and encouragement.

I am indebted to the staff of the Centre for Humanities and Health at King's College London, in particular Derek Bolton, for their flexibility and patience during the final stages of writing up, and also to my colleagues for their invaluable advice and generous support.

Finally, I wish to express my eternal gratitude to my parents, Peter and Lynsay. They have resolutely nurtured me throughout this process, tolerated my anti-social hours, provided endless tea and project management skills, and gone well beyond the call of duty in supporting me in whatever way they could. I dedicate this thesis to them.

INTRODUCTION

The implementation of the Mental Capacity Act (2005) in England and Wales has heralded a surge of interest in the concept of mental capacity, its role in medical decision-making and its assessment in healthcare (Owen et al., 2009b). As mental capacity becomes an increasingly important concept in law, medical ethics and particularly psychiatry, further research will focus on the way it is assessed and the prevalence of incapacity judged according to the criteria of the Act. It is therefore important to establish whether the test for capacity employed in the legislation is able to successfully differentiate between vulnerable individuals who lack the capacity to make particular decisions about their welfare and treatment from those who retain this ability, even if their decisions go against prevailing medical opinions (Jones, 2005b).

This thesis exploits a number of different literatures in law, clinical psychiatric research, bioethics, cognitive psychology, and themes in analytic philosophy of mind, language and psychiatry to consider how the criteria for capacity ought to be understood, and whether there are any universal normative standards underpinning judgements of capacity. The conceptual methodological approach taken situates this research within the field of philosophy of psychiatry, which is developing as a specialist domain of interest with the overarching aim of better understanding the field of psychiatry and its complex, diverse and contentious subject matter (Banner & Thornton, 2007). The methods and arguments employed here are largely derived from analytic philosophy, but the conclusions drawn are potentially compatible with conceptions of intentionality and interpersonal understanding drawn from a more Continental perspective.

The first major claim of this thesis is that procedural criteria do not successfully distinguish capacity from incapacity, particularly in certain psychiatric cases. The presence of beliefs and values perceived to be pathological frequently lead to judgements that capacity is undermined even when the process of reasoning from the

provision of information to a decision outcome is logically intact. Arising primarily from bioethical literature on informed consent and the nature of autonomy, much philosophically-informed research into capacity has focused on the role of substantive values and whether particular kinds of desire or value ought to be held by a person deemed to possess the capacity to make decisions (Holroyd, 2010; Tan & Hope, 2008; Charland, 2001). However, no previous research has investigated the substantive epistemic conditions on capacity, namely whether there are constraints on what kinds of beliefs it would or would not be reasonable for a person to hold if he is to be judged to possess capacity.

To address this question an original conceptual link is forged between judgements about a person's decision-making capacity and evaluations of the connection between one's actions and the reasons one has for those actions. Exploiting the wealth of literature in cognitive psychology on reasoning and decision-making, an informative parallel is drawn between the notions of procedural and epistemic rationality, and procedural and substantive criteria for capacity, which is of conceptual use in gaining traction on instances of irrationality in psychiatry that might undermine capacity.

The theoretical project of Radical Interpretation (Davidson, 1973b), influential in the philosophy of mind and language, is uniquely employed in framing questions about third-person judgements of a person's reasons for a decision. The method is used to examine the conditions of possibility for the attribution of reasons and intentional states more broadly. This is in order to ascertain whether there are universal standards of rationality by which judgements about a person's reasons, and specifically his epistemic commitments, can be made. Emerging from this discussion is the claim that the intentional realm is intersubjectively constituted and structured by norms of rationality. In defending this view, which I term "rational interpretationism", a novel argument in defence of the constitutive link between rationality and intentionality is developed based on the claim that opponents of this connection mistakenly conceive of

the demands of rationality in overly stringent, principled terms. It is argued that judgements about the rationality or irrationality of an agent's decision-making process can only occur against an assumed background of beliefs that are largely true and hang together in a broadly coherent whole.

A therapeutic reading of Wittgenstein's remarks on rule-following forms the basis of two claims about the norms of intentional behaviour. Firstly, the idea of there being radically different forms of rationality is an empty one, as we are bounded in our interpretive capacities by our own mindedness and form of life. This renders unintelligible the possibility of utterly different ways of going on that are nonetheless perceived as intentional. Secondly, beliefs do have an intrinsically normative structure but the constitutive standards of normativity cannot be abstracted away from the context and practices in which they are manifest. Although well-established, the rule-following dialectic has not previously been applied to considerations about the normative standards underpinning third-person judgements of the rationality of beliefs and evaluations of decision-making processes.

Two original conclusions about capacity judgements are drawn from these conceptual considerations. Whilst there may be variations in the details and emphasis of standards underpinning judgements about a person's reasons for a decision, the conceptual possibility of there being radically different forms of reason-giving between patient and clinician is closed off. Secondly, in seeking reliable standards for the assessment of capacity it is a mistake to focus judgements on the determination of context-free procedural or substantive criteria, as it is only by acknowledging the relational content, history and background of the decision-making context that judgements about a person's decision-making process can be reached.

This thesis represents an initial step towards achieving a fuller grasp of what underpins judgements about a person's capacity with respect to a particular decision. The

philosophical exploration of the concept of capacity and judgements about a person's decision-making conducted here fill a niche that has previously been neglected in the empirical and conceptual literature on capacity. The conclusions presented thus form the basis of an enriched understanding of judgements of capacity that potentially could be applicable to the development of clinical guidance for training and policy in capacity assessment. It is my contention that notwithstanding the myriad ethical and legal complexities in determining capacity, it is imperative that the balance between protecting vulnerable individuals and preserving the right to individual autonomy is struck with as much conceptual sophistication as possible.

1. ASSESSING MENTAL CAPACITY

1.1. THE LAW AND CAPACITY

The Mental Capacity Act

The Mental Capacity Act (2005) (MCA hereafter), fully implemented in England and Wales in October 2007, provides a statutory framework for dealing with individuals who may lack the ability to make decisions regarding their treatment, welfare or finances. It covers a wide range of protocols pertaining to, among other things, court powers, advance decisions, independent advocacy and powers of attorney. At its heart the legislation is concerned with preserving the autonomy of the individual as far as possible whilst allowing protection and care to be provided in the best interests of those who are unable to make their own decisions[1] (Jones, 2005a).

In this chapter I introduce the main provisions of the MCA, together with a brief examination of the clinical and legal background in which the legislation was developed. The principles underlying the legal framework will be explored and the commentary will focus on the conceptual basis of several key tenets of the MCA. Rather than provide a general overview (which can be found elsewhere, e.g., Ashton et al., 2006) particular attention will be paid to areas that, I will go on to argue, benefit from a close philosophical analysis to reveal the nature of the judgements involved in an assessment of the decision-making process. This discussion will raise a potential concern about the objectivity of clinical judgement that, if correct, raises an intractable problem for capacity assessment and the implementation of the Act, particularly in psychiatric practice.

[1] Capacity and autonomy are intricately related concepts (Owen et al., 2009). Despite the prevalence of literature, particularly in bioethics, on the notion of autonomy and its role in law, I wish to focus solely on the conception of capacity as presented in the MCA free from the theoretical and conceptual apparatus of theories of autonomy.

The Act is accompanied by a detailed Code of Practice (CoP hereafter), designed to assist clinicians and health professionals in utilising the Act to ensure that capacity is properly assessed; decisions on behalf of those lacking capacity are made in their best interests; and to provide legal protection for those responsible for such decisions. The tenets of the MCA are based on common law principles established in key landmark legal rulings and several resultant consultation papers by the Law Commission (No. 129, 1993; No. 128, 1993; No. 119, 1991). In the early 1990s a number of difficult cases arose that highlighted a gap in legislation for determining when an adult was incapacitated and how he or she should be treated if found to be incapable of making a decision. Against the background of the European Convention on Human Rights (ECHR) and an increasing focus on the rights of the individual, respect for personal autonomy has become the predominant governing principle in health care law (Gunn, 1994, p.8). This is the case so long as the individual retains the capacity to make a decision about his or her treatment. Common law is clear that treating a competent patient involuntarily amounts to the clinician committing a battery (Grubb, 2004, p.161). In this respect the law has been heavily influenced by developments in bioethics regarding the notion of informed consent, which promotes the fundamental right of the individual to free choice and autonomy (Charland, 2008).

Where a person is not capable of making a particular decision, a paternalistic decision on his behalf is legitimate and may even be ethically demanded to protect the individual's health and welfare. The common law principle of necessity dictates that there is a duty of care towards incompetent patients to save life, ensure the provision of beneficial medical treatment or prevent deterioration, or to act in the patient's best interests as determined by prevailing medical opinion (Raymont, 2002). The assessment of capacity is thus of crucial importance in medical decision-making and it is essential that the way it is tested sets a standard that strikes a balance between protecting vulnerable adults from harm and respecting personal autonomy where it is intact.

The MCA and its associated Code of Practice explicitly state that capacity is a functional notion, comprising two distinct but related aspects. Most importantly, the concept of capacity as functional means that its assessment should be based on evaluating the processes a patient uses to arrive at a decision rather than the content of the decision itself: *"What matters is* [the] *ability to carry out the processes involved in making the decision – and not the outcome"* (CoP, section 4.2). The concept of capacity employed here can be distinguished from two possible alternatives: the status and outcome approaches.

A status approach to capacity would entail that if a person were deemed to lack capacity, this assessment would apply to all decisions that person could make. This kind of approach is typically taken with young children (Stauch et al., 2006, p.115) and is also implicitly employed in mental health legislation when a person is detained under a section for treatment of a mental disorder. In these cases decisions are paternalistically taken on the patient's behalf on the basis of a clinical judgement of best interests. The functional conception of capacity is thus distinct from the formal legal category of incompetence which implies a person is incapable of making any decisions at all for legal purposes (Nys et al., 2004).

The outcome approach focuses on the result of the decision-making process. Any outcome deemed to be unreasonable, unwise, against conventionally held values or against medical opinion could be considered as evidence of incapacity[2]. This would undermine the very basis of requiring valid consent and vitiate the ideal of patient autonomy that health care law strives to protect (Stauch et al., ibid.). However, whilst indicating that assessing outcome is an inappropriate way to approach determining capacity, the Law Commission highlighted that it was in fact common in clinical practice

[2] The distinction between process and outcome based judgements of decision-criterion has been explored more fully in bioethical literature on autonomy, where the dichotomy is set up as between procedural and substantive accounts of autonomy.

"…if the outcome is to reject a course which the doctor has advised then capacity is found to be absent" (Law Commission, Report No. 231, 1995, para. 3.4). This raises a conceptual and ethical issue regarding the way capacity assessments are conducted in practice and what assumptions clinicians might implicitly make about what it means to possess capacity when judging individual cases.

The second aspect of the functional approach concerns the time and situation specificity of the assessment process. The CoP states that use of the term "capacity" refers to *"a person's…capacity to make a particular decision at the time it needs to be made"* (CoP, p.19). This stipulation aims to ensure that a status approach is not taken to an individual whose capacity is in question. This is particularly important in contexts where an individual's capacity fluctuates depending on, for example, the time of day, cycle of mood or the effects of medication. Capacity is therefore not concerned with general cognitive or reasoning abilities abstracted away from the specific context in which it is being assessed.

Capacity, Psychiatry and the Mental Health Act

Mental health law and the MCA have in common the provision of statutory powers by which an individual's right to make decisions regarding his welfare and treatment can be revoked. However, there is a clear distinction between the two sets of legislation, reflecting their very different purposes. The purpose of capacity legislation is to protect the interests of people with mental impairments. Mental health law, by contrast, has a dual concern with public safety and operates on the basis of managing risks to the self or others (Fennell, 2007).

In 2006 Richards and Mughal predicted that there would be significant interplay between mental health legislation and the MCA, particularly in cases concerning severe learning disabilities or dementia. The Mental Health Act (2007) supersedes the Mental Capacity Act for patients who qualify for a section under the former, so that if a

person is deemed to be seriously mentally ill and posing a danger to himself or others he may be detained irrespective of considerations about his capacity[3]. The MCA therefore applies to the large proportion of patients who need not be detained under a Mental Health Act section but who nonetheless may require others to make treatment decisions on their behalf[4].

Patients who do not fall under the jurisdiction of the Mental Health Act may be at a greater risk of being deemed to be lacking capacity than other clinical populations. Evidence for this phenomenon is rife (Raymont et al., 2004; Wong et al., 2000; Cairns et al., 2005ab) and it has been pointed out that *"until recently it was commonly presumed that serious mental illness, by definition, rendered a patient incapable of consenting to treatment"* (Cairns et al., 2005a, discussing a study by Grisso et al., 1997). The functional concept of capacity used in the MCA is intended to avoid discriminating against patients merely on the basis of a diagnosis of mental disorder or learning disability (CoP, p.57).

Van Staden and Kruger (2003) discuss the various concomitants of certain mental illnesses that may compromise a patient's capacity: indifference, ambivalence or indecisiveness during major depressive or manic episodes, problems with memory in dementia and disorganisation of thoughts in schizophrenia or other psychotic illnesses may clearly render an individual unable to make a decision regarding his treatment. However, Carpenter et al. (2000) demonstrate that impairments in decisional capacity were only modestly related to symptom prevalence in schizophrenia, with a much stronger correlation between poor performance on decisional tasks and cognitive

[3] Efforts to bring mental health and mental capacity legislation together (Dawson & Szmukler, 2006) aim at closing the gap between the conceptual and normative questions, arguing that if someone possesses capacity with respect to a decision, there are no circumstances under which this capacity should not be respected, but current mental health legislation in England and Wales contradicts this view.

[4] It will also apply to patients sectioned under the Mental Health Act who are in need of treatment for a physical disorder, which is not covered by the involuntary powers of the Mental Health Act. Only treatment for the mental disorder for which a patient is detained can be given compulsorily, and thus a test of capacity under the MCA will need to be conducted even for patients who are detained.

impairments that are not specific to psychiatric conditions. Presence of a severe mental illness does not therefore entail one's capacity is impaired.

This assertion is reflected in the judgement of Justice Thorpe in a landmark case (*Re C* [1994]). An institutionalised patient suffering from schizophrenia, who believed that he was a world-famous doctor, refused to consent to his gangrenous leg being amputated. He did not believe the gangrene would kill him and he held a strong desire to die with his body intact. His doctors believed it was extremely likely that he would die if the amputation did not occur, and sought permission from the courts to carry out the operation in spite of the patient's refusal. The judge ruled that although the patient was suffering from delusions resulting from a severe mental illness, this did not compromise his ability to make a specific treatment decision about the amputation of his foot[5] (Stauch et al., 2006, p.117).

Owen et al. (2009b) describe two categories of psychiatric patients to whom the MCA might apply. Patients with organic psychiatric diagnoses such as dementia or severe learning disabilities, which involve clear cognitive and memory impairments, form one group. For such patients capacity assessments may be reasonably straightforward and surrogate decision-making in their best interests is legitimised on the basis of their inability to engage with the process of decision-making at all. The other group comprises patients whose incapacity might not be thought of in cognitive terms: severely depressed patients who are ambivalent about treatment; paranoid schizophrenic individuals who do not trust or appreciate the significance of information given to them about their treatment; or patients with anorexia whose thought processes are dominated by fears and desires about their weight, for example. In such cases, surrogate decision-making is predominantly conducted on the basis of risk of harm, but if the risk does not qualify as sufficiently severe to warrant detention under section, an

[5] Despite the refusal to permit amputation the patient survived and his leg responded to alternative treatment, reminding us that medical opinion about what constitutes the best form of treatment is indeed fallible (Gunn et al., 1999).

assessment of capacity is required, and if capacity is intact the patient's decision must be respected. Capacity assessments for these patients are considerably more complex and represent the most challenging cases for clinical judgement, and are the most interesting for a philosophical exploration of the concept of capacity.

The Presumption of Capacity

The MCA contains five guiding principles, designed to emphasise the underlying ethos of the Act and make clear that the legislation is concerned with balancing autonomy and dignity with protection for those who lack capacity (Ashton et al., 2006). The first of these principles is the presumption that all adults have capacity unless it is established that they do not (MCA s 1(2)). The onus of proof is on the assessing clinician or health professional to show 'on the balance of probabilities', (which is the usual legal standard in civil proceedings) that a patient lacks capacity to make a specific treatment decision (British Medical Association & The Law Society, 2004). A patient's incapacity with respect to a particular decision cannot be presumed or inferred on the basis of his appearance, age, behaviour or medical or psychiatric diagnosis (MCA, s 2 (3a,b)). Judgements influencing the development of the MCA have been explicit in their efforts to endorse this approach in standard clinical and legal practice.

Unwise Decisions

Another guiding principle of the Act stipulates that: *"*[a] *person is not to be treated as unable to make a decision merely because he makes an unwise decision"* (MCA, s 1(4)). The competently made unwise decision should stand even if family members, friends or clinicians are unhappy with that decision. Such a right has been enshrined in English common law since 1850 (Ashton et al., 2006). Its purpose is to avoid as far as possible the threat of medical paternalism, which would lead to a patient's right to autonomy being overruled if a clinician does not agree that the patient has made the right or best decision.

Under normal circumstances where the ability of a person to make a decision is not in question, he has an inalienable right to decide whatever he wants, even if this is likely to result in his own death or disability. No reasons, justifications or rationalisations need to be provided to substantiate or explain his decision:

> "…the patient's right of choice exists whether the reasons for making that choice are rational, irrational, unknown or even non-existent" (Lord Donaldson in *Re T* [1992] at 653).
>
> "A mentally competent patient has an absolute right to refuse to consent to medical treatment for any reason, rational or irrational, or for no reason at all, even where that decision may lead to his or her own death" (Re MB [1997] at 426)

The case *Re MB* established at law that decisions based on irrational beliefs do not indicate a lack of capacity unless the belief is caused by a mental impairment. The implication here was that even a decision that is *"so outrageous in its defiance of logic or of accepted moral standards… [that]…no sensible person…could have arrived at it"* (*Re MB* at 437) does not indicate the person should be found to lack capacity. It is only if a mental disturbance or impairment to mental functioning looks to be compromising this decision-making ability that capacity comes into question.

Even where doctors and the majority of people might consider a particular decision to be unwise, the patient retains the right to make that decision if the process of decision-making is intact. If the stipulation regarding the wisdom of decisions were revoked, patients would be restricted to making decisions in accord with medical opinion. The converse situation may also arise if this outcome approach to capacity is taken: agreement with medical opinion would lead clinicians to believe the patient does possess capacity, even if the patient is merely passively complying and not actively consenting to the proposed treatment (Gunn, 1994). A seemingly wise decision may also not necessarily be underpinned by a reasonable decision-making process. A patient could agree to risky surgery for the wrong reasons, such as a deluded belief in his own invincibility. Thus even a 'wise' decision outcome would not be indicative of the patient possessing capacity. Placing the unwise decision principle on the face of the MCA amounts to an explicit rejection of the outcome approach to capacity, although as

the Law Commission noted we should recognise that such an approach *"is almost certainly in daily use"*[6] (Report No. 231, para 3.19).

In spite of the assurances laid down in the MCA and its CoP, clinicians may well consider treatment refusal as grounds for suspecting the patient may lack capacity: *"doctors faced with a refusal of consent have to give very careful and detailed consideration to the patient's capacity to decide…"* (*Re T* at 662). This suggests that a decision clinicians consider to be unwise, such as refusing treatment, may give rise to suspicion that the patient lacks capacity even in the absence of prior evidence of mental impairment. The patient's capacity would then be formally assessed under the criteria of the capacity test and his decision-making process scrutinised: a stringent demand considering that ordinarily no reasons or justifications are required. Such reasoning hints at an implicit outcome approach to capacity, leading to the burden of proof being placed upon the patient to prove he does have capacity or to provide reasons for his refusal, rather than there being a presumption of capacity (Kennedy, 1997, p.320-1). This illustrates the difficulties posed in practice of attempting to divorce decision outcome from the process of decision-making.

The principles of the MCA point towards a conception of mental capacity that has important legal, ethical and political dimensions. Determining that a patient lacks capacity enables a clinician to make decisions on the patient's behalf in his best interests, thus investing the clinician with considerable power and responsibility over that individual's life, temporarily at least. To avoid an overly paternalistic approach much emphasis has been placed in the MCA on evaluating the process of decision-making and not its content or outcome.

[6] Nonetheless, the Code of Practice states that there may be concern if somebody repeatedly makes unwise decisions that put them at risk of harm or exploitation or makes a particular unwise decision that is obviously irrational or out of character (CoP, p.25).

1.2. ASSESSING CAPACITY

Mental Impairment and Incapacity

The test for capacity is intended to aid clinicians in determining whether a patient's decision is autonomous, and thus ought to be respected, or indicative of a lack of autonomy and can thus be overruled because it is made on the basis of a mental impairment:

> "...a person lacks capacity in relation to a matter if at the material time he is unable to make a decision for himself in relation to the matter because of an impairment of, or a disturbance in the functioning of, the mind or brain." (MCA, s 2(1))

Sections 2 and 3 of the Act set out a two-stage process for testing capacity. The first stage checks the inclusion criterion that the person must be suffering from an impairment or disturbance to his mental functioning, whether this is temporary or permanent[7] (CoP, section 4.11). Only if this criterion is fulfilled does assessment proceed to the second stage, which stipulates that for a person to come under the powers of the Act, the impairment of mental functioning must be causing an inability to make the relevant decision. Examples of such impairment include delirium, coma, severe brain damage, dementia and severe learning difficulties. A further category of inclusion is the deliberately worded *"conditions associated with some forms of mental illness"* (CoP, loc. cit.) although the Code is clear to emphasise that a psychiatric diagnosis alone does not constitute incapacity. Informed largely by empirical research, investigations into the relation between impairment and incapacity have focused mainly on the physiological and neuropsychological effects some mental disturbances have on cognitive functioning. Doubts about capacity commonly arise with respect to individuals suffering from psychotic disorders or severe cognitive impairment. In these cases it is often clear that there is a causal relation between mental impairment and disruption of at least one of the component abilities of capacity. Delirium, which is a common source of incapacity in hospitalised patients, conspicuously affects a person's decision-making

[7] In earlier versions of the legislation, a mental disorder as defined under the Mental Health Act (1983) was a necessary precondition for capacity assessment, but this was dropped in part to avoid discrimination and in part to broaden the scope of legislative powers to include non-psychiatric populations (Gunn, 1999, p.282).

ability by altering cognition and disrupting thought processes to render the actions and utterances of the person incomprehensible (Raymont, 2002). The severe cognitive impairment typical of advanced dementia is perhaps the classic example of a clear indication that a patient may lack capacity (Kim et al., 2002; Nygaard et al., 2000). Capacity may also be impaired by external factors that have only a temporary effect. Shock, confusion, sedation, fatigue, panic, pain and medication may all potentially undermine capacity by diminishing the person's ability to take in information or engage in a coherent process of decision-making (Grubb & Laing, 2004, para 3.91).

The early Law Commission reports (Report No. 231, 1995; No. 129, 1993; No. 128, 1993; No. 119, 1991) stated that the presence of mental impairment must be established as the *sine qua non* of capacity assessment and subsequently it must be determined that this impairment causes an inability to make the decision at hand. Mental impairment is thus a necessary but not sufficient condition of incapacity (Nys et al., 2004). The inclusion of this criterion was influenced by the judgement of LJ Butler-Sloss in *Re MB* [1997]. In this case a heavily pregnant woman consented to delivering her baby by Caesarean section for medical reasons and because the baby was at risk. However she suffered from an extreme needle-phobia and panicked at the last moment, withdrawing her consent to the procedure. The judicial judgement was made that she was suffering from a temporary mental impairment in virtue of her phobia at the time of decision-making, and was thereby rendered incompetent to refuse. This case established at law that a judgement of incapacity can only be made if there is a causal relation between a mental impairment and the suspected incapacity.

The Test for Capacity

The second stage of assessment sets out a test of capacity, introduced to assist in the judgement as to whether a person is unable to make a decision:

> "…a person is unable to make a decision for himself if he is unable-
> (a) to understand the information relevant to the decision;
> (b) to retain that information;

(c) to use or weigh that information as part of the process of making the decision; or
(d) to communicate his decision" (MCA s 3(1))

If an individual cannot demonstrate any one of these four abilities in spite of sincere efforts being made to assist the decision-making process then he can be deemed to lack capacity[8]. Crucially, the assessment of these abilities is not supposed to inform the judgement about whether or not the patient is suffering from a mental disturbance or impairment, or the test would risk being circular. In order for the test to be non-question begging, a person's failure on any of the four criteria ought not to be used as evidence for mental impairment[9] (Kennedy, 1997), p.322). I will consider these conditions for incapacity in turn.

Understanding and Retaining Information

A necessary requirement for capacity is that one understands and retains the information relevant to the decision. This includes having an awareness of the purpose of the treatment, an idea of what it will involve and the consequences of deciding to receive or refuse the treatment, or of not making a decision at all (Ashton et al., 2006, para 2.57). Every appropriate effort must be made to assist in communicating this information to the patient. These criteria aim to minimise the gap between potential and actual understanding so that the patient is able to participate in the decision-making process to the best of his abilities (Gunn, 1994, p.18).

Whilst the requirement of imparting relevant information to the patient appears straightforward, Gunn et al. (1999) reported difficulty in ascertaining what the important relevant information should be in a research setting in which patients were assessed for capacity to consent to the simple procedure of a blood test. The authors indicate that somewhat to their surprise it was not easy to reach consensus about what patients needed to demonstrate to be assessed as understanding the relevant information.

[8] Efforts to maximise the patient's ability to comprehend the information and decide for himself include using simple language, visual aids, conducting the assessment when the patient is most cognitively alert, and so on.

[9] As of yet here is insufficient empirical evidence from the early implementation of the MCA to determine whether or not this potential circularity arises in practice.

Thus although these criteria looks as though they are tapping into cognitive capacities, namely memory and attention, there are ambiguities and subtleties in their application that will impact on how they are judged in practice[10]. For instance, there has been conflict in judgements as to whether it is necessary that a patient believes the information being presented to him about his condition in order to be deemed to possess capacity (*Re C*; *Re MB*). This may be because appreciation can be considered a necessary part of understanding: unless a patient appreciates and believes that the information about treatment being presented applies to him and his situation, he could not be said to understand that information.

The presence of a mental illness may prevent a patient from being able to acknowledge the need for medical treatment. It may also be the case that a patient is capable of hypothetically understanding the proposed treatment and reasoning about it whilst denying that he needs any intervention. This could occur when patients simply do not accept their doctor's diagnosis or opinion about their illness (Stauch et al., 2006, p.121). It would be undesirable for legislation to require that in order to have capacity, a patient must believe everything a clinician tells him about his illness and treatment. However, *"patients who refuse treatment in a psychiatric setting are particularly likely to be judged as lacking capacity"* (Hotopf, 2005, p.582), indicating that in practice at least clinicians may consider treatment refusal to indicate a failure to understand and appreciate the facts of their condition and their need for treatment[11].

The issue of appreciation extends beyond the question of whether or not understanding incorporates an element of acknowledgement that the information presented applies to and is relevant to oneself and one's situation. Decision-making is not a purely abstract or intellectual exercise but contains an important affective element: the decisions made impact upon the person's life, health and relationships. Emotional value may be

[10] See Manson & O'Neill (2007) for an account of the implications a fine-grained understanding of these criteria have for the notion of informed consent in bioethics.
[11] Appreciation is taken here to be similar to the concept of insight in psychiatry.

attached to various risks and benefits, the capacity for appreciation of which differs from a mere factual or abstract understanding of the situation (Gunn et al., 1999, p.19). For example, a person suffering from a severe episode of depression may be able to understand the facts relating to proposed courses of treatment, but he attaches no emotional significance and is ambivalent towards any proposal (Rudnick, 2002). It would be questionable whether such a patient unable to appreciate the significance of the decision-making situation and its potential impact on his life, health and welfare had the capacity to make this treatment decision.

Using or Weighing Information

The criterion of using or weighing information in the process of coming to a decision is perhaps the most conceptually difficult to understand and empirically difficult to test or measure. In employing this criterion the functional approach aims to direct capacity assessment towards an evaluation of the process of decision-making rather than its outcome. A person is able to use or weigh information in coming to a decision insofar as he can consider the risks, benefits, consequences of receiving or not receiving treatment and take into account his own beliefs and system of values in determining what to do (Ashton et al., 2006, para 2.61). Little further specification has been given for what constitutes a threshold for its fulfilment.

It is natural to assume that in assessing whether someone is using or weighing information, one needs to be aware of what information is entering into that process. Whilst the relevant treatment information imparted by the clinician will be an important part of this, other factors such as a value system and personal beliefs will also be influential in determining the decision outcome (Stauch et al., 2006, p.126). A hypothetical clinical vignette drawn from the CoP illustrates some of the relevant considerations. A patient with learning disabilities needs to have regular blood tests to monitor his medication for a minor heart condition, but has a phobia of needles. His doctor explains the purpose of the blood test and demonstrates what the procedure will

involve. In the process of decision-making, the patient will need to handle a number of different factors that may influence that process and its outcome. Highly relevant will be the patient's beliefs about a variety of things including that: he has a heart problem; it needs treatment; the blood test is an important part of his treatment; his doctor is trying to help manage his condition; the test won't harm him; there are possible negative consequences for his health if he refuses the test; he is scared of needles. He may also hold a particular set of values and have certain desires: the desire to stay healthy; to avoid being hurt; to prove he is capable of independent living; to be a good patient; to avoid situations that make him anxious. Whilst not all of these factors will be evident to the assessing clinician, the patient must be able to display some kind of process leading from the acknowledgement and understanding of the given treatment information to the reaching of a decision. This superficial outline indicates that using or weighing information entails taking available evidence and testimony together with an appreciation of one's beliefs and values to produce some kind of decisional outcome. The terminology I am using to describe the process is deliberately vague at this point: it is by no means clear what counts as using or weighing information other than there is an as yet unspecified connection between all the 'input' factors and the decision outcome. One prominent legal commentator queries why this standard is in place: *"why does the law insist on evidence that information, once comprehended, retained and believed, must be weighed in the balance?"* (Kennedy, 1997, p.321). The functional approach takes it that this very process, marking the transition from taking on board information to reaching a decision, is of central importance in the determination of capacity. Examining the nature of and evidence for this process will be a major focus of this thesis.

Communicating a Decision

A patient must be able to communicate his decision, whether verbally, in sign-language, through the use of visual symbols or by some other means, if he is to be considered to have capacity. This criterion is particularly relevant for patients who have

suffered strokes or a paralysis-inducing neurological deficit. A person might be able to understand, retain, assess and reason with the information about his condition and treatment options but if he is unable to express his decision, he lacks capacity. The significance of this criterion will be brought to bear in later chapters when I consider the importance of the interface of communication for third-person judgements about a person's reasoning process.

This discussion of the MCA has outlined the development of the ethical and legal principles central to the legislation. I now turn to examine how the challenges of assessing capacity have been met in clinical research and what conceptions of capacity and its measurement underlie the predominant approach. In doing so I question some fundamental presuppositions about the possibility of generating impartial, objective judgements of capacity and consider the implications this has for the assessment of patients' decision-making processes.

1.3. EMPIRICAL APPROACHES TO CAPACITY

The Concept of Mental Capacity

Several questions need to be distinguished if we are to understand the concept of capacity and how it functions in clinical practice[12]. Firstly, there is a conceptual question: "what is capacity?" What does it mean to say that someone possesses or lacks the capacity to make treatment choices? This question is primarily non-empirical, as it asks after the presuppositions underlying any theory of capacity or its assessment. Capacity is a legal, clinical, ethical and social construct (Hotopf, 2005) and as such cannot be understood by reference to one field of expertise alone. Secondly, there is a normative question: "what rights does capacity confer on a person?" This concerns the

[12] The term competence is used more frequently in legal literature, often to confer a status upon a person, whereas capacity is more usual in clinical usage. Bielby (2005) warns against conflating these two concepts, and in the USA this separation between clinical and legal usage of the terms is maintained. However throughout this thesis I will use the terms 'capacity' and 'competence' interchangeably, in line with common usage in the UK.

legal, moral and political normative obligations that influence how we, as a society, treat individuals with and without capacity. The third question is an epistemic one that has been addressed in the empirical clinical literature: "how can we determine whether or not an individual has capacity?" There are many clinical tests and diagnostic tools designed to aid the assessment of mental capacity, all of which strive for increased reliability across raters and different contexts.

Charland (2001) distinguishes two aspects of capacity, referred to as the *"dual nature of competence"* to illustrate the different concerns in play when an assessment of capacity needs to be made. It is first important to understand what capacity is and how it can be measured. Charland refers to this as the "descriptive" dimension of competence, understanding of which aims to provide objective, valid definitions or descriptions of what it is to possess or to lack capacity, and to develop diagnostic tools to help clinicians make this determination[13]. Additionally however, an ethical question also comes into play concerning whether or not a person should retain his right to personal autonomy. This is determined by the legal and ethical framework within which one is working. Charland considers the ethical question to form the second and normative dimension of capacity. Normativity here concerns what it is appropriate, correct or right to do or think, construed in terms of moral obligations or imperatives[14]. I will say considerably more about the notion of normativity later, arguing that the sphere of normativity is not limited to the realm of ethics, but here the term serves as a placeholder to describe the specifically ethically oriented dimension of capacity.

The overriding principle of respect for autonomy in health care law ensures that only the first dimension of capacity is usually relevant. If a person has capacity, it follows that he ought to have the right to make his own choices. For the vast majority of cases

[13] At this point I am using the term 'objective' in a lay scientific sense to signify something that exists independently of an observer.
[14] This conception of the dual nature of capacity was originally noted by Freedman (1981, p.55), an influential writer on mental capacity, who queried whether competence is an empirical or moral form.

the “descriptive” aspect of capacity is the sole determinant of whether a person should be permitted to make his own treatment choices. However, if a patient is deemed to pose a risk to himself or to others on the grounds of mental disorder he may be detained under the Mental Health Act and compulsorily treated irrespective of his capacity to make treatment decisions (MHA 2007, CoP, sections 4.9-4.10). Here, respect for autonomy is overridden by the ethically-motivated determination that the patient should not be permitted to make treatment choices.

Taking both ethical and descriptive dimensions into account suggests that determining whether a person should be entitled to make his own decision requires a two-stage evaluation: firstly, a descriptive account of whether or not the patient does in fact possess capacity, and secondly a normative ethical judgement about whether the patient’s capacity (if intact) should be respected. Carving up the dimensions of capacity assessment in this way provides an insight into the way in which research into mental capacity and its assessment has proceeded. If we are seeking to test mental capacity the factual, descriptive component appears to fulfil this role sufficiently, and independently of the ethically normative dimension. Separating out the ethically normative issues from the descriptive project ensures that research into what capacity is and how it can be measured, evaluated and assessed can continue independently of the ethical complexities of balancing autonomy with protection. However, I suggest that the distinction between descriptive and ethically normative components of capacity is misleading: it assumes that the only normative considerations relevant to capacity are ethical norms that impact after the fact of making a descriptive judgement about a person’s capacity status. In the following discussion I avoid the considerable and complex ethical debate surrounding problems in the ethics of mental capacity (see for instance Buchanan & Brock, 1986), instead focusing on the descriptive dimension of capacity. The question I wish to consider is whether the definition and determination of capacity are indeed free of normative considerations. This will firstly require an

examination of the underpinnings of the predominant approach to capacity and its assessment in contemporary research.

The Operational Approach

An influential paper by Appelbaum and Roth (1982) developed what has become known as the "cognitive conception" of capacity, claiming that capacity is comprised of a set of cognitive abilities or faculties[15]. This approach assumes that capacity can be operationalised through measuring what are presumed to be its constituent psychological processes. These processes are taken to be, in principle, amenable to observation and measurement by a third party observer, via the construction of scales or indices of observable behaviour or functions concerning memory, inferential reasoning and information processing (ibid). This is common in empirical psychology where it is often difficult to gauge abstract psychological constructs or processes through direct measurements, and *"the fact that these abilities are characterised as cognitive is usually deemed essential to the objectivity of the proposed operationalised standards for assessing them"*[16] (Charland, 2001, p.136). On the basis of an assumption that capacity is constituted by certain kinds of psychological processes the epistemic project of determining a person's capacity can be pursued: using operationalised criteria for the fulfilment of capacity, the right kinds of tools can be devised to measure the constitutive processes. The cognitive view supposes that the relevant processes and mechanisms are the ability to store information, to identify its relevance and to process it in order to make a decision, broadly construed as the capacities of memory, understanding and reasoning. Established tests of neuropsychological functioning such as the mini mental state examination have demonstrated strong correlations with judgements of incapacity in studies involving organic psychiatric disorders, but only weak and inconsistent correlations where it is

[15] Note that the term 'cognitive' is here employed as it is used cognitive science and psychology: to refer to internal mental processes. It is not used in the philosophical sense of cognitive states as being truth evaluable.

[16] For instance, fear is not necessarily directly observable to a third person observer but the concept may be operationalised by measuring galvanic skin response, on the assumption that increased perspiration rate is taken to be a reliable measure of a person's degree of fear.

not so clear that cognitive functioning is impaired (Okai et al., 2007). This suggests that incapacity cannot be modelled on standard tools for assessing neuropsychological deficits alone (Owen et al., 2009b).

The shift towards operationalising criteria for testing and diagnosis of mental illness has been mirrored by progressive editions of the American Diagnostic and Statistical Manual of Mental Disorders (DSM) from DSM-III (American Psychiatric Association, 1980) onwards. Progress and improvement in the utility of assessment tools is achieved as they become increasingly reliable and less susceptible to the idiosyncrasies of individual clinical judgements, yielding similar results when used by different clinicians. Significant efforts have been made to minimise the role of the clinician in the diagnosis of mental disorders by a clear preference instead for scales of symptoms rated according to observed behaviour, self-report questionnaires and, if possible, biological markers. Indeed, papers written for the American Psychiatric Association (APA) that aim to set out a research agenda for DSM-V and beyond explicitly argue for reducing the role of clinical judgement to *"remove items that cannot be determined reliably through patient self-reporting or through objectively observable signs of behaviours"* (Kupfer et al., 2002, p.21). Clearly, reliance on the clinician's judgement is considered to be inferior and undesirable compared to objective physiological markers and supposedly impartial psychological assessments based on rating scales. The intention here is to increase the reliability and validity of diagnosis. Reliability can be construed as the extent to which different clinicians agree on an assessment, whereas validity refers to the extent to which the tools measure what they purport to measure. This shift towards operationalising diagnostic criteria rests on the assumption that what such tests are measuring is in fact an objectively verifiable psychological construct or process. To this end the clinician is merely a trained impartial observer making descriptive judgements about what is or is not the case about a patient's mental states and cognitive functioning.

There are obvious advantages to such an approach to capacity assessment. Reliable, valid criteria and testing procedures create a transparent process, reducing the risk of abuse by clinicians and standardising assessments irrespective of patient background, or cultural or religious diversity. However, operational diagnostic criteria overlook the essential role of the clinician in determining what counts as fulfilling the capacity criteria. Furthermore, in what follows I argue that what is being assessed is an inherently normative process, and this normativity potentially generates difficulties for the idea that decision-making processes can be objectively assessed. I now turn to a brief survey of the diagnostic aids and tools that have been developed to assist in capacity assessments to support the first claim.

Most research in the area of mental capacity is based on unified, underlying assumptions about the cognitive nature of capacity and takes a broadly similar approach to the generation and analysis of tools used for its assessment. Appelbaum and Roth (1982) established four operationalisable criteria which they deemed should be met if a person is to be considered to possess capacity:

- Understanding of information;
- Appreciation of relevance of treatment options to one's own situation;
- Reasoning using the information provided;
- Communicating a decision.

The first three of these criteria are explicitly concerned with testing patients' memory, information processing and inferential reasoning capabilities, the assumption being that a good test of capacity would enable clinicians to provide a binary 'yes or no' judgement about whether patients could demonstrate each of these abilities in relation to a particular decision. Based on these criteria the MacArthur treatment competence study (Appelbaum & Grisso, 1995; Grisso et al., 1995; Grisso & Appelbaum, 1995b; Grisso et al., 1997) was designed to develop standardised tools for assessing capacity, and from these initial studies the MacArthur Competence Assessment Tool-Treatment (MacCAT-T) was devised. It consists of a series of adaptable questions that can be tailored to the specific situation and administered by interview with a patient. The

MacCAT-T assists in determining capacity by rating each of the constitutive abilities on a 3-point scale (inadequate, partial, adequate). The virtue of this approach is that it is straightforwardly applicable in practice (Grisso et al., 1997) and generates a high level of concordance among clinicians when using the same tools. The MacCAT-T has been used frequently to determine the prevalence of incapacity in general medical (e.g., Raymont et al., 2004) and psychiatric settings (Cairns et al., 2005b; Vollman et al., 2003) and it provides a high degree of inter-rater reliability (Okai et al., 2007). The developers do, however, stress that the diagnostic tool is intended to be used as an aid to assessment alongside clinical judgement and there are no absolute cut-off or threshold scores for distinguishing capacity from incapacity (Grisso & Appelbaum, 1995a).

The utility of such diagnostic tools has been investigated by comparing the assessments clinicians make when using these tools to the assessments of an expert psychiatric clinician. Here the judgement of the expert is used as the standard against which the tools are assessed. For instance, in a study by Janofksy et al. (1992) the authors explicitly judged the accuracy, reliability and validity of an assessment tool similar to the MacCAT-T on the basis of the goodness-of-fit it produced relative to expert psychiatric assessment. Similar methodologies have been employed to evaluate other diagnostic tools: their ability to discriminate patients found by an expert psychiatrist to lack capacity from competent individuals has been reported as evidence for their utility in clinical settings (Bean et al., 1994; Tomoda et al., 1997). Yet in none of these studies is any further detail provided about how or by what process these expert judgements are made, nor by what criteria a clinician was deemed to have such expertise.

When comparing assessments made by different methods, either with diagnostic tools or based in expert clinical judgement, there is often a significant discrepancy between clinical and tool-based evaluations (Kitamura et al., 1998; Mukherjee & Shah, 2001;

Vellinga et al., 2004). Thus when using diagnostic tools to assist in assessment, clinicians often appear to reach different conclusions about the capacity of patients than when they make a judgement that is not guided or constrained by the concept of capacity upon which the tools are based. This finding strikes a *prima facie* blow to the assumed validity of assessment tools as it indicates their criteria do not match up to general clinical intuitions about what constitutes capacity. Indeed, some commentators have acknowledged that criterion validity, which concerns how closely different measures of the same hypothetical construct match up, is impossible to test for capacity assessments (Cairns et al., 2005a).

Unpicking the Assumptions

Two specific issues are raised by this brief examination of the development and underlying assumptions of capacity assessment tools. The first concerns the role of the expert clinician in determining capacity, which the operationalising of criteria seeks to minimise. The tools are based on the fundamental assumption that the determination of capacity is a matter of an impartial, passive observer ascertaining whether a patient has certain abilities in relation to decision-making. However, even the determination of capacity made using the diagnostic tools is calibrated during their development by the judgement of an expert clinician. Although there may be considerable agreement between clinicians, difficult cases appear to yield different assessments from experts (ibid). This is not an issue unique to capacity assessment, since all tools and instruments of measurement rely to some degree on a human perspective both to gather and interpret relevant data: there are many situations in which experts may fail to reach consensus, even where the judgement is about an objective matter of fact[17]. The essential role a clinician plays in making the assessment does not therefore necessarily undermine the objectivity of what the judgement is about, if that judgement is merely a matter of making a potentially difficult binary decision about the fulfilment of a set of operational criteria by the patient.

[17] Concordance rates between radiographers in reading x-rays are a case in point (Sackett et al., 1991, p.30).

However, the second issue raised by this survey of diagnostic tools impacts on this assumption and queries the foundations of the cognitive approach to capacity. Acknowledging the difficulty in establishing a "gold standard" for capacity assessment instruments Kim (2001; 2006) highlights that the cognitive conception relies on the assumption that what is being investigated and assessed is an objectively measurable phenomenon. Yet assessment tools are ultimately calibrated by the judgement of an expert clinician, not a perspective-neutral measure of cognitive functioning. Low concordance rates recorded when clinicians are not guided by diagnostic tools (e.g., Kitamura et al., 1998) suggest that clinicians may be intuitively employing different concepts of capacity or its constituent abilities, further undermining the idea that capacity assessment is a matter of the impartial, theory-neutral measurement of observable behaviours and functions. Some authors have criticised the cognitive bias in research, and the MacCAT-T in particular, for underplaying the complexity of decision-making and overstating the operationalised elements of capacity (Breden & Vollmann, 2004; Higgs, 2004; Silberfeld, 1994). Such criticisms do not merely attack the level of refinement of the tools but seek to undermine the assumption that capacity can be determined by ascertaining scores or indices of the patient's mental functioning.

Capacity Judgements on Cognitive Criteria

One strategy for dealing with this potential problem for the objectivity of capacity assessments is to modify the criteria for capacity to measure only those elements of capacity that can be descriptively ascertained. This approach is outlined by Culver and Gert (2004). They moot the possibility of characterising competence (taken here to be synonymous with capacity) solely in terms of understanding and appreciating relevant treatment information. This strategy is based on the standard in bioethics that understanding information and appreciating its relevance to oneself are prerequisites for providing valid informed consent to treatment. Both criteria can in theory be determined independently of the outcome of the decision, satisfying one of the key

desiderata of capacity assessment. Furthermore, establishing if a patient fulfils the criteria does not require a difficult judgement on the part of the clinician: briefly quizzing the patient about the content and relevance of the given information is thought to be sufficient to determine whether or not he does, as a matter of fact, understand and appreciate it. The authors cite the ease of use and objectivity of these criteria (ibid. p.261) for their popularity in determining capacity in practice.

A brief examination of typical psychiatric scenarios reveals that the pure understanding and appreciation conception of capacity generates some highly counterintuitive results. The following examples are similar to those in Culver and Gert's commentary, and illustrate the inadequacy of a concept of capacity restricted to these criteria:

1. A severely anorexic patient refuses naso-gastric feeding owing to an extreme fear of gaining weight. He understands what the treatment involves, and that it could in fact save his life. He appreciates that it is his decision to make and despite agreeing that without the procedure he risks death, he will not consent.
2. A patient with chronic depression needs surgery to remove an ovarian tumour. Doctors tell her that without surgery she is likely to die. The patient is able to grasp this information and is capable of discussing the options with her clinical team, but she is entirely ambivalent about whether she lives or dies and thus will not consent to the operation.

These cases would not necessarily fall under the jurisdiction of mental health legislation and therefore the assessment of capacity would be the determining factor in deciding what treatment the patient is subject to. In both cases, the understanding and appreciation criteria are satisfied. Culver and Gert suggest, however, that as clinicians they would be reluctant to allow that either patient possessed the capacity to make the treatment choice at hand (ibid. p.264). They speculate that in such cases these criteria are not sufficient to determine capacity, and they resort to appealing to the "rationality" of the decision outcome to assess whether or not the patient should be deemed competent. But requiring that a patient reach a "rational" decision in order to be deemed to possess capacity undermines the attempt to provide purely understanding and appreciation-based criteria. On the basis of the understanding and appreciation criteria alone, Culver and Gert acknowledge that patients with capacity cannot be

reliably and consistently distinguished from those lacking capacity. What then of the assessment of using or weighing information in the decision-making process: can this provide an objective criterion to justify the denial of capacity without appealing to the rationality of the decision outcome?

Evaluating the Process

> "A person is unable to make a decision for himself if he is unable…to use or weigh that information as part of the process of making the decision…" (MCA, s 3(1)c).

This criterion is consistent with an element in many tests of capacity that the MacCAT-T refers to as the ability to "reason". The way this requirement is characterised varies with different tools, but the implication is broadly similar. Assessors need to determine if the decision has been reached through a process that indicates the patient has taken account of the relevant information and options, and weighed this information in the balance. There is surprisingly little discussion in the theoretical and empirical literature on capacity detailing either what is meant by using or weighing information, or what constitutes fulfilment of this criterion. In some of the legal precedents underpinning the MCA reference is made to "balancing" information (e.g., Re *MB*; Re *C* 'Eastman' test of capacity) but there have been no direct efforts in research of mental capacity to explore this concept further. Nonetheless, I suggest that such analysis is crucial if we are to understand how this criterion is judged and unpack the implicit standards that guide or constrain this judgement in practice.

One suggestion as to how best to evaluate the decision-making process has been made in the capacity literature, but has not been followed up with any degree of conceptual sophistication. In an early paper on the notion of decisional competence to consent to treatment, Freedman (1981) argues that what it means for a person to be using and weighing information is that he is capable of providing *"recognisable reasons"* for his decision. Therefore, in order to be deemed competent a person must have reasons that are relevant to the decision he makes. He argues that such reasons

must provide a strong justification for the decision, but does not, as Charland (2001) points out, elaborate any further on what is means for reasons to be recognisable, nor what constitutes justification in this context.

In spite of Freedman's vagueness we can speculate as to what a recognisable reason might look like on his account. The requirement of justification suggests that a reason ought to take the form of an argument or inference, whereby a conclusion is drawn or decision reached on the basis of a starting set of premises. If the premises are capable of supporting the conclusion, one's reasons count as recognisable (Charland, 2001 citing Freedman, p.137). Thus despite the rhetoric of the *Re MB* ruling, if a mental impairment is suspected the reasons a patient has for refusing treatment may be subject to scrutiny in an assessment of capacity (Stauch et al., 2006, p.128). In order to establish how the requirement of recognisable reasons impacts on capacity assessments, it will first be necessary to understand the implications that an appeal to reasons has for the concept of capacity and the nature of the judgements upon which its assessment is based.

1.4. CAPACITY ASSESSMENT AS A NORMATIVE JUDGEMENT

Reasons and Normative Judgements

The very idea of appealing to recognisable reasons for a decision introduces an inherent normativity to the notion of capacity. Normativity concerns what ought to be the case, or in the case of decision-making, how one ought to think and reason. Although there is little consensus over the nature of normativity and the role it plays in judgements of another person's beliefs and actions, following Wedgwood (2007, p.22) I suggest that we do not need a definition of normativity in order to provide a theoretical account of our everyday understanding of normative terms. When we invoke normative statements about what one ought to do or think, either in the course of decision-making or in judgement about another's reasoning process, they make claims on us: they

oblige, justify, constrain or guide intentions and actions and do not merely describe what we do (Korsgaard, 1996, p.8). To say that there is a recognisable reason for a person's decision is to say that in light of the information given to him, the decision made is appropriate; it is to say that the decision is one that ought to have been made. The question we need to ask is: is the outcome recognisable as a response that follows in light of the information feeding into the process (including the relevant information, along with one's beliefs, desires, hopes, fears and so forth)? As Owen et al. (2009b, p.101) point out, understanding what the normative elements of capacity assessment are and how they are structured is an increasingly important area of philosophical, legal and psychiatric research.

The Law Commission reports that preceded the MCA were keen to allow that even with evidence of intact cognitive functioning some individuals may lack capacity, through adopting a "true choice" approach to assessment (Consultation Paper No. 129, 1993). This is the stipulation that whatever decision a patient makes, it ought to reflect his desires and choices free from coercion, compulsion or external influence. The process of using or weighing information can be construed as a matter of how the patient is appropriating that information, along with his beliefs and values, in a way that is recognisable as supporting the decision outcome. Determining fulfilment of this criterion therefore involves the normative dimensions of reasoning well or badly, correctly or incorrectly, appropriately or inappropriately or as one ought or ought not to: only if the process is normatively appropriate will it supply a recognisable reason for the outcome[18].

Does the inherent normativity of capacity judgement undermine the possibility of providing reliable, accurate and objective assessments? If the operational approach

[18] A stronger constraint on the normativity of rational processes requires that the individual be capable of reflecting on his own reasons for belief and action (Hurley & Nudds, 2006, p.12) and this is necessary for the assessment of capacity as a patient needs to be able to provide his reasons for coming to a certain decision. This is quite a stringent requirement since in normal cases of decision-making, one's reasons or justifications for choosing a certain option are not under scrutiny.

depends on being able to make impartial observations about a person's mental functioning, it appears that capacity cannot be tested in this way once we acknowledge that judgements are dependent on what counts as recognisable reasoning. This potentially undermines the descriptive cognitive approach to capacity testing and measurement. If determinations of capacity are contingent on an assessor making a normative judgement about the patient's decision-making process, a fundamental difficulty to the assessment of capacity is introduced, because what constitutes evidence for the using or weighing information criterion is only evidence in virtue of its normative import. That is to say, we can only say a person is using and weighing information if he is using and weighing this information as he should: if there are appropriate or reasonable relations between what he values, believes and decides. If such relations are not in evidence or they seem unwarranted or inappropriate, then the person cannot be said to be using and weighing the relevant information in the process of making a decision.

Clinical judgement is an inextricable part of the assessment process in spite of the drive towards operationalising criteria of cognitive functioning. This is not to deny that one can make a descriptive judgement about a normative process: the question of whether or not a person does have recognisable reasons for his decision can be answered with a binary 'yes or no' response. But although the judgement made takes the form of a statement that the patient either possesses or lacks capacity, it will be based on the clinician's understanding of what counts as (among other things) using or weighing information in the decision-making process. The distinction Charland (2001) draws between normative and descriptive dimensions of capacity implies that the normative dimension is an ethical one that can be decided after the fact of determining whether or not a person possesses capacity. However, ascertaining if a patient fulfils the "using or weighing information" criterion is a matter of making a normative judgement about the patient's process of reasoning with the relevant information he's been given, together with his epistemic and evaluative commitments. Thus a normative

element is implicit even in the descriptive dimension of capacity, distinct from the ethically oriented normative dimension suggested by Charland's analysis.

Substantive Requirements for Capacity?

If the standards by which a decision-making process are judged are content-neutral and universally applicable, the normative nature of capacity assessments need not undermine their reliability and objectivity. To use terminology borrowed from bioethical literature on autonomy, we can question whether capacity can be judged through procedural criteria alone, irrespective of the content of the decision-making process, or whether there is a substantive element to the assessment that cannot be discarded. We can thus distinguish two important questions in determining what counts as a recognisable reason for a decision:

1. Are there normative constraints on the *process* of reasoning from the relevant information and one's beliefs and values to the decision outcome? If so, do capacity judgements rely on standards governing what ought to follow from the premises of one's decision-making process to its conclusion?

2. Are there normative constraints on the *contents* of the beliefs and values that enter into the decision-making process? Again, does this impact on our notion of capacity by implying that it requires one to have particular *epistemic* and *evaluative* commitments?

The process-based approach assumes that capacity can be distinguished from incapacity solely on the basis of the standards gestured towards in the first question. Conceiving of the decision-making process in terms of what Charland refers to as its *"internal rationality"* (2001, p.136) entails evaluating the form of the process rather than its content. Although the content of the information and the decision is relevant insofar as it specifies what the decision is about, it is not epistemically or evaluatively judged: the patient's beliefs and values are taken as premises for the decision-making process and are not themselves scrutinised. This requirement suggests that in order to have capacity one must be capable of making epistemic and evaluative commitments and acting upon those commitments. It does not, however, dictate what such commitments ought to be. An assessor need not share the specific beliefs and values of the patient in order to recognise how they enter into the decision-making process, providing

reasons for the patient's decision. On this procedural construal of capacity, what is important is that the structure of the process from premises to conclusion is intact. Understanding assessment as an evaluation of the integrity of the form of the decision-making process clarifies how this demand it supposed to be met: it turns on the idea that the outcome is consistent with the person's beliefs and values, which enter the process like premises in an inference or argument. Kennedy (1997) suggests that a decision can be identified as having been made due to a pathology of belief or reasoning insofar as it is inconsistent with the patient's own previously expressed beliefs and values, irrespective of what the contents of those beliefs and values actually are. Evaluating the integrity of the reasoning process in this way might be a successful strategy in cases where capacity is temporarily impaired or fluctuating: here, expressing a choice that is out of kilter with one's own beliefs and values indicates that the decision might not have been made on the basis of a process of using or weighing the relevant information.

This approach is consistent with the principle that patients should not be deemed to lack capacity because they make a seemingly unwise decision. The principle implies that assessment is purely procedural and ought not to be substantive: no judgement is made about the particular epistemic and evaluative commitments a patient has. The intention here is to ensure that a paternalistic view of what is reasonable does not become an inextricable part of the assessment (ibid. p.322). If, on the other hand, epistemic and evaluative standards do partially determine what counts as a recognisable reason for a decision, then the judgement is contingent on what beliefs and values the assessing clinician deems it is reasonable to hold. If a patient holds seemingly bizarre beliefs or values that influence in the decision-making process, this could potentially discount the reason for the decision from being 'recognisable' by the clinician, thus introducing an element of medical paternalism to the assessment process, regarding what the patient ought to believe, want or decide.

The question I am interested in here is whether what counts as a recognisable reason can be determined independently of substantive evaluative and epistemic commitments about what one ought to believe and desire. My contention is that evaluating the process of decision-making without regard to its content does not, contrary to the rhetoric of the MCA, successfully capture all and only those cases for which a judgement of incapacity would be appropriate. Although capacity legislation attempts to minimise such paternalism, the process approach alone does not successfully track the notion of capacity. If we examine typical cases in psychiatry in which capacity is brought into question it is clear that part of what drives the instigation of an assessment of capacity in the first place is the suspicion that the patient holds particular beliefs or values that are in some way pathological, and it is to such cases that I now turn in order to argue that capacity judgements do depend on substantive and not merely procedural criteria. Here we must acknowledge that epistemic standards governing what one ought to believe and evaluative standards governing what one ought to assign value to are intrinsic to the assessment of one's decision-making capacity.

Beliefs

The presence of a mental disorder or impairment may cause a person to hold beliefs that are delusional or simply untrue. The case of Norfolk and Norwich NHS Trust v W (1996) is a pertinent example: a female patient was admitted for an emergency Caesarean section, but refused treatment because she denied that she was pregnant. Here, a misperception of reality provided strong evidence that the patient lacked the capacity to make a decision about her treatment, as she was either incapable or unwilling to acknowledge an uncontroversial fact that endangered her life. Similarly, a patient suffering from *anorexia nervosa* who was in imminent danger of death was deemed to be incapable of acknowledging facts about her weight, insisting that she was still fat and refusing naso-gastric feeding on that basis (SW Hertfordshire Health Authority v KB [1994b]). In these cases, the patients denied empirical truths that are

irrefutable to the outside observer, and were deemed to lack capacity on grounds that their mental disorders were impairing their ability to comprehend obvious facts about the world. In clinical practice, delusions are also strongly associated with assessments of incapacity (Owen et al., 2009a). In a judgement granting a hospital the right to override a patient's refusal for a medically necessary hysterectomy, the court stated: *"a compulsive disorder or phobia may prevent the patient's decision from being a true one, particularly if conditioned by some obsessional belief or feeling which so distorts the judgment as to render the decision invalid"* (*Trust A and Trust B v H* [2006] at 965). Here the falsity of the patient's belief that she was childless (she had two grown up children) meant that her refusal to consent on the grounds that she wanted children undermined her capacity.

There is a clear substantive condition at work here in judgements about the particular epistemic commitments of patients: in order to possess capacity, one ought to believe facts relevant to one's treatment decision that are manifestly true, or in any case reasonable and understandable in light of the evidence and information available. If one breaches this standard, it is questionable whether one is understanding, using or weighing the relevant information in the process of decision-making. This characterisation of epistemic requirements for capacity is deliberately vague at this point and serves as a placeholder for a further analysis of the constraints on belief to be developed in the following chapters.

The substantive epistemic conception of capacity is not without problems, not least because it threatens the commitment to pluralism in and freedom of beliefs, taken to be a central liberal ideal of our legislature. It also leads to the question of whether having capacity requires that a patient ought to believe the information given to him about his condition and the available treatment options. For many instances of mental disorder in particular, a denial by the patient that they are ill does not necessarily constitute a rejection of obvious truths. Psychiatric diagnoses are highly controversial and in the

absence of clear physical evidence of illness (in the way that a gangrenous foot is objectively verifiable), rejection of a medical opinion or proposed course of treatment does not imply the patient is failing to acknowledge incontrovertible facts. In these cases, a mental illness diagnosis combined with rejection of or disbelief in medical opinion might lead clinicians to presume there is a causal link between the two. The process of decision-making, in principle, can be influenced by unusual or eccentric beliefs without detriment to the presumption of capacity. There is an intuitive difference between disbelieving a doctor's opinion because one is detached from reality and impervious to reason owing to a mental impairment and the natural tendency to critically reflect upon and assess the medical expertise one is offered (Grubb, 2004, para 3.85). But where impairments to capacity may be more subtle, there is no line of demarcation between a rejection of medical advice made with capacity from one made lacking capacity. Hence, it is not clear whether merely unusual beliefs can be distinguished from those that potentially undermine capacity: a conceptual problem mirrored throughout the domain of psychiatric diagnosis and practice[19].

Values

Decision-making ability can be impaired subtly by disorders such as depression, where a person may be perfectly capable of fulfilling the capacity criteria as an abstract exercise in cognitive functioning whilst attaching no emotional significance to the process or outcome (Charland, 2006). The ability to understand, use and weigh the relevant information is intact but the patient is indifferent about the outcome and lacks the motivational ability to express a personal choice. In such a case it is questionable whether the patient possesses the requisite evaluative capacities to make the decision at hand, because he is ambivalent about the status of his own health or life. It is not, however, merely a capacity for making evaluative commitments that is in question in a judgement of capacity. The issue here is whether capacity requires that patients actually hold any particular values.

[19] Differentiating between legitimate religious beliefs and pathological ones is recognised as a particularly problematic area of judgement (Waldfogel & Meadows, 1996).

A series of qualitative studies examining the decision-making competence of patients suffering from *anorexia nervosa* provide a useful insight into the role played by the particular values a patient holds in the process of reasoning and weighing up information about a treatment decision (Tan & Hope, 2008; Tan et al., 2003; Tan et al., 2006). These studies document that patients suffering from *anorexia* are often insightful, coherent and able to understand the hypothetical reasoning for the necessity of forcing a dangerously underweight, malnourished person to ingest food in order to save his life or prevent serious disability. The patients involved in Tan's research were frequently able to weigh up all the relevant considerations and communicate this process of reasoning: *"all the participants were already highly conversant with the facts of their disorder, the exercise of going through information about anorexia nervosa and its treatment"* (Tan et al., 2003, p.704). On measures of understanding, reasoning and expressing a choice all patients performed comparably with normal subjects, although appreciation was unclear for two of them (ibid). Despite the imminent danger to their lives, health and physical mobility, the patients refused to consent to naso-gastric feeding. On the process-based view of capacity employed by the MCA criteria, such patients appear to fulfil the formal requirements for capacity and therefore ought to be entitled to make treatment decisions. Nonetheless, it was evident to the researchers that for decisions regarding treatment (usually related to feeding and weight-gain) many of these patients should not be deemed competent to refuse treatment.

Whilst many psychological and biological factors may contribute to this continued refusal to eat, Tan et al. (2006) established that the evaluative commitments of these patients underpin their reasoning and motivations to continue to lose weight. The authors present the myriad complexities of the reasoning behind treatment refusal in a sample of patients, pertaining to a wide range of values that are only revealed and pieced together through detailed interviews. Positive evaluations associated with *anorexia* include feelings of control, safety, distraction from other problems, a sense of

identity and community with others with *anorexia*, and feeling special or different, all of which are associated with a positively valued and valuable identity. Additionally, the significance of positive evaluative commitments to life, health and well-being are downplayed and considered less important. Thus, *"treatment refusal may occur, not because the patient wishes to die, but because of the relative unimportance of death and disability as compared to anorexia nervosa"* (Tan et al., 2003, p.704). As a consequence, patients' decision-making processes are based on assigning significant weight to the positive values associated with *anorexia*, which has the effect of outweighing the negative connotations associated with disability and death. It is not that these patients do not acknowledge or believe the risks involved in continued treatment refusal, but when weighed against their strong evaluative commitments to thinness, control and so forth, these concerns are insignificant: *"these… are not products of a lack of understanding, but instead influence the use of understood information and the weight placed on it in coming to a decision"* (Tan et al., 2006, p.277). Patterns of evaluation serve as highly significant weights in the decision-making process, providing reasons or justifications in themselves for making a choice to refuse treatment.

What is perplexing and perhaps indicative of incapacity in these cases is that the overriding evaluative commitments clash with or undermine some of our most deeply held values: it is difficult to comprehend how the value of thinness could be accorded greater weight than the value one places on one's own life, for instance. We cannot avoid concluding from this that patients expressing such commitments are not merely deviating from values that are usually held, but rather that they are in error: they are not making the particular valuations that they ought to. Tan and Hope (2008) conclude that capacity does in fact require certain evaluative commitments to be made by the patient. Thus the values a person attaches to relevant information such as risks and benefits of a treatment play a significant role in determining how that information is used and weighed in the process of decision-making. The implication is that there are evaluative

standards impinging on the judgement of whether a patient fulfils capacity criteria: it requires that one's values ought to be reasonable, good, or otherwise normatively appropriate.

It is a hallmark of a liberal, broadly inclusive democratic society that wide variations in religious and cultural values are not only tolerated but embraced as expressions of diversity. However, whilst many values are deemed acceptable or reasonable even if they are not shared, others are not and may be perceived as being indicative of psychopathology, thus potentially undermining the capacity of the individual who holds them. A comprehensive account of the nature and role of substantive evaluative commitments in capacity judgements can be found elsewhere (e.g., Holroyd, 2010), and it is interesting to note the convergence among authors (Tan & Hope, 2008; Martin, 2007) that an understanding of patients' values is critical to the determination of their capacity.

Compulsion

Some mental disorders or impairments may disrupt decision-making ability owing to a compulsion or inescapable drive to behave in a particular way. In the case of *Re MB* the patient consented to a procedure only then to refuse repeatedly once the operation was imminent, on account of her extreme and overwhelming fear of needles. The judge ruled that her phobia temporarily impaired her capacity to make a decision: *"…a panic fear of needles dominated everything and at the critical point she was not capable of making a decision at all"* (*Re MB* at 427).

Tan et al. (2003; 2006) found that even in patients who were determined to recover from their illness, a strong compulsion to reject food was overriding their sincere desires for recovery. The notion of compulsion is familiar in Obsessive-Compulsive Disorder, which is characterised by repetitive compulsively driven behaviour that is beyond the control of the individual to cease or alter. Thus certain mental disorders

may deprive a person of the ability to make a genuine choice, independent of compulsion or coercion. In these instances the patient is not expressing a decision arrived at competently and although free from external influence he will not be making an autonomous decision. In terms of the MCA criteria the disorder diminishes the ability to weigh information relevant to the decision, as the compulsion may obviate any other considerations from entering into the reasoning process and generate an impediment to the patient's will. A refusal of treatment under these circumstances is not a competent refusal, and may thus be overruled (Grubb, 2004, para 3.87). Only if the compulsion or phobia acts to *"paralyse the will and thus destroy the capacity to make a decision"* will a judgement of incapacity be justified (*Re MB* at 437). The compulsion may not necessarily be acknowledged by the patient in order to be considered an impediment to the will, as evidenced in the case of a depressive woman refusing to eat (B v Croydon Health Authority [1994a]). The patient's acute self-awareness, insightful self-analysis and rigorous articulation of her reasons for refusing to eat actually assisted the judge in deciding that her desire to refuse treatment was the result of a compulsion induced by her mental illness. Again, whilst I will not be focusing on the way in which disorders of volition may compromise capacity, this brief analysis of compulsive behaviour supports the claim that judgements about using or weighing information intrinsically involve an appeal to the reasonableness or appropriateness of the factors influencing the decision-making process.

Norms of Judgement in Capacity Assessments

For some mental disorders then, serious doubts about the capacity of patients are legitimate even when their cognitive functions appear to be intact, or they display consistency and coherence in their beliefs, desires and decisions. This is particularly the case when patients make bizarre or dangerous choices about their treatment. This could be the result of the decision being made on the basis of abnormal or patently false beliefs, or strong evaluative commitments that potentially indicate the presence of

a mental impairment. Epistemic and evaluative standards do come into play when assessing whether a person is using or weighing information.

To take the example of the patient suffering from *anorexia*: her continued refusal to eat or to permit naso-gastric feeding does logically follow from her deeply held beliefs and values. In a sense, she has a reason for her decision. It makes sense that she refuses treatment if she sincerely believes she is, for example, overweight; or she places such a high value on thinness that this overrides concern for her physical health or even her own life. However, it is questionable whether these reasons, whilst recognisable as having the form of reasons (i.e., she makes this decision because of them), are indicative of a competent decision-making process. Indeed, the very fixation and conviction she exhibits suggests some mental impairment or pathology might be at work that would undermine her capacity to make autonomous decisions.

This insight demonstrates that in attempting to assess the decision-making process, an implicit normative standard is operating that disciplines judgements about what reasonably ought to follow from the information provided about a treatment decision. In this case, despite the logicality of the decisional process, we might consider the fact that the patient does not seem to place due weight on the severity of her disability or risk of death to indicate she is not using or weighing the relevant information. Crucially, this is a judgement that is based in part on a conception of what one ought to believe and want: in this case, that she ought to believe that she is dangerously underweight and place a high value on her own life. Procedural criteria alone are thus insufficient for distinguishing capacity from incapacity.

There are substantial political, ethical and legal issues associated with the claim that capacity requires substantive evaluative commitments but these are beyond the scope of this thesis (see e.g., Martin, 2007). I will focus instead on the procedural and epistemic requirements for capacity, examining the structure of the decision-making

process and the constraints on what one ought to believe in order to possess capacity. Expanding upon the idea that capacity concerns one's reasons for a decision, I turn in the next chapter to discuss the large body of empirical and philosophical literature on the topic of rationality. I examine the normative structure of reasons and seek to ascertain whether there are objective, universal standards underpinning normative judgements about the process of decision-making.

2. RATIONALITY AND REASONS

2.1. RATIONAL PROCESSES

Reasons and Decision-Making

The possibility of defining capacity according to purely procedural criteria has been undermined by recognition of the fact that ascertaining whether an individual has recognisable reasons for his decision requires some evaluation of the beliefs, values and desires that enter into the decision-making process. In this chapter I unpack this claim further by discussing some influential empirical and philosophical views on what it means to engage in a process of decision-making and how we form reasons for our decisions and actions. The aim is to shed light on how we are to understand what counts as using or weighing information in a deliberative process of decision-making, or conversely, what circumstances would undermine the claim that a person was using or weighing information in coming to a decision. Part of my concern is to ascertain if our decision-making processes are subject to any universal normative standards, and to query the implications for the reliability and objectivity of capacity assessment if not.

There are strong parallels between the procedural conception of capacity and a prominent view in research on reasoning and decision-making referred to as the "standard picture" (Stein, 1996). I utilise this established literature firstly to examine why the procedural conception of human reasoning is inadequate and secondly to determine whether any content can be given to the idea that there are substantive epistemic standards governing normative evaluations of the decision-making process. Procedural norms do possess an advantage over substantive norms, in that they are amenable to codification in principles. This means they can be applied in any relevant situation to prescribe appropriate moves or proscribe inappropriate ones. The universal nature of logical standards renders them ideally suited to the task of providing clear, objective standards by which to judge a person's reasoning. By contrast, epistemic standards do not appear to be codifiable in this way. Towards the end of the chapter I

advance a relativist worry that if epistemic standards are intrinsic to judgements about a person's reasons and decision-making as I have suggested, without codification there is scope for such standards to differ across individuals or groups: a problem that raises a *prima facie* obstacle to attempts to provide objective standards for assessments of decision-making capacity.

Rational Relations

In everyday interpersonal encounters we unreflectively and spontaneously see meaning, intention and motivation in the movements and utterances of others. This interpersonal understanding rarely takes the form of an explicit deliberation as to what people mean by their words or why they decide or act as they do. However, evaluating a person's decision-making process for the purposes of determining whether or not he has capacity requires a more critical consideration of his decisions, which necessitates an attempt to grasp his reasons for that decision. This is a matter of seeking an explanation as to why the patient decides as he does, and when we seek to identify a person's reasons we may make reference to the beliefs and desires we take him to have[20]. Whilst many different psychological, affective and motivational factors will affect a person's decision-making processes, here I wish to focus on those elements that enter into explanations for a decision in terms of that person's reasons.

Appealing to an agent's beliefs and desires may serve to generate a reason explanation for a particular instance of behaviour. A reason explanation is one that makes the patient's behaviour intelligible to the clinician, but this kind of explanation differs from explanations of phenomena found in the natural sciences. To use Dennett's (1987) terminology, we can adopt the "Intentional Stance" towards agents to explain at least some of their behaviour. Why a person behaves as he does is not

[20] The approach I am adopting does not imply that understanding the behaviour of another involves a perception of physical movements devoid of psychological intention, followed by an inferential attribution of meaning and significance. Rather, behavioural phenomena are expressive and visible to others (Zahavi, 2005, p.151). Much writing in the phenomenological tradition takes this view of intersubjectivity and although this discussion is not directly informed by such approaches, nothing in what I suggest is incompatible with this view.

explained using the vocabulary of physical-causal concepts but by employing the folk psychological concepts of belief, desire, intention and so forth. This folk psychological approach has dominated philosophical theorising about interpersonal understanding and explanation, and although I consider the framework to represent an impoverished view of ordinary, everyday understanding of intentional action, the concepts it employs are nonetheless a useful starting point from which to begin the project of establishing whether there are epistemic constraints on what counts as a recognisable reason for a decision[21]. In this chapter I start from the assumption that judgements about capacity require the clinician to grasp an explanation of the patient's decision in terms of his reasons. The evidence upon which the clinician must make this judgement comes from the patient's linguistic utterances and his behaviour, together with whatever knowledge the clinician has about such relevant contextual information as his background, history of previous choices and value system.

Constructing reason explanations from attributions of beliefs and desires serves an explanatory purpose that will help clarify what having a reason for a particular decision or action might entail. A few caveats are in order here. I am concerned with the question of whether there are normative epistemic standards constraining the attribution of beliefs, and I will therefore not consider the role of evaluative judgements in intentional attributions of desires and values. Furthermore, I will not touch upon the considerable philosophical and psychological literature concerning self-knowledge and first person reason-giving (see e.g., Bortolotti & Broome, 2008; Moran, 2001) as I wish to examine the attribution of reasons from the position of the third-person making judgements about another's intentional behaviour. Although for the purposes of exposition and tractability we need to focus on simple chains of inference and belief/desire pairings, seeking to grasp an agent's reasons is a project best construed more broadly. This is especially true for judgements of capacity: an observer needs to

[21] See Bermúdez (2005) for a detailed account of the influence of folk psychological concepts in both philosophical and empirical models of the mind.

be aware of a great number of relevant factors entering into the decision-making process if he is to make a well-informed judgement as to whether this process is sufficiently reasonable to allow the decision to stand.

Reason explanations function by enabling us *"to see the events or attitudes as reasonable from the point of view of the agent"*[22] (Davidson, 1982, p.169). Conversely, in the absence of recognisable reasons we might think of an action or utterance as being irrational. Reasons are in principle recognisable to the person himself: a requirement that reflects the need for an agent to be able to articulate and weigh up his own reasons in order to be deemed to possess the capacity to make a decision. In the first person case, if one wishes to act in one's own best interests there is an intrinsic connection between comprehending what one ought to do in order to fulfil that wish and making the decision[23]. In normal circumstances intentions are formed on the basis of what one thinks one ought to do: the normative judgement (for example, that one ought to φ) supplies the motivation for intending to φ (Wedgwood, 2007, p.33; Broome, 1997, p.141-2). There is thus an essential connection between the judgements one makes about what one ought to do and one's reasons and motivations for action. Whilst this is a claim that is open to debate, I shall presume that it suffices to outline the sense in which one's own reflective judgement might normally play a role in motivating intentions and deliberatively making decisions. The purpose of seeking reason explanations is thus to attempt to grasp the first-person point of view to make sense of the reasons a person has for his decisions and actions from a third-person perspective[24].

[22] I shall leave it an open question as to whether such reasons provide the cause of the action, although see especially Davidson (2001a) for some considerations about the relation between reasons and causes.

[23] Contemplating what one ought to do is a deliberative question and Wedgwood (2007, p.25) suggests that if one is rational it is a question that arises when considering a decision about what to do (footnote 10).

[24] Frankfurt (1977) outlines a similar requirement of second-order self-reflection in his ethically oriented philosophical account of autonomy.

Gerrans (2004) and Bayne and Pacherie (2004) criticise the reason approach to explanation, particularly in psychiatry, arguing that focusing solely on person-level concepts such as beliefs and desires excludes cognitive and neurobiological facts about a person's psychological functioning from entering into explanations of his behaviour. This is because they presume that non-propositional content (which they take to include perceptual experiences) cannot be accommodated within a reason-based account of behaviour. However, the central importance I am placing on the role of reason explanations in the understanding of decision-making behaviour is not incompatible with the idea that there are sub-personal causal factors involved in the formation of intentions and the performance of intentional action, nor that ordinary behaviour is not disciplined by careful reflection on one's own reasons. The reason approach to explanation does not preclude the causal relevance of sub-personal processes, cognitive biases, non-propositional content or even postulated Freudian unconscious drives. Rather, this emphasis on reason explanation is based upon the claim that an explicit and deliberative understanding of the processes by which an agent comes to a decision is necessarily pitched at the intentional level, and that this understanding is normatively structured. Indeed, the prime focus of this thesis is on distinguishing situations in which a person's decisions and behaviour are explained by reasons from those that might instead be explained or accounted for by some mental disturbance, cognitive deficit or neuropathology that impairs the normative structure of the reasoning process. Accepting that reasons provide explanations of intentional behaviour does not commit one to any particular position regarding the metaphysics of mental content. Whatever one's view regarding the constitution of the psychological realm folk psychological vocabulary has a clear practical utility if we are seeking to understand a person's deliberative process: we can explain reasons for behaviour, and indeed decisions, by reference to what agents think and want (Bortolotti, 2004a, p.360).

What lies at the heart of the utility of reason explanations is the idea that propositional attitudes are capable of bearing semantic and logical relations to one another (Millar,

2004). Beliefs can serve to justify, support or undermine other beliefs; they may lead to the creation or cessation of desires; doubts about the truth-value of a particular proposition may make one reluctant to form a belief about it; a fear may undermine a desire and thus thwart action, and so forth. Understanding the nature of these inter-relations between propositional attitudes is necessary to the project of exploring what constitutes a recognisable reason for an action or decision. Let us take a few hypothetical examples to clarify what I mean by the relatedness of propositional attitudes.

1. A patient is convinced his wife is being unfaithful to him. When asked why he thinks this, he says "I know because the number 23 bus just went past the window."
2. A patient with a minor scalp wound believes the FBI has sewn a radio into his skull and repeatedly attempts to sue the government (described by Gold & Howhy, 2000).
3. An artistic patient with bipolar disorder refuses anti-psychotic medication because it diminishes his creativity.

In example 1 a belief is resolutely maintained, but the basis upon which the agent asserts his reason for holding that belief is irrelevant to the belief itself. There is no rational relation between the supposed marital infidelity and the timing and route of a bus, nor any intelligible significance of the bus number to the agent's relationship with his wife. In short, the agent's avowed reasons cannot in fact be reasons for his belief: there is no conceivable logical or semantic internal connection between the belief and the reason the agent gives for it. This example demonstrates how a lack of obvious rational relations between intentional states precludes the provision of intelligible reason explanations for actions and utterances.

Example 2 is of a type that occurs in individuals with delusions of persecution, often diagnosed with schizophrenia. The belief that is formed is fantastical, delusional in its intensity and immune to counterevidence or argument. Nonetheless, if we take the delusional belief as a given, perhaps evidentially supported in part by the scar on the patient's scalp, then his persistent efforts to sue the government are at least intelligible to the observer: if I had such a belief then such a course of action would, even if ill-

advised, be a recognisable response to my predicament. Thus, although the belief is bizarre and the resultant action potentially detrimental to the individual, from the outside we can at least formulate a reason explanation for the person's actions: he is attempting to sue the government because he believes that the wound in his head is evidence that the FBI is persecuting him. In this instance whist the process of reasoning is at least intelligible, it is based on poor epistemic commitments.

In the final example not only can a reason explanation be provided for the action or decision, but also it seems a reasonable and appropriate outcome. This is so even if it goes against prevailing medical opinion. In addition to being recognisable the reasons given for the action are understandable, particularly in light of the high value the patient places on his creative flair. Thus while perhaps deviating from medical norms in his evaluative commitments, the patient is not considered to have made a normative mistake in his valuations and beliefs. It is unlikely in this kind of scenario that the patient's capacity would be undermined.

This brief anatomy of reason explanations suggests that there is a normatively rich structure to the reasons that explain or account for one's actions and decisions. I turn now to discuss the myriad theoretical positions relating to human reasoning and the notion of rationality, to ascertain what normative commitments underpin judgements about a person's reasoning and decision-making process.

The Concept of Rationality

Thus far I have avoided use of the term 'rationality' in discussing the idea that normative standards underpin judgements of capacity. This is because the polysemous concept of rationality has a wide range of applications and no concise definition even within the literature of a single academic domain. Here, I use the term 'rationality' with the caveat that it is intended loosely to denote a pre-philosophical sense of picking out processes that provide reasons for beliefs and actions, from an observer's point of

view. I suggest then that what counts as a recognisable reason is determined by the standards of rationality imposed on the process of decision making. These standards concern whether or not there is a rational relation between all the input factors such as information, beliefs and values, and the decision outcome.

In day-to-day life our thinking, actions and decision-making rarely follow a clear serial structure. Much of what we do is not the result of a well-reasoned argument or process that takes into account all or even most of the relevant information, and actions may be performed and decisions made unconsciously or reflexively, particularly when in familiar situations. Many of the factors involved in making a decision may not be known to us unless we are interrogated about our reasons. It is only at this point that these factors, or our self-conscious perceptions of them, may be linguistically described in the form of a reason explanation using the concepts of folk psychology. Dreyfus and Dreyfus (1986) refer to this folk model of thought and reasoning as the “Hamlet model” (p.28), reflecting the idea that in decision-making a person analyses, weighs up and self-consciously considers the available options in a deliberative fashion. They point out that this model does not necessarily mirror the structure of thought and reasoning, arguing that much of what we know and how we go about making decisions does not take the form of explicit propositional knowledge that would be a candidate for being a reason. Furthermore, we may construct post-hoc rationalisations of our behaviour that do not accurately reflect our motivating reasons for a particular decision. However, a determination of capacity does require evidence of a deliberative reasoning process leading to the decision, and so my concern here will be with the beliefs, values and intentions that the patient himself is capable of acknowledging and that are observable from a third-person perspective, even if there are additional implicit influences on his decision-making.

In exploring the notion of rationality I am not seeking to claim that decisions that are not made on the basis of a clear-cut process of reasoning are thereby irrational or

indicative of incapacity. For a start, there are significant differences between the concepts of capacity and rationality. Rationality is a global concept which, in the philosophical and psychological literature at least, reflects the broad underlying reasoning competence of an individual across different contexts and that need not be manifested consistently in all situations. A judgement about capacity is, on the other hand, a judgement about a patient's ability to make a decision in a particular instance on the basis of an information-weighing process. For the purposes of framing the present discussion it will suffice to say that rationality requires an abstract, domain-general and flexible ability to act for reasons; and the term refers to the active exercise of this ability to act for reasons, rather than as a latent capacity. I will also not be making use of the well-established distinction between theoretical and practical reason, which distinguishes what it is rational to think or believe from what it rational to do or intend (see e.g., Mele & Rawling, 2004). This is owing to the fact that decision-making capacity encompasses both of these types of reason and I will not be pitching any claims at a level that distinguishes theoretical from practical rationality.

The Rationality of Reasoning Processes

The tasks of understanding and assessing the processes of decision-making have spawned a vast research literature throughout several disparate academic disciplines. The project predominant in psychology and biology has focused on exploring how humans do in fact make decisions and form beliefs, whereas the project more common to logic and economics aims at setting out ideal standards and models of inferential reasoning. Philosophical discussions of the nature of rationality have tended to conceptualise it as an ability to draw inferences from premises to conclusions and to recognise the validity of such inferences (Raz, 1999, p.357). Whilst rationality is conceived as an ability or capacity, it is manifested through the process of reasoning; hence the concepts of rationality and reasoning are closely related. Reasoning can be thought of as a process that generates behaviour (Hurley & Nudds, 2006, p.5), and which might require conceptual and linguistic abilities. It can be explicit and capable of

being articulated or it may take the form of a tacit process the mechanism of which is opaque to the subject. Kacelnik (2006) considers the focus on process to belong primarily the domain of philosophical and psychological discussion on the topic of rationality. He terms this “PP-rationality” and distinguishes it from two other conceptions of rationality. Firstly, “E-rationality”, most frequently used in economics, takes rationality to consist in behaviours that ensure the maximisation of expected utility for an individual. E-rationality deals with observable actions and outcomes and it is these that determine whether an individual is rational or not. Secondly, rationality as construed by biology is referred to as “B-rationality” and this is concerned more with whether the behaviour is evolutionarily adaptive or beneficial to the individual (ibid. p.17). Like E-rationality, this conception of rationality focuses on behavioural outcomes. The possibility of scrutinising someone’s ends presupposes the existence of objective standards or values regarding what it would be rational to do or think. This is precisely what judgements of capacity are supposed to avoid, although as I have suggested the possession of capacity does in fact depend in part on one’s epistemic and evaluative commitments. By contrast, the PP-rationality of a behaviour or action is evaluated in terms of the process that led to it being performed, irrespective of the appropriateness of the ends or the outcome. As Kacelnik puts it, in the language of cognitive psychology PP-rationality is about the rationality of information processing rather than the rationality of actions themselves. This approach mirrors the process-based conception of capacity, in which it is the structure of the inference from premises to conclusion that is subject to being normatively judged irrespective of the decision content. Additionally, whilst PP-rationality is concerned with the formal relations between the thoughts and values that enter into a process of reasoning, it is contrasted with decisions arrived at through emotion, religious or spiritual faith, conformity to authority or arbitrary choice (Hurley & Nudds, p.5).

It is important to distinguish those kinds of psychological process that might be subject to judgement by normative standards of reasoning from those that are involved in

causing behaviour and action, but that nonetheless are not evaluated in an assessment of the rationality of an agent's decision-making process. Some processes leading to the generation of behaviour are 'arational', by which I mean that it is not possible for the individual himself to evaluate rationally such processes as being appropriate, reasonable or otherwise. Behaviour may be generated by physiological or neurochemical processes that do not fall within the domain of rationality, or there may be tacit, unconscious psychological processes governing action that are not subject to any degree of conscious control, intervention or analysis. Heuristics and cognitive biases are two such examples that have been heavily researched in the empirical psychology literature (Gigerenzer & Selten, 2001; Gigerenzer et al., 1999) revealing implicit shortcuts and resource-saving strategies we normally use when engaged in reasoning tasks (Tversky & Kahnemann, 1974; Kahnemann & Tversky, 1972).

On the conception of rationality I am employing, which requires the agent to act for reasons that are recognisable as such, heuristics and cognitive biases do not fall within the range of processes that could in principle supply an agent with reasons for his action. The unconscious employment of such strategies may well explain an agent's behaviour but the heuristic is not a reason for the behaviour: its use is not a reason the agent himself has for making a particular judgement or choice[25]. Thus I do not consider such heuristics to be part of the reasoning process that is open to evaluation by a capacity assessment, although their use may in fact cause behaviours or outcomes that can be interpreted as rational or irrational.

Alternatively, processes that are in principle rationally describable may fail to be rational, in the sense that they produce mistakes or errors, effecting the wrong or inappropriate kind of behaviour or action. Such processes may be termed 'irrational,' denoting that they remain rationally evaluable in principle and that something appears

[25] Whilst there are arguments that identify the normativity of such heuristics in evolutionary and adaptive functions (e.g., Danielson, 2004), it is beyond the scope of this discussion to argue that these constitute rationally evaluable reasoning strategies.

to have gone wrong in the process of reasoning. These kinds of rational process are of interest to clinicians assessing capacity, particularly where cognitive functioning appears intact. The relationship between arational and rational processes is a complex issue that has concerned philosophers of mind and psychologists alike and I do not intend to impinge upon debate about the interface between reasons for and causes of action (see e.g., Bolton & Hill, 2004; Thornton, 1997; Davidson, 1982). Here I merely wish to draw attention to the relationship between reasoning and rationality, and highlight that the pertinent focus for understanding what underpins judgements of capacity is on the kinds of psychological processes that are amenable to rational evaluation.

The PP-rationality of processes of reasoning has been termed “instrumental rationality” in philosophy and psychology, although the term has different connotations in economics (Hurley & Nudds, p.7). Instrumental rationality concerns the way in which agents select the means necessary to achieve a given end or outcome. An agent would be instrumentally rational to the extent that he takes the means necessary to achieve his ends, irrespective of what those ends are (Wallace, 2008). There is a benefit to thinking of capacity assessment in terms of instrumental rationality, in that focusing on process ensures that an ‘appropriate’, ‘correct’ or ‘good’ outcome is not automatically accepted as being indicative that the patient possesses capacity, or indeed the converse. Instrumental rationality requires that the outcome is reliably arrived at by a rational reasoning process: it is not sufficient that the patient complies with a doctor’s recommendation as this may be accidental or resulting from a warped reasoning process. We are getting closer to being able to frame the question as to what standards determine whether a decision-making process provides recognisable reasons. This instrumental approach to rationality is entirely compatible with the intention behind capacity assessment, in which the perceived wisdom or rationality of the actual decision is not supposed to influence the capacity judgement.

2.2. PROCEDURAL RATIONALITY

The Standard Picture of Rationality

Empirical research into reasoning, particularly in the cognitive sciences, has for the most part been guided by a framework whereby the degree to which participants are rational is determined according to whether their reasoning, judgements and decision-making accord with the dictates of a normative ideal. The experimental methodologies, analyses and theorems developed from reasoning task experiments are framed around the idea that there are normative standards of rationality that one ought to obey and that failure to do so results in one committing an error, often in the form of a logical fallacy (see Stein, 1996 for an overview of the field). For instance, cognitive psychology in this area has tended to focus on conditional and syllogistic reasoning, with participants' judgements in the tasks being taken to denote success or failure in reasoning correctly: clearly a normative analysis on the part of the researchers (Eysenck & Keane, 2005). Gigerenzer (2006) refers to the ideal of rationality upon which this idea is based as the "LaPlacean demon" of unlimited time, cognitive resources and omniscience, and it is a model based on the laws of probability and logic that has pervaded thought on human reasoning and inference at least since the Enlightenment (ibid. p.117).

Two camps have emerged from this empirical literature, based on opposing assumptions about whether or not humans can be considered rational. The rationality thesis asserts that humans in general are rational and that errors in reasoning can be attributed to performance problems such as accidental mistakes or limitations in computational capacity. By contrast, the irrationality thesis asserts that humans are fundamentally irrational, and that errors made in reasoning tasks are due to poor competence in adhering to the principles of reasoning rather than merely resulting from performance errors. The pioneer of the Wason selection task, which is used in countless reasoning experiments, argues that *"irrationality…is the norm. People all too readily succumb to logical fallacies"* (Wason, 1983, p.59). Stein points out that for all

their differences, both of these theses are based upon the same assumption: that what it is to be rational is to reason in accordance with the rules of logic and probability theory. This is referred to as the "*standard picture of rationality"* (Stein, 1996, p.4) and it is generated from the conjunction of two ideas. Firstly, that reasoning is based on principles that dictate how we ought to draw inferences, and secondly that these principles are derived from formal logical rules and axioms of probability theory.

This conception of rationality is not, however, a modern phenomenon resulting from advancements in logical theory. Hume noted the isomorphism between deductive reasoning from propositional premises to a conclusion and the transition from one psychological state to another, and this led him to believe that the rationality of reasoning processes was explained by the validity of deductive arguments (Smith, 2004). Research based on the standard picture has largely concerned inferential relations: drawing conclusions from given sets of premises, recognising the validity of stated inferences and enabling an observer to make judgements of correctness or to make choice preferences on the basis of those principles. This view of rationality has an historical precedent quite independent of empirical research, in the writings of Aristotle on the practical syllogism and more recently by Davidson, who discusses the reasons an agent has for his actions as being akin to an inferential argument:

> *"*If we can characterise the reasoning that would serve, we will, in effect, have described the logical relations between descriptions of beliefs and desires, and the description of the action, when the former gives the reasons with which the latter was performed. We are to imagine, then, that the agent's beliefs and desires provide him with the premises of an argument" (Davidson, 1978, pp.85-86).

The conception of rationality implicit in the standard picture is referred to by Bermúdez as "procedural rationality": *"subjects are procedurally rational to the extent that they reason in accordance with familiar deductive principles as modus ponens, modus tollens, contraposition…together with…basic principles of probability theory…"* (Bermúdez, 2001, p.496). What logic and probability theory have in common with regard to rationality is their basis in considerations of the truth-functions of propositions:

they are concerned with what would be logically entailed by the truth of a proposition, irrespective of its actual truth-value. Principles of procedural rationality deal in relations of implication and entailment, providing a formal structure for setting out what follows from a given premise or set or premises. For example, if we take the proposition 'that-*p*' to be true, it follows that 'not-*p*' cannot also be true: this would contravene the law of non-contradiction. Similar rules are derived from probability theory. For example, the conjunction rule states that the probability of the conjunction of two events occurring is always less than the probability of a single one of those events occurring. These logical and probabilistic principles provide a formalised framework for studying inferential reasoning and the relations that obtain between propositions.

On this view, the extent to which agents are deemed rational is determined by their behavioural conformity to what this normative ideal dictates. A classic example from the empirical literature on human reasoning demonstrating this presupposition can be seen in the way the conjunction experiment devised by Tversky and Kahnemann (1983) is formulated and discussed by the researchers themselves. The hypothetical scenario drawn up for this experiment concerned a woman, Linda, and gave some details of her history such as a passion for left-wing politics in her youth. The reasoning task required the participants to rate the probability of Linda being a feminist, a bank teller, or both. Many participants committed a conjunction fallacy, believing the probability of the conjunction of the two conditions (bank teller and feminist) to be greater than the probability of a single condition (bank teller): a mistaken inference. The principle being violated is the conjunction rule, part of probability calculus, which states that the probability of a conjunction being true cannot be greater than the probability of its individual conjuncts. To suggest that participants commit an error when they violate the conjunction rule is to derive a specific normative prescription on what one ought or ought not to believe from the rules of statistics and extensional laws of probability (ibid. p.294). The fact that the probability of a proposition being true cannot be greater than

the probability of its constituents being true translates into a normative constraint on what beliefs one ought or ought not to hold.

On a procedural conception of rationality, logical axioms dictate what it is correct to infer from a given set of premises, and conforming to the principles derived from these rules in drawing inferences is just what reasoning correctly is[26]. Correctness here is determined by logical relations between the premises rather than in facts about the world: it is derived purely from the internal relations between the propositions that comprise the premises for the reasoning process. The correct option or choice preference in the experimental scenarios devised is identified algorithmically by the application of procedural principles to the starting premises[27]:

> "...often rationality is taken as equivalent to logicality. That is, you are rational just in case you systematically instantiate the rules and principles of inductive logic, statistics, and probability theory on the one hand, and deductive logic and all the mathematical sciences, on the other" (Flanagan, 1984, p.206).

If a given set of beliefs and desires is taken as the starting premises, the application of a system of logic and a theory of decision-making can define functions from these to other beliefs and intentions that are entailed by the original premises (Heal, 2008, p.49). The implication is that logical validity from premises to conclusion is a necessary and sufficient condition for the normative correctness of the reasoning process, such that making a valid inference guarantees that the inference is rational, irrespective of the truth-values of its propositions.

There are significant epistemic and cognitive resource limitations on our reasoning capacities, and we seldom follow through all the logical entailments of our beliefs or

[26] Principles of reasoning are based on the rules of logic but are not identical to them: reasoning is based on relations between beliefs whereas logical rules apply to statements and propositions. This distinction is perhaps best maintained by referring the relations dictated by logic as rules and those of reasoning as principles (Stein, 1996, p.5).

[27] Several theories of reasoning have been developed on the basis of procedural principles and these epitomise the claim that a logical ideal of rationality forms a normative standard by which actual reasoning competence can be judged. Abstract-rule theory (Braine et al., 1984) suggests that people implicitly possess 16 mental rules, which account for performance on syllogistic reasoning tasks. Oakesford and Chater (2001) consider that a probabilistic approach provides for a more realistic normative theory or reasoning. In both cases rationality is conceived of as a matter of the inferential structure of the reasoning process.

exhaustively weigh up the expected utilities of different choice preferences before coming to a decision. Yet whilst it is clear that we do not have the information processing capacity to adhere to the principles of logical reasoning and probability theory, the standard picture takes the dictates of procedural rationality to be a normative ideal[28]. In practical terms, we may still be procedurally rational to the extent that our beliefs are consistent with one another, or at least, not obviously inconsistent, and that we seek to correct our decisions or choice preferences if a breach of procedural principles becomes salient. There are empirical precedents for this approach to rationality, for example, a decision-theoretic approach takes rationality to be a matter of maintaining internal consistency within one's mental economy (Mele & Rawling, 2004, p.4). Bermúdez (2001) cites this *"norm of consistency"* as retaining the standards of procedural rationality, as it generally ensures we do not hold openly contradictory beliefs or commit obvious logical fallacies in our reasoning.

According to the procedural conception of rationality one's ability to reason, conceived of as an ability to conform largely to logical principles, has two clear characteristics. Firstly, it has what Hurley and Nudds refer to as *"flexible generality"* (2006, p.11): the reasoning process can be applied in different contexts and environments, used in counterfactual thinking and in the formation of beliefs and actions. The content of the inference, what the agent is reasoning about, is therefore irrelevant to any determination as to whether or not the agent is reasoning correctly, as it is only the logical relations between premises and conclusion that matter. Secondly, the standard of correctness for the inference is supplied by its logical validity, derived from the instantiation of abstract, universally generalisable rules that stand irrespective of the specific content of the inference (Searle, 2001, p.21). Being instrumentally or 'PP' rational requires that the processes one uses in decision-making adhere to these rules.

[28] Stich (1999) argues that Dennett subscribes to a view of rationality as an ideal, to the extent that the ideal is so far removed from what we are actually capable of that one is forced to adopt an instrumental view of mental entities such as beliefs and desires: the applicability of such concepts to human minds and behaviour cannot possibly be true if they are subject to such stringent constraints of rationality.

At first glance this procedural conception of rationality looks to be aligned with the intention of the MCA to focus on evaluating the process of decision-making as opposed to the judging whether or not the outcome is objectively good or wise[29]. Because procedural rationality is silent on the matter of the truth-values of propositions it cannot determine either the B- or E- rationality of an individual, as it makes no reference to the appropriateness of an outcome in relation to the broader situation or environment in which the process is occurring[30]. What is more, appealing to generalised, context-free principles appears to ensure that the normative standards of rationality by which an individual's reasoning is evaluated are objective and universally applicable. Codified principles of logic are as well established and as secure as mathematical truths, possessing the same fixed and universal status (Stein, 1996, p.4). If human reasoning is governed by the normative standards dictated by procedural rationality, there is *prima facie* plausibility in the idea of codifying these principles as benchmark standards that would enable us reliably to distinguish between those who are successfully engaged in a reasoning process and those who are not. It follows that if the way we ought to reason is determined by adherence to the principles of procedural rationality, we have a strong basis for an algorithmic and content-neutral normative standard for judging the decision-making process.

Limitations of Procedural Rationality

Ascertaining whether or not we do actually reason in accordance with, and only with logical principles has spawned a significant philosophical and empirical research literature. One of the most famous and empirically robust reasoning experiments demonstrates that we consistently commit the fallacy of affirming the consequent. The

[29] Gunn (1994) comments that despite the Law Commission report discussing approaches to capacity choosing not to make rationality a component of capacity, the identification of logical fallacies in the decision-making process may well serve as evidence that a person's capacity is impaired (p.25), indicating that the formal standards of procedural rationality may play an implicit role at least in capacity judgements.

[30] Reasoning could be procedurally rational but nonetheless fail to result in a behaviour that fulfils substantive criteria such as the maximizing of inclusive fitness (Biological rationality) or expected utility (Economic rationality).

Wason selection task (Wason, 1983) was designed to test participants' ability to reason according to the principle of *modus tollens*, requiring them to make selections that would disprove a rule of the form 'if *p* then *q*'. A high percentage of people commit the logical error of thinking '*q*, therefore *p*'. When the task is set up with realistic content rather than abstract premises, participants are less likely to commit the fallacy (e.g., Johnson-Laird et al., 1972), suggesting that reasoning competence may not be grounded in adherence to abstract principles but rather be attuned to the actual situations and contexts in which decisions are made and intentions formed.

Evans and Over (1996) discuss specific criticisms levelled at research on reasoning that is based on a procedural conception of rationality. The "normative system problem", first brought to light by Cohen (1981, cited in Evans and Over, 1996, p.4) is that the standards and styles of reasoning used by the experiment participant may differ from those expected or explicitly stipulated by the investigators. Typically researchers intend for participants to use only reasoning based on (for instance) extensional propositional logic to guide their decision making processes, assuming that the information given in the task will be taken to provide factually isolated premises and that contextual factors and information extrinsic to the logical structure of the inference or formation of a choice preference will not impinge upon the reasoning process from premises to conclusion. Any deviation from outcomes dictated by this use of standard logic would therefore appear to the researcher to indicate a failure of rationality on the part of the participant, or provide evidence for a reasoning bias infiltrating the inferential process. Aside from practical computational difficulties, Evans and Over argue that reasoning researchers have assumed the tasks they investigate tap into participants' abilities to reason according to abstract context-free principles of logic and probability. The "interpretation problem" further indicates a potential divergence in understanding between the investigator and participant, who may interpret the premises or requirements of the reasoning task differently. Although appearing to contravene the dictates of formal logic, an individual may in fact reason perfectly logically from his own

perspective or representation of the problem at hand. He may have interpreted the premises in an unforeseen way or have perceptual experiences that influence his reasoning process in a way that is invisible to those evaluating his performance on the task.

Both of these criticisms highlight the fact that the reasoning process being tested is supposed to be devoid of content. Of interest for the experimental task is the structure of the process and its conformity with procedural principles: the specific premises used in the task are merely hypothetical examples that are intended to permit this process to be manifested in a familiar way. Rather than presenting participants with a series of logical schema involving propositions of *'p'* and *'q'*, the tasks are given content to make them more understandable. Nonetheless it is the abstract formal structure of the inferential process that is under scrutiny. The reasoning process being investigated is therefore a highly intellectualised one that is abstracted away from real-life contexts.

The "external validity" problem (ibid.) queries whether the controlled laboratory scenarios set up and evaluated by such research actually tap into anything that could be termed a psychological construct of rationality. It incorporates the concerns of both the normative system and interpretation problems by suggesting that the concept of rationality imposed by researchers is artificial and should not form a normative standard against which human reasoning ought to be judged. Traditional interpretations of the notoriously low proportion of adults passing the Wason selection task testing conditional logic have generally focused on the notion that we are imperfectly rational because of our tendency to fail to apply the principle of *modus tollens* correctly (Stein, 1996). The predominance of 'errors' has been attributed to failures to adhere to the logical normative standards to which we should aspire. Yet the 'irrationality' exhibited by many adults on failing the Wason task may in fact be explained not as a failure of rational inference but as a normal, explicable and entirely reasonable inferential process.

It would be irrational for a person to calculate every eventuality and probability before making simple everyday decisions. At times it may be rational to ignore the inconsistency of one's beliefs, for example, if it would be unwise to devote the cognitive resources demanded to such a task. Some authors (such as Foley, 1993; Nozick, 1993) have argued that being rational may be perfectly compatible with holding inconsistent beliefs (Bermúdez, 2001). Rationality needs to be tempered by practical and epistemic considerations such as the level of interest one has in resolving fallacies in one's reasoning process (Harman, 2004, p.50). It is not therefore merely dictated by the logical relations obtaining between propositions held true, and the principles of procedural rationality are not indefeasible constraints.

There are evident problems with setting adherence to logical rules as a normative standard by which to judge the correctness of a process of reasoning and one argument for this, advocated by Gilbert Harman, is that reasoning has very little to do with relations of logical consistency and implication (ibid. p.46-7). Harman argues that deduction, the epitome of logical reasoning, is not an instance of reasoning at all: *"Logic, the theory of deduction, is not...itself a theory about what to believe (or intend); it is not a theory concerning how to change your view"* (Harman, 1999, p.28). Logic and probability theory are concerned only with what is entailed by the acceptance of given premises, not with how those premises are established in the first place or with whether or not they are in fact true: *"Logical powers, in the absence of suitably grounded beliefs to provide rationally held premises, are like an engine without fuel"* (Audi, 2004, p.41). This remark brings us to the crux of the limitations of procedural principles as grounding a process-based conception of rationality. What counts as a rational process depends in large part on content of the propositions entering into that process. Rationality is essentially world-involving: in deciding what reasonably follows from a set of premises some attention must be given to the question of whether or not those premises actually obtain. More precisely, it concerns the truth-values of those

propositions comprising the premises. It might not be rational to make a procedurally correct decision or choice preference if one's starting premises were false or unreliable. The point of decision-making is to enable one to act in the world; hence any conception of rationality that overlooks what is actually going on in the world must be inadequate. Ascertaining the truth (or otherwise) of beliefs brings us to the heart of epistemology and a concern with the normative standards that constrain what our epistemic commitments ought to look like.

Reasoning is concerned with the process by which we justify, change and revise our beliefs, desires and values (Harman, 2004, p.47). This point is emphasised by the fact that while the dictates of propositional logic govern relations between propositions they do not, Bermúdez points out *"have anything to say about how one should revise one's beliefs"* (2001, p.466). The principles of procedural rationality say nothing about the world or how we ought to form beliefs and make decisions that are reasonable in light of the way the world is. Recognising the limitations of conceiving of the reasoning process in purely procedural terms brings us back to the issue I began to draw out in the previous chapter, namely that normative judgements about a person's capacity to make a decision incorporate an evaluation of his epistemic commitments as well as the formal structure of his process of reasoning.

2.3. EPISTEMIC RATIONALITY

Reasons and Context

Being engaged in deliberative decision-making requires us to draw on available sources of information and pragmatically utilise vast amounts of background data. To be able to reason one must have the capacity to accept particular premises, reject others, consider the testimony of others, appeal to one's knowledge of the way the world is and call one's prior beliefs into question in light of counterevidence. In this regard we face the classic problem of epistemology, concerning the justification of

beliefs and what we have grounds to hold true. Non-inferential beliefs may form the premises for logical inferences that lead to the acquisition of other beliefs, desires and so forth, but whether I am justified or reasonable in holding them is not a matter of my adherence to logical principles. Other beliefs may be acquired via inference from prior beliefs, desires and values, or from the testimony of others, but the relationship between these is not akin to that between premises and conclusion: reasoning does not necessarily take the form of an argument or proof[31] (Harman, 2004, p.47).

Bermúdez (2001) suggests that reasoning about evidence requires "epistemic rationality", a term he uses in contrast with procedural rationality to refer to the norms of good reasoning that dictate how we ought to treat the information we have available when forming intentions and making decisions:

> "The norms of good reasoning are principles that govern the processes of drawing conclusions; weighing up the balance of evidence for and against a particular proposition; deciding upon a particular course of action; judging the likelihood of a particular event and so forth. These are all psychological processes that result in either changes of belief or alterations in one's plans" (Bermúdez, 2001, p.465).

Epistemic rationality concerns the capacity one has to test hypotheses, revise beliefs in light of available information, decide upon one action rather than another on the basis of evidence, favour certain beliefs and values over others and make judgements about how best to achieve one's ends, whatever they may be. It is about how one revises and alters one's beliefs, generates intentions and makes plans in light of the information that is available, and the way one uses and weighs the evidence from perception, knowledge and the testimony of others in forming beliefs and intentions. This kind of rationality and its converse, irrationality, appears to be much closer to a lay usage of the term than the dictates of procedural principles:

> "We use the words 'irrational' and 'unreasonable'…for those who refuse to accept 'obvious' inductions, or for those who jump to conclusions on insufficient evidence…or for those who are uncooperative" (Harman, 1999, p.45).

[31] The roles of testimony, persuasion and counter-argument by other people and figures of authority will not be considered separately here, although they undoubtedly play a role in influencing a person's decision-making.

We would consider someone irrational if he did not ground his beliefs on sufficient evidence, typified by the jumping-to-conclusions bias that has been identified as a robust trait in individuals with schizophrenia (Garety et al., 1991). Also, and of particular interest to those concerned with the role of psychiatry as an agent of social control, an uncooperative agent might be considered irrational: here it is the agent's ends and actions that are being evaluated against society's view of how one ought to behave. These examples of irrationality reflect the idea that there is a failure in the reasoning process, but not because of a normative error of failing to conform to principles of procedural rationality. Rather, they suggest failures of epistemic rationality, namely through having mistaken or poor epistemic and evaluative commitments.

Epistemic Rationality in Psychiatry

Some psychiatric phenomena may be thought of as involving impairments to epistemic rationality. Langdon and Coltheart (2000) argue that deluded individuals suffer from an impairment in the way the evidence for different explanations of an anomalous perception is evaluated and weighed up. Spitzer similarly argues that the distinction between a delusional belief and one that is rationally held is that the person entertaining the former is unable or unwilling to reason about, justify and be open to the possibility of revising that belief in the face of counterevidence or argument[32] (1990, p.391). Again we see the appeal of understanding standards of epistemic rationality in terms of the relationship between the beliefs and experience, as the truth or plausibility of beliefs play a role in determining whether they are being rationally held.

Often delusions take the form of a highly elaborate set of beliefs and desires, each mutually consistent with one another and entirely inferentially valid, and patients may deliver procedurally intact reasoning in defence of their claims (Bermúdez, 2001, p.471; Kemp et al., 1997). Indeed, individuals with schizophrenia frequently perform better

[32] A significant philosophical and cognitive psychological literature has evolved around the question of whether delusions possess the status of beliefs (see e.g., Campbell, 2001). Insofar as their attribution serves a useful function in the explanation of action, I will assume that they can be considered to be beliefs.

than healthy controls on tasks of formal reasoning and logic (Owen et al., 2007). Nonetheless, ordinarily we expect beliefs impacting on one's decisions to have at least a partial grounding or basis in experience, whether this is sensory experience of the external world or reflective, introspective experience. A decision made on the basis of a belief, where an outside observer can grasp no justification for acquiring that belief, either from experience or through inference, may be considered irrational[33]. It is not only that the belief content is bizarre, but that often when faced with an epistemic inconsistency between the delusional belief and sources of counterevidence or the testimony of others, the delusional individual constructs elaborate and implausible reasons to immunise the delusion against all contrary evidence:

> "… the individual with the irrational or delusional belief will most likely postulate a very complicated and unfalsifiable explanation to resolve the inconsistency. Consequently, he makes the entire belief set less intelligible, less testable, more unwieldy, and thus even more immune to falsification. What seems to make this mode of thinking irrational is, in part…making it more impervious to counterargument and refutation" (Leeser & O'Donohue, 1999, p.691).

What strikes us as unusual in such cases is a failure to afford due weight and significance to the available evidence that runs counter to the delusional conviction. This is not a matter of agreeing or disagreeing with the truth values of the beliefs an individual holds, but rather a perceived deficit in the processes by which normatively significant facts about the world are grasped, accommodated and used in the formation, maintenance and revision of beliefs, and in the forming of intentions to act. Even if the delusional beliefs are the result of anomalous perceptual experience, as proposed by Maher (1999), they could not be considered to be 'rational' responses to these experiences, as the resultant beliefs do not cohere with any (or in some instances, many) other beliefs the individual holds. There therefore appears to be a breakdown between the individual's normal, general beliefs and those that appear to

[33] This characterisation excludes those beliefs that may be culturally sanctioned, such as minority religious beliefs, though I make no claim as to how such beliefs may be distinguished from pathological convictions. Such culturally sanctioned beliefs might include instances of what Wittgenstein calls "framework propositions" (Wittgenstein, 1969).

be guiding decision-making. Somewhere in the reasoning process relations between beliefs have gone awry[34].

Other authors propose that a deficit in the processes of belief evaluation does indeed account for the transition from anomalous experience to delusional belief (e.g., Davies et al., 2001). Specifically, deluded individuals often lack the ability to reject candidates for belief derived from first-person perceptions, despite the implausibility of their content and inconsistency with all other beliefs and knowledge they possess (ibid.). The supposition that there is an abnormality in evidence evaluation is supported by the startling lack of insight frequently shown by patients suffering from delusions. For instance, individuals suffering from the Capgras delusion insist that their spouses, family members or significant others have been replaced by an impostor, in the form of a clone or sophisticated robot. The delusion is impervious to counterargument, the testimony of others or even recollections of shared memories or prior knowledge. These Capgras patients do not complain that it is "as if" their relatives have been replaced by impostors: they persistently claim that they actually have been replaced (Stone & Young, 1997). Perceptual abnormality and cognitive biases cannot account for this striking lack of the reality-testing of such beliefs (Langdon & Coltheart, 2000), and this indicates that a problem in appropriating and weighing evidence from experience and other beliefs is what typifies such monothematic delusional beliefs[35].

Delusions are perhaps a special class of psychiatric phenomena that are not particularly frequently observed in normal clinical practice with such clarity and persistence as the example of Capgras delusion. However, more common conditions such as depression and *anorexia* could also involve similar impairments of epistemic

[34] It must, however, be noted that delusions may serve useful psychological functions, such as bolstering self-esteem and preventing the comprehension of perceived personal failings, particularly through an exaggeration of the self-serving bias (Bentall et al., 2001).

[35] It is less clear that we can gain a rational handle on florid, polythematic delusions. Since I am interested here in the kinds of cases that could help identify a distinction between rationality and irrationality, I will focus on the pathologies of reasoning at the borderline rather than extremes of apparent irrationality.

rationality where procedural rationality appears to be intact. Tan et al.'s (2006; 2003) in-depth interviews with anorexic patients revealed a complex picture of coherent, intact logical reasoning from premises to conclusions, but which frequently involved distorted evaluative commitments and false beliefs about weight (Viglione et al., 2006). In such cases the inference from premises to conclusion is procedurally valid but the starting premises are in breach of what I have described as epistemic rationality. There is of course much more to the problem of understanding the pathologies involved in the types of case I have mentioned, but it is plausible to suggest at this level of generality that something is going awry with patients' ability to form, maintain and act upon epistemically and evaluatively appropriate commitments. There are problems in the way that beliefs are weighed against evidence, experience and each other, with the way values are prioritised and desires are related to intentions to act.

A significant difficulty for providing any substantive account of rationality is that the norms governing good reasoning do not appear to be amenable to codification in a formal theory or even hierarchical ordering[36]. Bermúdez comments that there is a *prima facie* difficulty in comprehending how the processes of acquiring and weighing up evidence and dynamically revising one's beliefs could be subsumed under formal principles (2001, p.467). How could we derive a formal principle from the general injunction to take one's prior beliefs into account when evaluating the weight and significance accorded to new evidence, for example? The most plausible candidate for such a role in normatively constraining what one ought to believe would be the Carnapian "principle of total evidence". This principle is that one ought to take into account the totality of the available evidence when forming a judgement of probability, taken to be analogous to the conviction with which one holds a belief. Yet as a context-free principle, this axiom provides nothing in terms of normative guidance: it cannot say what counts as evidence, how one ought to weigh up the contributions of different

[36] Bermúdez (2001) suggests that proponents of Bayesian epistemology would seek to underpin all norms of reasoning with formal principles but this approach presumes that the probabilities or degrees of conviction attached to beliefs can be taken as a given.

forms of evidence, how one ought to apportion degrees of belief, and so on. Thus, it cannot dictate what one ought to do or believe, in light of the nexus of beliefs and desires one possesses. Whereas the formal demands of procedural rationality are amenable to codification in terms of principles subsumed within a formal theory of rationality, once we incorporate norms of good reasoning that constitute epistemic rationality into our conception of rationality, its normative requirements do not appear to be so easily defined and axiomatised. Codification permits an algorithmic application of standards to evaluate a particular piece of reasoning or decision-making; but if the norms of epistemic rationality are uncodifiable, it is not clear if judgements about reasoning are disciplined across the board by the same standards. The question I wish to address now is whether this lack of codification via abstract principles entails that standards of epistemic rationality might not be universal and shared across groups.

2.4. JUDGING BY A DIFFERENT STANDARD?

The Possibility of Divergent Norms

Consider two hypothetical cases of decision-making that are similar in form and process, differentiated only by the particulars of the patient's beliefs:

1. A Jehovah's witness with a life-threatening condition refuses to have an urgent blood transfusion because she believes that she will be condemned to eternal damnation if she receives blood[37].

2. A patient with a schizophrenic delusion refuses to have an urgent blood transfusion because she believes that the secret services will poison the blood.

What is it that motivates the intuition that only in the case of the schizophrenic there has been a normative mistake in the epistemic commitments the patient has, as opposed to a different but valid way of understanding the world, which is how we ought to think of the first case?[38] A fine-grained understanding of rationality must be capable of discriminating between processes considered irrational and those that signify

[37] I shall not discuss specifically the problematic issue of judgements about capacity where a religious belief is influential in decision-making, though it has been raised as a difficult area of clinical judgement (Waldfogel & Meadows, 1996).

[38] I will return to this example in chapter six, to consider how the nuanced understanding of capacity assessment I advance might better accommodate these two cases.

conformity to different but intelligible forms of reasoning, particularly if it is to shed light on the complex notion of capacity. We could appeal to the idea that the religious belief is one that is culturally sanctioned, thus providing the individual with reasonable and secure grounds for her belief, whereas the delusional belief has no external support from a community. This alone does not, however, provide a criterion for distinguishing acceptable variations in epistemic standards from unreasonable or inappropriate ones. So-called 'pro-ana' websites are home to online communities that encourage the kinds of beliefs about *anorexia* and evaluations of body image and health that would be considered wrong and unhealthy by the majority of the population and by psychiatrists, suggesting that the communal nature of particular epistemic or evaluative commitments alone ought not to ensure their acceptance as indicating divergent but acceptable standards in reasoning and decision-making.

The notion that different groups or individuals exhibit differing degrees of rationality dates back in Western thought at least as far as Aristotle, whose politically motivated conception of rationality supplied the criteria for full citizenship of the Greek state (Lloyd, 2007, p.152), notably criteria that women and slaves failed to meet. At the time, Aristotle's conception of rationality was not in a position to be challenged: it was assumed to be superior and provided a clear, fixed normative standard for reasoning ability, whether or not it was universally agreed upon. Early anthropologists conducting ethnographic studies of alien tribes made sweeping generalisations about the rationality exhibited by members of different cultures, exemplified by Lévy-Bruhl's (1923) attribution of "pre-logical" mentalities to seemingly primitive peoples. Such arguments were of course heavily criticised for invalidly imposing post-Enlightenment Western standards of reason and logic upon the behaviour and practices of a culture in which they might not necessarily apply. Furthermore, there seemed no reason to suggest that the understanding and interpretation of behaviour ought to rest on such standards applying across the board to all human cultures (Lloyd, ibid). Put simply, why

should we think it is the case that all humans conform to the same standards of reasoning?

Liberal Western democracies are at pains to avoid attributions of inferior rationality or reasoning competencies to other groups, mindful of the consequences that may follow from such perceptions of inequality. Yet it is clear that there are differences in the way people assimilate information, weigh this up with their own beliefs and values and reach decisions. It appears that what count as the norms of good reasoning may vary significantly between different populations and cultures. For example, the idea that we should place value on empirically established scientific evidence and statistics may be entirely disregarded in a society where beliefs are weighted according to the dictates of one's spiritual faith rather than medical opinion. The priority accorded to the evidence of one's own perceptual experience may be subordinate to the authority of a religious leader. In a society that values the welfare and health of a community over and above that of the individual, the value placed on one's own life may be overridden by a concern for the success and stability of the community or family. In short, the norms of epistemic rationality that I have thus far taken largely for granted do not look as though they necessarily converge in different communities. Divergences in evaluative commitments, what is taken to be evidence and the priority accorded to different forms of evidence in the decision-making process may arise across different populations. Yet whether or not they have similar institutions in place to our own, every society at least in some way distinguishes between behaviour that is deemed irrational from that which is normal and acceptable, wherever it is that these normative boundaries are drawn.

Returning to focus on psychiatry within English law and culture, the significance of the potential for such divergence becomes clear. In the 1960s R.D. Laing attacked the presumption that in the absence of a coherent, sense-making explanation from a third-person perspective, a person's actions and utterances are nonsensical or meaningless, in his critique of psychiatry and the conceptualisation of psychiatric conditions as

medical disorders. For Laing, behaviour that is considered to be symptomatic of disorder when there is no physiological evidence (such as a lesion), is *"without exception ... a special strategy that a person invents in order to live in an unliveable situation"* (Laing, 1967, p.114). In other words, behaviour that is unintelligible to an outsider has structure, purpose and meaning for the individual concerned, and it is his way of surviving or coping with his experiences. The fact that this strategy may not be discernable to others does not undermine the possibility that to the individual himself there are logical, rational connections between his beliefs and actions: from his own point of view there are reason explanations for what he thinks and does (Hunt, 1990). It is not clear whether Laing would have considered the problem to be one of a lack of epistemic access by clinicians to the beliefs and thoughts that would make the behaviour coherent, or of individuals reasoning in fundamentally different ways, but either way the implications for the judgements about agents' reasons are the same: we are not in an epistemic (or indeed moral) position to form judgements about the intelligibility of the reasons, experiences and beliefs of others whose behaviour might appear baffling to us.

The way we accommodate and weigh up information depends at least in part on the evaluative commitments and attitudes we hold, and these are largely influenced by our cultural, social and psychological context. There is therefore much diversity and variability in the way we reason. Such differences pose a problem if we are seeking to formulate general normative standards of rationality. We might freely accept that we may be unqualified to judge the reasoning process of an individual whose entire worldview, language and belief and value system is dissimilar to our own, since there would undoubtedly be a host of implicit factors to which we would be culturally insensitive. In light of this, the question arises as to whether what counts as a recognisable reason for one's decision relies on a contingent set of standards that happen to guide one's own reasoning but that need not impose any constraint on

interpretations and judgements of the behaviour of another. Davidson puts this worry succinctly:

> "If you deviate from my norms of rationality, and you do not share my sense of what is reasonable, then are you really irrational? After all, fully rational agents can differ over values. If rationality is just one more value or complex set of values, then calling someone irrational would seem to be no more than a matter of expressing disagreement with his values or norms" (1985a, p.189).

If the norms of good reasoning are contingent only on conventions or epistemic and evaluative commitments that are particular to a community, the search for a reliable, objective set of criteria for evaluating reasoning processes inevitably will be futile. On the other hand if certain epistemic standards are necessary for judgements about decision-making processes, how are they to be characterised if not in terms of universally applicable procedural principles?

Relativism About Rational Norms

I have argued that judgements about capacity cannot be disciplined by a procedurally defined set of rules operating independently of the epistemic content of the decision being made. Establishing whether a person has a recognisable reason for his decision in light (in part) of the beliefs he has necessarily involves making a judgement about whether those beliefs themselves are reasonably held. But incorporating this substantive epistemic element to judgement ostensibly generates problems for the objectivity of its standards. This lends support to a philosophically motivated concern, that whatever we happen to consider the standards of reasoning to be, we have no justification for believing that what counts as good reasoning in our own case ought to be the same for others. It may of course turn out that such agreement is necessary for one's belonging to a certain community, but the point is that it could have been otherwise, that there is a possibility of doing things differently.

I take this view to represent a relativist conception of rational norms: *"relativism is the view that cognitive goals and virtues, especially rationality, are relative to persons, situations and purposes"* (Mišcevic N., 2000, p.47). This means that the normative

standards governing the way one individual or group reasons about the world, weighs up evidence and so forth, may be indexed only to that particular group. A similar idea is found in Davidson's notion of a "conceptual scheme" (1974b), which is taken to be a system of concepts that enables one to organise or make sense of one's experience of the world. It is a point of view that may be specific to a culture or time, and the essence of the idea of such a scheme is that there is potential for other, rival schemes to exist. This possibility forms the basis of the doctrine of conceptual relativism, which I consider to be informatively analogous to relativism about rational norms[39].

Various sociological theories of knowledge (e.g., Bloor, 1983) have sought to embrace relativism about rational norms, rejecting all reference to the supposed truth or falsity of beliefs, or any objective standard of rationality regarding their formation and maintenance. Instead, by citing sociological and psychological causal conditions that lead individuals to embrace certain beliefs and forms of reasoning, knowledge is construed as consisting in whatever that particular community takes to be knowledge, and good reasoning as whatever community consensus dictates is an acceptable or correct process of inference or decision-making.

One implication of this view is that the behaviour and actions of agents could not accurately be judged by those who are not part of the same community, as the standards governing what count as acceptable reasoning and decision-making processes for the observer may differ from those by which the observed agents form their intentions. Within his own community, an agent's reasoning process may conform adequately to the norms of good reasoning that have achieved consensus in that community, but judged from the perspective and standards of an outside the process is poor, inappropriate or even unintelligible. Such a judgement would be an illegitimate imposition of standards onto the behaviour of an agent who is not bound by those

[39] Conceptual and rational relativism are not synonymous, but the close connection between one's concepts and one's reasons for action ensures parallel insights about their epistemic claims can be drawn for them both.

same norms. This kind of imposition appears to have occurred when early anthropologists studied new found tribes: their behaviour seemed strange, “pre-logical” and irrational as it was judged by the researcher’s own conception of rationality (Lloyd, 2007, p.154).

This is not merely an epistemological problem regarding how we know and can judge the rationality of others. It is an ontological concern about whether or not there are any universal standards governing the decision-making process. If there are not, the conceptual possibility is open that there may be fundamentally different ways of going on in the world. By this I mean that there may be such different ways of understanding facts, of considering evidence, and of reasoning to form decisions, that there is no possibility that an observer could form a judgement that other intentional agents outside of his own linguistic community have in some sense erred in their beliefs about the world. On this view, there are no universal normative constraints on the formation of beliefs or concepts; there is no objectively right or wrong way of doing things; people may simply be said to see the world differently and we are in no position to judge them as being in error. Indeed, we are in no position to subject their decision-making to a normative evaluation at all. This is the position most forcefully advocated by Laing in his discussions of mental illness. For him, the behaviours and delusional beliefs typical of some types of mental illness carry with them no normative consequences above and beyond the fact that we as observers or clinicians deem behaviour characteristic of mental illness to be, for instance, socially inappropriate. Normativity here goes no deeper than contingent social or cultural convention.

At this stage it is worth considering how the *prima facie* plausible relativism of rational epistemic standards I have sketched may impact on judgements of capacity, in order to understand why a lack of universally shared standards would be epistemically problematic.

Standards Governing Capacity Judgements

In light of the preference in psychiatry for minimising the idiosyncrasies of clinical judgement and increasing reliability of assessment across raters, whatever the standards by which clinician judges capacity are they ought to be universally applicable, used reliably and consistently by clinicians across the board. Codification of standards in terms of explicitly statable general principles would be one way of ensuring this kind of transparent regimentation of clinical assessment, but as I have argued what counts as fulfilling the essential criterion of using or weighing information cannot be fixed by a set of principles. Alternatively, standardisation of clinical assessment would be assured if the norms governing capacity judgements were universally shared. If standards are the same across humanity the norms of the assessor and those of the patient would coincide. Consequently these normative requirements need not necessarily be set out and articulated but instead could be implicitly relied upon to guide capacity judgements and delimit the boundaries of reason-guided decision-making processes. However, in mooting the idea that different cultural groups may possess fundamentally different norms of reasoning a doubt has been generated about the universality of rational standards.

The operational approach to capacity treats assessment as a matter of ascertaining fulfilment or failure on a number of cognitively-based descriptive criteria of mental functioning. These checklists will be perfunctory and misleading at best if what clinicians deem to count as fulfilling the criteria has the potential to differ either from other professional opinion or from the standards of reasoning by which the patient himself reaches his decision. The veneer of objectivity and neutrality afforded by the checklist approach to ascertaining capacity belies the complex normative judgement underpinning these criteria and if there is scope for the standards governing this judgement to differ, significant implications for the possibility of reliably and fairly assessing a patient's decision-making process follow. Naturally there is always scope for disagreement among clinicians but the concern here is an ontological one with

epistemological ramifications. The ontological concern is that there may simply be radically different standards of epistemic rationality governing how individuals accommodate, assimilate and utilise information, their own beliefs and values and so forth, in making a decision. Consistently convergent inter-rater judgements about particular cases will support the claim that standards among clinicians might be shared, but the concern is more pressing if we consider the potential for discrepancy in standards between a clinician and patient. A clinician will judge a patient's decision-making process according to his own implicit conception of what counts as a recognisable reason based on the available information relevant to the decision at hand. However, if the standards governing the actual process for the patient are different, the clinician may perceive a failure or pathology of reasoning when by the patient's own lights the decision is reached via a perfectly legitimate, normatively appropriate process.

Although I will go on to argue that this concern about the possible relativism of rational standards is misplaced, it is interesting to consider how the notion of judging the behaviour and actions of one group by imposing the standards of another has a particularly troublesome resonance for psychiatry. The demands of rationality could be interpreted as tools of social control to justify taking the decisions and deliberations of one authoritative group seriously whilst ignoring or undermining the voices of others on account of their apparently inferior rationality. If virtue is an invention of politicians to keep human cattle in line (Korsgaard, 1996, p.8, citing Mandeville), then so too could rationality be thought of as a paternalistic invention to quell the eccentric and bizarre, preventing decisions made on the basis of seemingly irrational reasoning from being implemented. Such hyperbole is not beyond the criticisms levelled at psychiatry particularly since the anti-psychiatry movement of the 1960s onwards. If judgements about rationality determine whose autonomy should be respected and whose is undermined, then even the hypothetical possibility of fundamentally different ways of reasoning generates a worrying asymmetry in which the authoritative standards of

rationality are those that belong to the psychiatric profession: predominantly white, middle-class males, in the UK at least. The spectre of relativism about rational standards therefore raises serious questions for the possibility of impartially judging the reasoning and decision-making process of others.

In the following chapter I seek to assuage this concern by examining the normative constraints on third-person judgements about a person's reasons and decisions. I use Davidson's arguments for the conditions of possibility on ascribing intentional states to others in order to undermine the notion that there could be radically different standards of rationality, and will go on to consider whether there are any broad necessary norms that could underpin reliable, consistent normative judgements about the decision-making process.

3. INTERPRETATION

3.1. REASONS AND THIRD-PERSON INTERPRETATION

Reconstructing Interpretation

I have argued that judgements about reasons are underpinned by standards of both procedural and epistemic rationality, and suggested that the two are not independent of one another. Despite the limitations of procedural rationality it does at least appear to supply a universal and codifiable normative standard by which to judge the rationality of a decision-making process that is abstracted from and irrespective of the content of the beliefs and desires entering into that process. By contrast, epistemic rationality is concerned with the way information and evidence is used and content of the decision within the context in which it is made. Epistemic standards are not obviously amenable to codification and I mooted the possibility that there might in principle be radical cultural differences between communities, thus generating a *prima facie* worry for the reliability and objectivity of third-person judgements about a person's reasons.

In this chapter I begin to neutralise this concern by broadening the scope of enquiry, shifting from considerations of rationality in particular instances of reasoning and decision-making to an examination of the conditions that render intentional attribution possible in the first place. To do this I draw on a theoretical reconstruction of interpretive judgement and the normative constraints underpinning the possibility of interpretation that emerge from Donald Davidson's project of Radical Interpretation (1973b). Having first sought to justify this technical approach to interpretation, I attempt to provide a robust philosophical account of the grounds of interpretive judgements to support two claims. Firstly, that the rational constraints on normative judgements about a person's reasons for a decision reflect a deep fact about our nature as intentional agents. This argument will form the basis for the second claim, to be developed in chapter five, which diagnoses the idea that judgements about the process of decision

making could be informed by radically different standards of rationality as an unintelligible notion.

The Davidsonian Project

I use the term 'interpretation' as a shorthand for the process of seeking to make sense of people's utterances and intentional actions from a third-person perspective using the language and concepts of reason explanations, that does not depend on a prior grasp of substantial theory. Interpretation can be understood as part of the process of forming judgements about a person's reasons for an action, enabling an interpreter to ascertain whether there is a rational process occurring that renders the action intelligible. Here I intend to consider how the project of interpretation can proceed on the basis of the behavioural evidence available, without any prior assumptions regarding shared language or social conventions. To do this I employ Davidson's project of Radical Interpretation (1973b), which provides a theoretical account of the inferences made by an interpreter from observing the behaviour of an agent to rendering it intelligible, by assigning meaning and intention that enable the behaviour to be amenable to reason explanation[40]. It sets out a view that an interpreter can only attribute propositional attitudes and thereby seek to explain behaviour if an assumption is made that the agent is rational, indicating a close connection between rationality and what it is to be an intentional agent. Davidson argues that the practice of interpretation is governed by constraints of Coherence and Correspondence, revealed through the methodology of Radical Interpretation (hereafter, RI), which he terms collectively 'The Principle of Charity'. The interpretive constraints imposed by Charity derive from the interpreter's own logical and epistemic resources.

The claim that Charity imposes a normative structure on the process of interpretation is contentious not least of all because in practice rationality and intentionality appear to be capable of coming apart, for example when a person's behaviour is best explained

[40] This usage is wider in scope that Davidson's own concept of interpretation, which he took to refer only to the understanding of linguistic utterances (1973b).

by attributing to him a false belief, or that his actions are self-defeating. In this regard Charity might at best be thought of as a tool that can assist interpretation on occasion but that may be abandoned if reason explanations can be better supplied by ignoring the constraints of Charity when attributing propositional attitudes: a view characterised by Richard Grandy's pragmatic 'Principle of Humanity' (1973). However, taking Charity to be a dispensable heuristic for interpretation is precisely the move that permits the threat of relativism about norms of rationality to take hold.

To mitigate against this possibility, in the latter half of the chapter I advance what Child (1996b) terms an "interpretationist" view of the relation between the methods of interpretation and the structure of the intentional realm. I argue that the applicability of the concepts of belief, reasons, and intentional action depends on their being attributable as such in interpretation for the purposes of reason explanation. This view is developed using Davidson's argument that in virtue of being an intentional agent one's behaviour is, in principle, intelligible to an interpreter. I then suggest that interpretation and intentionality are normatively structured by the demands of rationality. These conditions on intelligibility are exposed by the constraints of Radical Interpretation: the rational standards of Charity. Intelligibility and rationality cannot come apart as Grandy suggests because the applicability of the concept of belief requires the attributed attitude to bear rational relations to other intentional states. Intentional concepts therefore cannot be applied to, and behaviour therefore cannot be intelligible for, a creature whose behaviour does not conform broadly to the normative demands of Coherence and Correspondence. In the next chapter I will address the most pressing objections to this constitutively normative view of intentionality.

The locus of interest in this thesis is in instances where ordinary understanding and interpretation may fail, where it is by no means clear if a person is acting for good reasons or in some cases at all intelligibly. In exploring the process of interpretation the aim is thus to scrutinise the boundaries of intelligibility and sense-making that are

breached when communication and understanding break down. One strategy for theorising about where these boundaries might lie is to attempt to abstract away from the potential variability and norms of social and pragmatic convention that constitute our ordinary interpersonal encounters, in order to ascertain whether or not there are necessary conditions underpinning the possibility of interpreting another intentional agent. If such conditions exist, they would delineate the boundaries beyond which another's behaviour could not be understood as intentional. The Davidsonian approach to interpretation seeks to do precisely this: his radical project attempts to strip away the assumptions of shared language, social and cultural norms and conventions that ordinarily pervade our interpretive practices, to examine the kernel of intelligibility that distinguishes an organism as an intentional being capable of acting for reasons from a non-intentional system. The project of Radical Interpretation is a thought experiment designed to uncover the necessary conditions of possibility for the attribution of intentional concepts.

It is not, therefore, immediately clear that the conditions of intelligibility Davidson uncovers have any direct bearing on our ordinary understanding of reasons and decision-making. In specifying the assumptions that would enable an interpretive theory to be generated from method of Radical Interpretation, Davidson is engaged in an explicitly theoretical as opposed to empirical enterprise (Evnine, 1991, p.76). RI is not therefore a direct claim about what evidence could suffice for interpretation in practice: *"My argument that RI is possible therefore does not depend…on pretending that the evidence and methods I describe represent the actual epistemic condition…of linguists and translators in the field*" [41] (Davidson, 1994, p.124). In themselves the constraints of Charity, emerging from a rational reconstruction of interpretation in a radicalised scenario of total epistemic ignorance, do not directly generate interpretive constraints for ordinary, everyday instances of interpretation and the provision of

[41] In contrast with the likes of Chomsky (1986, p.28-9) for his theory of Universal Grammar, he is not seeking to determine the actual implicit structures underlying our linguistic competence or ability to acquire language.

reason explanations. In utilising the resources of RI I am intending to provide a philosophical account of the deep normative structure of the grounds of interpretive judgement rather than supplementing psychological theories of human reasoning or providing specific standards of judgement in ordinary practice. However, towards the end of the chapter I will consider how the insights into this structure exposed by RI can be brought to bear on the narrower project of attributing reasons in particular instances of action and decision-making.

3.2. INTERPRETATION AND THE PRINCIPLE OF CHARITY

Radical Interpretation

Radical Interpretation takes the observable behaviour of an agent to provide all evidence there can be for an interpreter to base his intentional attributions upon. It aims to generate an idealised method for understanding the linguistic utterances of a radically unfamiliar speaker without any prior knowledge of the speaker's language or his mental states. RI was originally devised to shed light on the nature of meaning and to ascertain the most abstract and general conditions on possessing a language, based on *a priori* knowledge about what must be the case for a creature to be considered a linguistic being together with publicly observable linguistic behaviour (Ludwig, 2004, p.350). Davidson's strategy here is based upon Quine's notion of "Radical Translation" (Davidson, 1970b, p.62), developed for the philosophical study of language to expose the minimal necessary conditions for languagehood without the usual contingent vagaries of culture, language or psychology[42] [43] (Quine, 1960, ch.2). The primary application of RI was thus as a tool in the philosophy of language, concerned with the understanding of linguistic utterances on the basis of extensional, behavioural criteria

[42] There are significant differences between Quinean Radical Translation and Davidson's project of RI. Whereas translation implies a mapping from one language to another, interpretation is more akin to a transition that clarifies the meanings of utterances and behaviour for the interpreter himself, by giving them expression in another way that is understandable or intelligible to him (Mulhall, 1987, p.320).

[43] For Quine, the evidence for such a theory could only take the form of behavioural dispositions to physical stimuli if the theory was to indeed be purely extensional, purged of intensional concepts and remaining 'scientific' in its approach. Davidson, by contrast, embraces the fact that the evidence for interpretation is intensional (Davidson, 2001b, Reply to Foster, p.175).

and stripped of the contingencies of social convention, heuristics and assumptions about language we take for granted in seeking to understand other people's behaviour (Davidson, 1995a). Here, however, I will look beyond this original function, instead focusing on the application of the methodology for understanding the nature of third-person attributions of intentional states based on the minimal epistemic resources available to a Radical interpreter.

With the method of Radical Interpretation, the interpreter is starting from a position of total epistemic ignorance and cannot make any assumptions about the meanings of an agent's utterances in order to facilitate intentional attributions. Neither is it possible to ground an interpretation of the agent's utterances on the intentions he has in using words the way he does, for example, by pointing and making a vocal utterance, since the interpreter has no prior knowledge of what he believes:[44]

> "beliefs and meanings conspire to account for utterances. A speaker who holds a sentence to be true on an occasion does so in part because of what he means, or would mean, by an utterance of that sentence, and in part because of what he believes. If all we have to go on is the fact of honest utterance, we cannot infer the belief without knowing the meaning, and have no chance of inferring the meaning without the belief" (Davidson, 1974a, p.142).

In the same way that Decision Theory seeks to delineate subjective probability and utility from the thin evidence of choice preferences, the challenge for an interpreter engaged in the project of RI can be expressed as the problem of ascertaining a speaker's beliefs and meanings simultaneously on the basis of the evidence of his utterances and actions. The dispositional facts that determine a speaker's hold-true attitudes, which for Davidson form the behavioural evidence for interpretation, are vectors of meaning and belief (ibid. p.142; 148). Using uninterpreted utterances as evidence alone the interpretive project cannot get off the ground, since neither can be determined prior to the other: *"interpreting an agent's intentions, his beliefs and his words are parts of a single project, no part of which can be assumed to be complete*

[44] This precludes the possibility of generating a theory of interpretation on the basis of non-linguistic intentions: an approach most thoroughly worked through by Paul Grice (e.g., 1957).

before the rest is" (Davidson, 1973b, p.127). In this idealised scenario, initiating the process of interpretation looks to be an impossible task unless some starting assumptions can be made about what the agent means by his utterances or what he believes.

Here the resources of the Radical Interpreter himself come to light. For it is on account of his status as an intentional being capable of holding beliefs, having reasons for his actions and speaking a language, that the interpreter can use his own standards of belief formation and reason-giving to enable him to make initial intentional state attributions. In spite of the relative paucity of the evidence before him, the interpreter can launch the process of interpretation by making fundamental assumptions about what the agent believes. Such assumptions are based on the fact of the shared perceptual environment in which the interpretive encounter occurs and particular presuppositions about the relationship between the meanings of the agent's utterances, the beliefs that he holds and the actions that he performs. In short, interpretation can only proceed in virtue of the assumption that the normative standards constraining the interpreter's beliefs and actions are relevantly similar to those constraining the behaviour of the agent: *"interpretation depends on reading some of the norms of the interpreter into the actions and speech of those he interprets"* (Davidson, 1994, p.123). Davidson's construal of what these norms consist of is known as the Principle of Charity.

The Principle of Charity

The term "Principle of Charity" is attributed to Wilson (1959) who developed the notion as part of a philosophical account of the determination of the semantic values of linguistic terms (Jackman, 2003), though it first came to prominence through Quine's pragmatic use of it as an arbiter between conflicting translation manuals (1960). The Principle can be formulated in several ways, each carrying different implications for understanding the theoretical reconstruction of interpretation pursued by Davidson.

Lepore and Ludwig for instance provide three different formulations derived from textual evidence in Davidson's own writings alone (Lepore & Ludwig, 2005). Nonetheless, in all its guises Charity can be thought of as a methodological presupposition for Radical Interpretation, which allows interpretation to proceed by dictating that an interpreter uses his own logical and epistemic resources to render the beliefs and utterances attributable to an agent intelligible as far as possible: *"[t]he clarity and cogency of our attributions of attitude, motive and belief are proportionate, then, to the extent to which we find others consistent and correct"* (Davidson, 1982, p.184). Its use therefore licenses assumptions about the beliefs of agents being interpreted, based on the normative structure of the interpreter's own beliefs. For its purposes in the philosophy of language this enables an interpreter to assign putative beliefs to the agent in order to establish what his utterances mean, to *"hold one factor steady while the other is studied"* (Davidson, 1975, p.167). In appealing to Charity Davidson thereby seeks to state conditions that would yield an interpretive theory for understanding an agent's language (Davidson, 1994, p.127). For my purposes here, however, I am concerned with the nature of the assumptions being made by the interpreter rather than their role in the development of a theory of meaning.

Davidson considers Charity to consist of two distinguishable constraints on interpretation: "*we must assume that a speaker is by and large consistent and correct in his beliefs"* (Davidson, 1973a, p.238). Broadly, these conditions of consistency and correctness are termed "Coherence" and "Correspondence" respectively (e.g., 1983; 1991, p.211; 1970a, p.221; 1967, p.21; 1969, p.48). The former consists in the claim that there are holistic constraints on the structure of the patterns of that speaker's thoughts and language; and the latter that a speaker's utterances indicate generally true beliefs. Although these components are interdependent, for clarity they will be considered separately here.

Coherence

In order to begin the process of attributing intentional states, an interpreter must assume that the agent's beliefs, desires, values and intentions are broadly consistent with one another and hang together as a more or less coherent whole. Hookway gives a succinct delineation of what this admittedly vague requirement amounts to for interpretation: *"We are constrained to look for…rationally coherent bodies of belief, to avoid ascribing inexplicable ignorance, to look for reasonable desires…"* (cited in Miller, 1998, p.273). It is difficult to specify how the requirement of Coherence provides a methodological injunction for the interpreter in practice, but the motivation underpinning this normative assumption is perhaps clearest in Quine's pragmatic demand for logical consistency as a constraint for the translation of truth functional operators[45]. He uses this notion extensively in Radical Translation. Although Coherence on Davidson's construal of Charity has a far broader scope and forms an essential rather than pragmatic constraint on interpretation, it is nonetheless instructive to take Quine's view as a starting point for considering the nature of Coherence, as I shall later argue that the non-semantic rule-based conception advocated by Quine is in fact the source of many of the problems Charity faces[46] [47].

For Quine, the evidence available to the field linguist takes the form of sentences assented to by a speaker, and he considers the law of non-contradiction to be a prime candidate for a rule of the translation of truth-functions. To take an example, for a field linguist the terms 'blip' and 'bloop' appear to be translatable to the English conjunction

[45] Quine does not elaborate on his reasons for restricting his use of Charity to truth-functions, although it would be consistent with his behaviourist intentions to suppose he considered them to be directly translatable through tests of behavioural assent without any semantic assumptions being necessary.

[46] Davidson insists that we need a much richer conception of Charity that applies *"across the board"* (1974a; 1974c). His formulation of Charity is therefore considerably thicker than Quine's, in that it explicitly emphasies the semanticity of interpretation rather than its non-semantic form. For the behaviourist Quine, the relation between belief and truth was ensured by the fact that the only evidence available for the translation was from stimulus meaning, which was created from a causal chain emanating from occurrent environmental events or conditions. In contrast, the behavioural evidence Davidson relies on as a platform from which to generate his account of RI is already explicitly semantic.

[47] It should be noted that Quine himself does not distinguish the elements of Charity as I have done here.

'and' and the negation 'not' respectively, and the speaker assents to certain sentences that are compatible with this translation. However, when querying '*p* blip bloop-*p*' (*p* and not-*p*) the speaker also assents, thus apparently violating the law of non-contradiction. The linguist could either argue that the speaker adheres to standards of logic that are bizarre, or alternatively could assume that he has mistranslated the terms 'blip' and 'bloop'. On the basis of the behavioural evidence alone, the interpreter cannot arbitrate between these two alternatives. The notion of logical consistency is thus invoked by Quine to constrain possible translations: it cannot be the case that a speaker defies such logical standards, and we must select a translation manual that does not entail this logical defect. This prevents translations from permitting, in Quine's words, *"silliness"* in the translations of utterances (Quine, 1960, p.59; Evnine, 1991, p.104). In every case, a translation will be preferred that, when abstracted to the form of an inductive or deductive argument leading from premises to a conclusion, adheres to the dictates of formal logical reasoning. We should err on the side of assuming we have mistranslated if a situation arises whereby a translation involves the attribution of illogical or inconsistent beliefs in order to cohere with the available behavioural evidence. Beyond this usage in the translation of truth-functional operators and logical constants (Miller, 1998, p.272) it is not clear precisely what Quine takes Charity to consist in or whether it extends beyond the requirement for translations to imply adherence to other logical truths. Irrespective, Quinean Charity can be thought of as a wholly non-semantic practical tool for translation.

Davidson's conception of Coherence differs in two ways that are important for my purposes. He acknowledges from the outset of the project of RI that the evidence for interpretation is explicitly semantic (Glock, 2003, p.174) and also applies the requirement of coherence far more broadly and applies it to the use of folk psychological concepts such as beliefs and desires[48]. This wider scope entails that the

[48] Davidson argues against the idea that perception provides us with non-propositional content (sense-data, percepts etc) that constitutes the rational basis of our beliefs about the world, in

meanings of utterances and the contents of thoughts are subject to the interpretive requirement of Coherence. That is to say, there is a normative constraint on the interpreter to attribute beliefs, desires, intentions and meanings that are coherent with one another. Coherence forms a holistic constraint on the totality of intentional attributions made by an interpreter to a particular agent. Rather than demanding deductive closure and complete logical consistency it counsels the interpreter to make attributions that ensure the beliefs, desires and intentions of the agent hang together in a broadly consistent whole. Such coherence is, for Davidson, a matter of adherence to logical and probabilistic principles derived from Decision and Probability Theory. This notion of Coherence has ostensible benefits when it comes to attempting to interpret the utterances and behaviour of others with whom we may not share certain beliefs: it looks as though it is only the logical structure of the reasoning process that Coherence concerns. However, I shall argue presently that the coherence aspect of Charity cannot be thought of as independent of the second aspect, Correspondence.

"Correspondence"

The Principle of Correspondence is derived from an assumption about the relation between a speaker's hold-true attitudes, his beliefs and his meanings, namely that he knows what he means by his sentences and that he infers the sentences expressing his beliefs are, by and large, true (Glock, ibid. p.188-9). Correspondence entails that truth conditions are assigned to a speaker's utterances that make the speaker *"right when plausibly possible"*[49] (Davidson, 1973b, p.137). I use the term "Correspondence"

---

line with Sellars and Austin (Davidson, 2003, p.695). This is because nothing non-propositional can have a logical relation with (and so provide a rational justification for) something propositional such as a belief. Davidson proposed that perceptions cause us to have true beliefs, therefore yielding knowledge directly and not via an epistemic intermediary: "I look and I believe". This contrasts with Quine's appeal to stimulus meaning serving an epistemological function. By contrast McDowell (1994) argues perceptual experience itself is conceptually structured and thus propositional, enabling it to serve as an epistemic basis for belief.

[49] Davidson does at times suggest (e.g., 1973a, p.239; 1973b, p.136; 1968, p.101) that Charity constrains interpretation by requiring that the interpreter and speaker merely *agree* about events or conditions in the environment. Lepore and Ludwig refer to this principle as "Agreement" (2005, p.190), which differs from Correspondence in not judging the truth of beliefs relative to the environment itself. Thus an interpreter and speaker could be equally deluded or deceived: they would hold the same beliefs about the environment but these beliefs could be false. This reading appears to contradict Davidson's overriding commitment to a constitutive

cautiously as it is not intended to indicate anything like the relation between thought and world implicated in traditional correspondence theories of truth (Lepore & Ludwig, 2005).

On Davidson's methodology of interpretation, Correspondence provides an injunction that the interpreter assumes a speaker's verbal dispositions to assert particular held-true attitudes in the same or similar environment largely match the interpreter's own (1991, p.211). Evnine describes Correspondence as requiring the assumption that other speakers *"find obvious what we (the interpreters) find obvious"* (Evnine, 1991, p.103). This is not an authoritarian empirical assumption about the possibilities of two agents happening to converge in their judgements and attitudes towards events and objects in a shared environment (Joseph, 2004, p.67). Rather, it points to the causal connection between the sentences a speaker holds true and events and objects in the external world: a subject's verbal dispositions are related to his environment (Davidson, 1983, p.150). The causal relation enables a radical interpreter to proceed by assuming that the attitudes a speaker holds about the world, exemplified by his utterances, are largely correct: *"if we understand a speaker, we know how her words are connected to the world"*[50] (Davidson, 1994, p.125).

Truth, on this view, is not a neutral, observer-independent arbiter of belief. Statements about the truth-conditions of sentences are always expressed in the language of the interpreter and it is thus always the interpreter's own perspective that determines the standard by which sentences uttered by the agent are judged. Davidson acknowledges that in RI we interpret a speaker *"…according, of course, to our own view of what is right"* (1973b, p.137). If we consider the project of interpretation to be akin to constructing a scientific theory, a parallel emerges between this conception of the

constraint on beliefs being largely true about the environment and I do not consider Agreement to constitute what Davidson needs the notion of Correspondence to consist in.

[50] For Davidson this requirement extends to the need for the interpretation generated to satisfy the holistic constraints of a Tarski-style Convention T, enabling T-sentences comprising expressions in the object and metalanguage to be generated that identify the extension of the truth predicate for a language (Davidson, 1967).

relationship between the observer of the data of the theory, and an influential argument in the philosophy of science put forward by Kuhn (1970b), among others. This is the view that there can be no theory-neutral vocabulary of sense-data to describe accurately the entities of and events occurring in nature, free from the perspective imposed by a human agent: *"Feyerabend and I have argued at length that no such vocabulary is available"* (Kuhn, 1970b, p.262, cited in ; Davidson, 1974b, p.191). In the process of RI the injunction for an interpreter to ascribe true beliefs to the agent is not a task that can be achieved without recourse to the interpreter's own language and perspective regarding what is true. Nevertheless, this idea will recur in chapter five as an important component of the argument against the possibility of relativism about rational standards.

Correspondence allows that the relation between an agent's utterances and the extra-linguistic reality of the world is tracked using the interpreter's epistemic vantage point and beliefs about the truth-conditions of sentences in his language (Joseph, 2004, p.63). This assumption allows the interpreter to go beyond the actual verbal behaviour of the agent in attempting to develop a theory of interpretation. The observed agent is not behaving in a vacuum: his movements and the sounds he utters occur within a social and environmental context part of which the interpreter himself will have epistemic access to. Whilst the interpreter may be unaware of the social dimension of the speaker's behaviour, in observing his actions during the interpretive encounter the two will share at least in part a perceptual environment. Thus the interpreter will form beliefs about the environment in which he is in that he can employ in making intentional attributions to others.

The Relation Between Coherence and Correspondence

The interpreter's beliefs about the shared environment necessarily inform his interpretation of a speaker's utterances and beliefs. But the attribution of true beliefs is not particularly informative for the interpreter unless he can extrapolate beyond the

directly available evidence and formulate hypotheses about what other beliefs the agent might hold, and thereby begin to enrich his understanding of the agent's language and intentional behaviour. An assumption is required that certain things follow from the attribution of a belief, and this is what Coherence warrants. Thus, if an interpreter attributes to an agent the belief 'that-*p*', this belief is not isolated from other beliefs that the agent might have: the interpreter is justified in assuming that the agent will not also believe 'not-*p*'. In this way, the interpreter can extend his interpretations beyond the available behavioural evidence, formulate hypotheses about the agent's actions and utterances, and initiate the process of seeking reason explanations for his behaviour. Both Correspondence and Coherence are necessary in order for this step to be taken.

A formative influence on Davidson's thinking about the two-factor essence of Charity was the development of Ramsey's strategy in Decision Theory for disentangling the elements of subjective utility and perceived probability that, he argued, sufficiently explained choice preference behaviour (Bermúdez, 2009). In order to extract rich information about propositional attitudes from the thin evidence of expressions of choice, Ramsey made the crucial move of assuming that the agent would act consistently to maximise the likelihood of achieving his goals and aims. An agent who made choices that aimed to maximise his expected utility was, on these terms, behaving rationally. This conception of rationality comprises two distinguishable components. Firstly, to maximise one's expected utility requires that one's actions appropriately follow from one's beliefs and desires, and secondly, if the expectation of utility is going to be fulfilled, the intended consequences must follow or be likely to follow from one's actions. The assumption that agents are broadly rational therefore formed a descriptive constraint on the interpretation of his actions (Joseph, 2004, p.51), acting as a presupposition that there would be a coherent pattern to his beliefs, desires

and actions such that he would seek to bring about a maximisation of his expected utility through his actions[51].

The idea that rational agents seek to maximise their expected utility does, at first glance, appear to align with the process-focused conception of rationality characterised as PP-rationality by Kacelnik described previously. PP-rationality is a matter of the internal coherence of a process of reasoning and the appropriateness of the action is considered solely in light of the agent's beliefs and desires. This is irrespective of the truth of the beliefs, the appropriateness of the ends given the environment in which the agent is acting and so forth. On this process-conception all that matters to the rationality of a decision-making process is the subjective weighting of beliefs and desires in the formation of an intention to act. However, it is only if one's intentions are appropriate to the environment that one would successfully maximise the likelihood of achieving one's goals. Thus the rationality of this process of decision-making cannot be considered purely as a matter of internal coherence: if they are to successfully guide action and satisfy his goals, the agent's beliefs must reflect the way the world is, that is, they must be true. The two components of Charity mirror this two-fold conception of the assumptions required for understanding the decisions of others based on their choice behaviour (Davidson, 1974a).

3.3. AN ARGUMENT FOR CHARITY

Charity or Humanity?

Several commentators (Lepore & Ludwig, 2005; Bortolotti, 2004b; Goldman, 1989; Mulhall, 1987; Grandy, 1973) have argued that the imposition of Charity rules out the possibility of explaining behaviour that we would perceive as intentional. Furthermore,

[51] Furthermore, Ramsey cited the need to take into account expressions of choice evinced by different experimental set ups, so that values and expected probabilities are scaled relative to one another: particular choices can be explained by examining other interestingly relevant choices (Bermúdez, 2009). Thus individual choices or actions cannot be considered in isolation from other beliefs, values and desires and agent holds: a holistic approach to understanding choice behaviour is necessary, constrained by the assumption that the agent will act to maximise the achievement of his goals.

they claim that if Charity is taken seriously, we would be forced to deny intentional status to agents whose behaviour, while odd, we can nonetheless apply the concepts of meaning and intentional thought to and generate reason explanations for. On this view, the requirements of Charity ought to be rejected as a constraint on interpretation. There are two strands to this rejection of Charity, both of which are discernible in Taylor's (2002) critique of Davidson's strategy for RI. Firstly, Taylor construes Davidson's Charity as epistemological, arguing that its employment in interpretation requires justification as a constraint on interpreters' attributions. Such justification could be obtained only if the application of Charity rendered fruitful explanations and predictions of intentional behaviour in every instance. He takes it that the adoption of Charity is contingent on the practical utility of the attributions it produces and could be construed as a pragmatic aid to interpretation, rather than essential to it. It will be the aim of the following section to reject this view of Charity and to argue that such criticisms misconstrue the nature of Davidson's project.

Secondly, if a Davidsonian requirement of Charity is accepted as a normative interpretive constraint, Taylor argues that it licenses ethnocentrism about interpretation and is thus inapplicable in intercultural contexts. He claims that there is nothing to suggest that epistemic virtues are universal (ibid. p.116-8), for example, the high value placed on ascribing true belief could be considered a direct consequence of the influence of Enlightenment and positivist thought in modern secular Western societies. Although Taylor is light on anthropological detail, as previously suggested it is plausible to conceive that in other cultures high epistemic value is placed on beliefs that ensure social cohesion or adherence to the dictates of religious authority, rather than on truth. It is therefore by no means a given that we are justified in thinking that the demands of Charity are universally valid, and this undermines their status as theoretical reconstructions of the grounds we have for interpretive judgement. However, I will set the groundwork here for the argument to be developed in chapter five that total unintelligibility of another culture is not an option if we want to continue regarding its

members as intentional agents (Davidson, 1974b). Whilst there may be differences in epistemic values and standards between different cultures, I will go on to claim that the normative demands of Charity are not ethnocentric but rather capture an important insight about the universal constraints and limits of intentionality.

There are numerous routes into arguing that Charity ought to be treated as a contingent epistemological strategy, all based on the overarching claim that generating reason explanations for behaviour is not beholden to Charity's restrictive demands. Indeed, behaviour is often most intelligible when attributions are made that explicitly violate the constraints of coherence and true belief. I will address the most compelling stance against Charity, involving the citation of clear-cut empirical instances of irrationality, in the following chapter, but here I consider the *prima facie* plausible view that the constraints of Charity are pragmatic and thus potentially dispensable for the project of interpretation.

The most well-known characterisation of this idea comes from Grandy (1973). Primarily attacking Quine's notion of Charity, he argues that if the purpose of translation is to produce the best possible predictions and explanations of an agent's behaviour (ibid. p.442), then Charity is neither necessary nor even particularly well-suited to the task. His alternative Principle of Humanity provides a pragmatic injunction that when translating another's utterances *"the imputed pattern of relations among beliefs, desires and the world be as similar to our own as possible"* (ibid. p.443). Grandy does not reject the dependence of intentional attributions on the epistemic and logical resources of the interpreter, but rather suggests that these permit of more flexibility than the requirement to attribute true and coherent beliefs entailed by Charity. Breaking the connection to truth that is so central to Davidson's account, Humanity straightforwardly allows scope for the attribution of false beliefs if in doing so an agent's behaviour is thereby rendered more intelligible. This permits, for example, an interpreter to attribute a belief to an agent that he (the interpreter) knows is false, but that which from the

perspective of the agent is either understandable (perhaps, for instance, he cannot see what is obvious to the interpreter on account of his spatial location) or indicative of the agent making an error in that instance. Preserving empathy with the agent's point of view is more important to interpretation than preserving truth in the beliefs attributed (Stein, 1996, p.120). Humanity counsels that when an interpreter is engaged in interpretation, he ought to recognise that the agent has basic perceptual and epistemic similarities to himself (Grandy, p.445), and this includes all the flaws of human reasoning, propensities to evidential bias and proneness to committing logical fallacies that typify our decision-making and intentional behaviour[52].

In arguing that constraints on intentional attributions are those of ensuring agreement, Humanity places a significant emphasis on the capability of the interpreter to empathise successfully with the interpreted agent. With the link to truth broken, the normative constraints on the interpreter are, Grandy claims, far looser than those of Charity, and the only resource the interpreter has to frame his interpretation is an extrapolation of his own consideration as to what he would believe or do in the situation of the agent. The Principle of Humanity therefore ties the intelligibility of behaviour solely to a third-person perspective that floats free of considerations of rationality.

Humanity looks to be most plausible when considering cases of understandable error. It is of course true that in the ordinary practice of interpretation, generating a reason explanation that makes sense of an agent's intentional behaviour might involve the attribution of false beliefs, particularly if the behaviour does not come about as the result of a carefully deliberated process of reasoning and evidence-checking by the agent. But the attack on Charity's concern with true belief as too stringent a constraint on interpretation neglects the balancing role played by Coherence in the attribution of intentional states. Charity does not dictate that an interpreter ought to attribute only true

[52] There are strong parallels between the Principle of Humanity and simulation theory, although Grandy takes Humanity to be an option for interpretation rather than a psychological theory of interpersonal understanding.

beliefs if doing so would do violence to the coherence of the agent's beliefs. For example, suppose I am in a bar and order a gin and tonic. The barman nods, turns to face the row of bottles behind him, picks up a bottle of vodka and begins to pour out a measure. According to Grandy, Charity demands that I ought to attribute to the barman the belief that he is pouring a measure of vodka, as this would be a true belief. However, to do so would be to ignore the inconsistency in his beliefs that would arise from such an attribution. Presumably, being a competent barman he knows that a gin and tonic requires gin, not vodka, and if he heard my order correctly he will believe I wanted gin and tonic and that he is fulfilling my request. To attribute to him the true belief that he is pouring vodka would undermine the coherence of his other beliefs. In this situation it is straightforward to attribute to him a false belief, that he is pouring gin, and thus to maintain the coherence of his beliefs and actions.

This move is a legitimate one to make within the constraints of Charity. If a false belief is consistent with what an agent would know given his evidence and the rest of his beliefs, Charity has no difficulty with permitting such falsity to be attributed. Correspondence does not trump Coherence: both are necessary for successful interpretation. In this respect, Grandy's argument for Humanity misconstrues what Charity requires, and his alternative principle does not provide any further resources for interpretation that Charity does not already utilise.

The rejection of truth as a constraint on interpretation does, however, differentiate Humanity from Charity. Humanity trades on agreements in judgements between the interpreter and interpretee, that is, the interpreter makes attributions according to what he himself would believe in a similar epistemic situation. Explicitly in contrast to Charity, Humanity does not take truth to provide a standard of interpersonal agreement. As an interpreter one could assume that if an agent points and utters the word 'cat' he believes there is a cat in the direction he is pointing, and if he utters 'dog' he believes there is a dog, just as the interpreter himself would do. But without reference to what

the interpreter takes to be true, given the epistemic surround in which the interpreter and agent are operating, how could these intentional attributions be made? Agreements in judgements depend upon a degree of co-ordination between interpreter and agent about what is true:

> "it is only when an observer consciously correlates the responses of another creature with objects and events of the observer's world that there is any basis for saying the creature is responding to those objects or events rather than any other objects or events" (Davidson, 1991, p.212).

In other words, agreement between interpreter and agent needs to be triangulated with the world if it is to tell the interpreter what the agent does in fact believe (Davidson, 1991, p.213). The assumption of truth is required to tie beliefs to specific events and objects in the world and without it, the notion of sameness in judgements and agreement between interpreter and agent is an empty one. I suggest then that Humanity's appeal to agreement in judgements implicitly relies upon the very constraint of Charity that it is seeking to reject, namely the requirement of truth. Although it assumes that the application of intentional concepts is independent of the interpretive constraints of Charity, the idea that interpreters ought to attribute agreement in beliefs is only given substance if it is co-ordinated by a connection to truth.

The argument from Humanity is pitched solely at an epistemological level, in that it focuses on the application of interpretive constraints by an interpreter without consideration of the metaphysical suppositions underpinning the use of these concepts. Humanity looks plausible if the concepts of belief, reasons and intentional action employed by the account do not themselves implicitly rely on the norms of rationality captured by Charity. However, it is my contention that the concepts of intentionality are intrinsically structured by these norms, and that therefore the pragmatic use of Humanity does not in fact succeed in eliminating the reliance of interpretation on the demands of Charity.

This kernel of an argument at first appears to conflate two distinct philosophical projects. On the one hand, we are concerned with the question of whether there are

essential normative constraints that are imposed on the interpreter, which is an epistemological issue regarding the acquisition of knowledge and understanding of intentional behaviour. On the other hand, I have speculated that Humanity, the pragmatic alternative to Charity, implicitly relies on a concept of belief that is bound up with the requirement of truth: an insight that concerns the metaphysics of intentional states. Indeed, the argument for Humanity, perhaps reasonably, takes it as given that the epistemological project of generating a theory of interpretation is divorced from an ontological thesis about the nature of intentional states and meaning. However, on Davidson's view of the relation between language, thought and the world, this distinction is not warranted. The epistemological projects of seeking reason explanations and justifying the constraints of Charity on interpretation are intrinsically bound up with the nature of intentional thought, action and meaning: a view termed 'Interpretationism' by one of its key proponents, Child (1996a).

Interpretationism and the Nature of the Intentional Realm

Interpretationism is a methodological approach to both the metaphysics of intentionality and the epistemology of interpretation that is based on the claim that the interpretive structure of intentional state attributions and reason explanations is constitutively bound up with the nature of the intentional realm. Its central claim is that the conditions on interpretation are not merely contingent features of the interpretive project that are intended to aid the process of intentional attribution, but rather that the structure of interpretation mirrors the structure of the domain of intentionality: *"When we attribute a belief, a desire, a goal, an intention or a meaning to an agent, we necessarily operate within a system of concepts in part determined by the structure of beliefs and desires"* (Davidson, 1973a, p.230). Its advocates (primarily Davidson, passim; Child, 1996a; 1996b; 1993) assert that we can seek to understand the nature of beliefs, desires, intentions and so forth by examining and interrogating how the process of interpretation functions, on the basis of agents' intentional behaviour and utterances. Understood in this way, interpretationism does not necessarily imply any particular thesis about how

the interpretive and intentional realm is structured. Davidson and Child do, however, make such a commitment, claiming that interpretation and intentionality are structured by the normative demands of rationality. I will term this thesis "rational interpretationism", to distinguish it from a weaker view that would not necessarily afford a constitutive role to rationality. It can be captured by the claim that rationality is intrinsic to the concept of belief, and this impacts both on the conditions under which intentional concepts can be applied and the constraints that underpin the possibility of interpretation. For the rational interpretationist, these conditions of rationality are described by the Principle of Charity, namely the requirements of Correspondence and Coherence.

If it is assumed that the demands of Charity are interpretive constraints imposed after the fact of having decided whether a piece of behaviour is intentional in the first place, then it is little surprise that Charity appears at best to be a potential aid to intentional attribution but is by no means constitutive of the interpretive process. Rational interpretationism, by contrast, seeks to shed light on the mechanisms of intentional thought by taking the epistemology of interpretation, structured by the norms of rationality, to reveal something about the metaphysics of mentality. By tying the rational epistemology of interpretation to the structure of the intentional realm, the objection that the constraints of Charity supply only pragmatic aids to interpretation is undermined. For, on the broadly Davidsonian view I am developing here, the normative constraints revealed by RI have their basis in real features of the mental realm, and are not merely addenda to the process of intentional state attribution. The very possibility of behaviour and beliefs being capable of falling under an intentional description depends upon their being interpretable as such by an observer[53]: *"what attitudes a subject has is a matter of how she can be interpreted, which is answerable to what she says and does"* (Child,

[53] Davidson does not wish to deny that we have privileged access to our first-person intentional states, but rather claims that this knowledge is not demarcated by the kinds of privilege often assumed of first-person authority in contrast to third-person access, such as its being infallible or non-inferentially acquired.

1996b, p.9). This claim, however, appears only to push the objection one step back, and the Davidsonian approach faces a further, more complex challenge: what justifies this bold assertion that the epistemology of interpretation mirrors the nature of the intentional domain? Moreover, why could it not be the case that the rational interpretationist is making a further illegitimate step here: what grounds the assumption that the theoretical reconstruction of intentional attribution described by Radical Interpretation tells us anything at all about the way the intentional domain is structured?

To answer this question the rational interpretationist appeals to a second claim, one that is considerably more contested but that, I suggest, acquits the thesis of the charge that the constraints on interpretation revealed by RI are dispensable. The view I propose will be set out here and unpacked and defended throughout the next two chapters. My central claim consists of a positive and a negative thesis. The negative thesis is that it is a mistake to conceive of intentional states as private mental entities with determinate content and to think that the process of interpretation is an attempt correctly and accurately to formulate hypotheses about these from an impoverished position of third-person epistemic access. Whilst I will not posit specific arguments either outlining this view, which is a remnant of Cartesianism about the mind, or the common objections to it, I aim to show that there are compelling reasons to prefer an alternative conception of the relation between mind, world and other people. The positive thesis, which I shall focus on in the spirit of a Wittgensteinian therapeutic exercise, comprises two claims: that the norms governing the application of intentional concepts are indeed those of truth and coherence, but secondly that these norms ought not to be construed as abstract codified principles that impose strong prescriptions on what we ought to do and believe.

Language and Thought

The interpretationist view of the metaphysics of intentionality I am advocating finds strong resonance with a similar and more established view about the nature of

linguistic meaning. The claim that meaning, like belief, is essentially public and intersubjective has its roots in a rich historical and philosophical tradition the most influential formulation of which is in Wittgenstein's argument against the possibility of a private language (1953, §243 onwards). With this in mind it is worth framing the constitutive claim about the nature of intentional states within the context of Davidson's particular understanding about the relation between thought, the world and language.

Davidson's central aim in the philosophy of language was to devise a method of generating a truth-theory for natural languages, in a way that was achievable for formal languages: a project rendered problematic by the fact that natural languages are grammatically complex and contain intensional contexts that do not straightforwardly admit of axioms that specify the truth-conditions of sentences (Joseph, 2004). Nonetheless, using the resources of a Tarksi-style theory, Davidson sought to provide an account of how the meanings of linguistic utterances could be determined. As we have seen from the methodology of RI, the evidence Davidson appealed to took the form of sentences held true by the linguistic agent under interpretation, evidence that necessarily implicates facts about what he believes to be true. However, in contrast to those who consider a theory of meaning to be either dependent on a prior understanding of speakers' beliefs and intentions (such as Austin and Grice) or on some mental object, (such as that picked out by a Fodorian language of thought), Davidson refuses to prioritise the philosophy of mind over language. Whilst we cannot appeal to an atomistic notion of meaning, neither can we *"hope to use thought to explain linguistic meaning"* (Evnine, 1991, p.74). Davidson does not wish to divorce his theory of meaning from facts about propositional attitudes and intentions, nor, by implication, from the social context which he acknowledges is essential to language. Nonetheless, at the same time he rejects the idea, advocated by Dummett, that language is analytically prior to mind and in this respect strays from the formal semantics approach of Carnap and Tarski (ibid.). The mutual interdependence between thought and language is not a relation at the level of epistemology but is

rather constitutive of belief and meaning themselves: *"belief and meaning are intertwined in such a way that there are not two elements there to be separated. This is an ontological point, not merely an epistemological one"* (Føllesdal, 1984, p.308).

It is a fundamental feature of Davidson's approach to thought and language that the meanings of sentences and speakers' intentions are constitutively interdependent (1975). Meaning is connected with truth conditions and is thus extensional, but is also indeterminate and must be construed in the context of the speaker's thoughts, intentions and actions, and vice versa[54]. For Davidson, the project of interpretation is *"epistemology seen in the mirror of meaning"* (Davidson, 1975, p.169) and therefore occupies a central role in our understanding both of the world and of other people. He is concerned with understanding belief and other propositional attitudes in terms of their relations to truth, in the light of meaning itself. This indicates that meaning and belief are thoroughly interpretive and neither can supply foundational conditions for understanding the other, as they are established only through an attempt by observer and agent to engage in a mutual process of understanding.

I do not consider that the inherent circularity in this strategy of argument to be a weakness, however. As will become clear, seeking a foundational bedrock for claims about belief and meaning is misguided: I am not attempting to set out necessary and sufficient conditions for the interpretation of another intentional agent but rather to develop a view of the relation between thought, language and the world and the nature of intersubjective understanding that stands or falls together but is, as I shall argue, no worse off for its lack of observer-neutral evidence or criteria of justification[55].

[54] Thus neither intention nor a reference-based relation can play a foundational role in constituting meaning (Malpas, p.68): an idea that will have resonance with the discussion of rule-following to follow in chapter five.

[55] This approach bears similarities to the hermeneutic philosophy of Gadamer, who regarded understanding as grasping the inter-relatedness of things within a common horizon, rather than acquiring an unshakeable foundation for knowledge (Malpas, 1992).

3.4. THE INTENTIONAL REALM

Rationality, Intelligibility, Intentionality

The rational interpretationist view contends there is a strong implicit connection between the nature of intentional states and the explanatory function they serve for interpretation. The intentionality of behaviour is a matter of its interpretability from a third-person perspective: it is the interpreter's judgements that dictate the conditions under which intentional concepts can be applied and the reason explanations that can be given for the agent's behaviour. Hence the role of the interpreter comes to the fore in a somewhat counter-intuitive way, as the implication of this thesis is that one's own beliefs and the meanings of one's utterances owe their very identity to the possibility of their figuring appropriately in a reason explanation for one's behaviour given by an interpreter: *"certain types of thoughts necessarily depend for their individuation on an individual's relation to others"* (Davidson, 2003, quoting Tyler Burge, p.698).

The thesis presented thus far is compatible with the idea that whilst intentional concepts depend for their application on a third-person perspective, this requirement does not exhaust the extension of such concepts. But the Davidsonian approach is predicated upon the claim that the facts available to and the constraints exposed by the project of Radical Interpretation are all that there is to interpretation and intentional agency[56] (Davidson, 1983). The intentional realm cannot be based on anything beyond the resources of Radical Interpretation: it is interpretively closed, despite being causally open to the non-intentional realm (Malpas, 1992, p.105), and only through the resources of interpretation can the concepts of language, thought and understanding find application (Davidson, 1994, p.126). Utterances and movements that are not interpretable as intentional behaviour are not, on this account, intentional at all. Hence there are no facts about an agent's beliefs, desires and intentions that are beyond the epistemic capability of an informed interpreter to know. All there is to propositional

[56] This claim carries the implication that intentional content is intrinsically indeterminate, since there is no evidence beyond what is available to an interpreter by which to arbitrate between potential competing interpretations of intentional behaviour.

attitudes so described is the kind of information that is available to an interpreter, namely the actions and utterances of an agent, the conditions or circumstances under which he exhibits these behaviours, and any prior knowledge the interpreter has about the agent's history and past behaviour.

One forceful criticism of this interpretationist view of mentality is that it appears to be anti-realist about mental concepts. The notion that the evidence available to an interpreter exhausts the conditions under which intentional concepts can be applied looks, at first, to be wildly implausible. Surely however they are described in folk psychological terms, intentional states are physiological states in the brain, whether these are conceived as atomistic representational mental entities or distributed patterns of neural firing? Whilst I am not denying the validity of the basic categories and concepts of the natural sciences, it remains the case that the concepts of intentionality and the explanations of behaviour that issue from them are not capable of re-description in physical-causal terms[57]. If we take the interpretationist view of the mental realm seriously, then what counts as an intentional state, intentional action or linguistic utterance is contingent on its being interpretable, in folk psychological vocabulary, as occupying a role in a reason explanation:

> "What qualifies a movement as an action is that it is explained by a reason explanation rather than by a purely causal explanation. Similarly, to be a desire or a belief is just to be a factor that can figure appropriately in reason explanation" (Føllesdal, 1984, p.312).

Let us consider a piece of intentional behaviour, for instance, grasping a cup and raising it to one's lips. This movement could be described in terms of the muscle contractions involved, the neural firing pattern that co-ordinates the motor activity, or in any number of ways using the physical-causal vocabulary of the natural sciences. What makes this motion the intentional action that it is, i.e., drinking, are the reasons that figure in an explanation of the motion, and these fall under a different vocabulary and set of concepts, the distinguishing features of which I outline below. The

[57] The qualitative distinction here echoes Sellars' distinction between the realm of law and the space of reasons (Rosenberg, 2008).

mechanistic language of cause and effect characterised by explanations in the natural sciences is simply not suited to the provision of reason explanations, and any attempt at reduction will entail the loss of the explanatory purpose these reason relations are supposed to serve: *"the capacity of one belief to explain another depends on relations that cannot be characterised except intentionally"* (McDowell, 1984a, p.394).

Before seeking to assess whether and how this conception of the relationship between intentionality and interpretation can be defended, two further characteristic features of this rational interpretationist view must be drawn out. These features have already been implicit in much of the discussion thus far, and here I aim to bring the former in particular to greater prominence than Davidson himself does, as I will go on to argue that as an intrinsic feature of the intentional realm it serves to temper the ostensibly stringent constraints imposed on interpretation by the demands of truth and coherence.

Constitutive Holism

Holism is the view that elements or items within a system are individuated only through the relations they bear to other items within that system. As an interpretationist thesis about the nature of the intentional realm it can be taken to comprise both an epistemological and metaphysical claim that the conditions of individuation of intentional states depend on their location within a whole network of other intentional states and attitudes. If we acknowledge that all propositional attitudes are part of an interconnected attitudinal system (Malpas, 1992, p.73), intentional attribution and the seeking of reason explanations for a particular instance of behaviour require an interpreter to take into account and attribute a host of related intentional states, utterances and behaviours. At this level, however, the holistic approach appears trivial, and is certainly not unique to theorising about the psychological realm. Holism is a feature of theory construction that is common throughout the natural and social sciences (ibid. p.54), and in theorising about language. Indeed, *"Holism merely in respect of how one might…arrive at a theory of meaning for a language is…almost*

banal" (Dummett, 1993, pp.25-6). Few would wish to challenge the claim that when we attempt to understand the behaviour and utterances of another person we should attribute intentional states on the basis of as much of the behavioural evidence before us as possible. However, holism about the intentional realm is not merely a claim about the constraints on evidence collecting in interpretation (Malpas, 1992, p.54): within an interpretationist system it is the metaphysical claim that it is the very nature of intentional states that they are constituted by the relations they bear to one another.

This view, which I shall term "constitutive holism", can be understood as a kind of functionalism that extends across the intentional realm and into the intersubjective world, one which dissolves the Cartesian divide between 'private' inner mentality and the social and physical context in which an agent is situated. Again, we can gain insight into the conception of intentionality advocated here by considering a related view about meaning. The interpretation of linguistic behaviour is a matter of understanding the way in which a speaker uses the words that he utters: *"we interpret a bit of linguistic behaviour when we say what a speaker's words mean on an occasion of use"* (Davidson, 1973b, p.141), and when we know what roles the constituent parts of an utterance play in an agent's language, we know what that utterance means. Thus, it is plausible to argue that what constitutes the meaning of a term is the functional role it occupies in a language.

If we consider the meaning of a word to be constituted by its use, not only by an individual but in the context of a language shared within a community, the functional determinants of meaning are distributed across the whole network of patterns of use and intentional behaviour; hence meanings have holistic identity conditions. Davidson explicitly advocates a brand of semantic holism, largely derived from Quine, that stands in opposition to atomistic, correspondence or reference-based theories of meaning that

have been particularly influential since Locke[58]. In line with the communication-intention theorists such as Austin and Grice (e.g., Grice, 1957), who sought to characterise meaning in terms of a combination of beliefs and intentions, Davidson takes sentences (as opposed to individual words) to be the primary bearers of linguistic meaning (Evnine, 1991, p.74). In contrast to theories that seek to ascertain the meanings of words as discrete entities, there is no *"single nugget"* (Davidson, 2003, p.699) which is the thing that is meant by a word, phrase or sentence: *"the meaning of a sentence, the content of a belief or desire, is not an item that can be attached to it in isolation from its fellows"* [59] (Davidson, 1982, p.183). With this approach, Davidson follows Frege in the assertion that words can only have meaning in the context of the sentences in which they are used, stating that meaning can be imputed to semantic components[60] of sentences *"only as an abstraction from the totality of sentences in which it features"* (1967, p.22).

The functional conception of language, articulated most succinctly by the aphorism "meaning is use"[61] can be applied *mutatis mutandis* to belief. In order to attribute a belief with a given propositional content to an agent, the role that it plays in agent's psychology and the relations to other propositional attitudes need to be taken into account. The attribution of a single belief rests on the supposition of many more (Davidson, 1982, p.183). This holistic requirement is particularly clear if we consider a case in which an individual sincerely asserts something bizarre such as 'there is an elephant in the desk drawer'. Here an interpreter would be justified in querying what the content of this belief is. If an agent holds beliefs about desk drawers (including, for

[58] He took the radical step of asserting that reference is nothing but a semantic abstraction that can play no explanatory role in a theory of meaning. The argument against reference is founded upon the rejection of the reification of meanings construed as entities (Davidson, 1977, p.215).

[59] This approach to meaning is broadly similar to that taken by Structuralist theorists of language such as de Saussure, who construed language as a system of relationships and differences (Malpas, p.55).

[60] In a defence of Davidson's position Ramberg asserts that the essential compositionality of natural languages is entirely compatible with denying the need for a wholly 'bottom-up' approach to meaning (1989, p.35).

[61] This phrase is often attributed to Wittgenstein but it does not do justice to the complex understanding of meaning that Wittgenstein develops throughout the *Philosophical Investigations* (McGinn, 1997).

instance, the fact that they are comparatively small) and about elephants (substantial, large creatures occupying significantly more space than a desk drawer) it is inconceivable that an agent could genuinely form such a logically inconsistent belief: *"since the possession of propositional attitudes involves the possession of relevant concepts, it implicates abilities to exploit these concepts in ways that respect their logical roles"* (Millar, 2004, p.7). In this instance the putative belief fails to respect the logical relations between the concepts of 'elephant' and 'desk drawer' and an interpreter inevitably would be baffled by a sincere assertion of this incoherent belief[62].

That thoughts are broadly consistent and integrated, and sustain relations of implication, confirmation, justification and reinforcement with one another, is an expression of the psychological unity characterised by the notion of holism (Malpas, 1992, p.74). Intentional states are *"individuated and identified by their relations to other beliefs"* [63] [64] (Davidson, 1997, p.124), but this holism extends throughout an agent's psychology and his interactions with the physical and social world. It is important not to underestimate the wide horizon over which this holism operates, as it undermines the idea that beliefs can be understood and attributed in isolation: *"when we first begin to* believe *anything, what we believe is not a single proposition, it is a whole system of propositions. (Light dawns gradually over the whole)"* (Wittgenstein, 1969, §141). As children, we acquire a whole aggregation of beliefs, rather than a system consisting in a series of single beliefs.

[62] An analogous theory in the philosophy of language is Conceptual Role Semantics (CRS), which is a sophisticated extension of the idea that a term's meaning is identified by its use in social interaction and communication (see e.g., Fodor & Lepore, 1992; Block, 1987 for a detailed critique).

[63] I am not here claiming that the entirety of the psychological realm is propositional in nature. Numerous attitudes, such as pain, are arguably non-propositional and are not therefore necessarily subject to the relational constraints I have given for the propositional attitudes (Malpas, 1992, p.88). However, as the main focus of my discussion is on attitudes that are relevant to reason explanations and interpretation, I shall not consider the impact of non-propositional content on this constitutive thesis of holism.

[64] There is a similarity here with Parallel Distributed Processing, or connectionist models of cognitive architecture, in which the identity of interconnected nodes is entirely dependent on their location within a network space of multiple other nodes (e.g., McLeod et al., 2006).

It is not simply the case that in interpretation it does not make sense to ascribe a singular belief or desire without a holistic attribution of other beliefs, intentions, preferences and so on to an agent, but rather that without such relations the very concept of belief would lose its application. We see, then, that holism shifts the locus of interpretation away from thinking of the identity conditions of an individual's intentional states as a matter solely of their internal relations and connectedness, and into an encounter between the individual and his interlocutor or community (Pettit & McDowell, 1986, p.14). Thus even though an interpretive encounter is an individual, circumscribed instance of interpretation, this does not isolate the interpretation from the world or the community within which it occurs (Malpas, 1992, p.100).

If thoughts are constituted by their holistic relations to one another and by the role they occupy in interpretations of an agent's behaviour, how can they be said to be objectively real? Davidson's response seeks to undermine the demands of realism when discussing his semantic holism:

> "It is only if we have a Cartesian, individualistic conception of meaning and the intentional that we assume a conflict between realism and holism. Realism about correct interpretation does not, for me, entail that what someone means by his words is independent of what is understood by others" (Davidson, 1994, p.126).

Given the close inter-relation between beliefs and meanings in Davidson's theory, this point extends to realism about propositional attitudes. On an interpretationist view the intentional agent is not a solipsistic entity possessing internal intentional states, speaking a language and performing actions that fall under the concept of intentional behaviours, bearing only a contingent relation to the interpretations of others. Yet just as the lack of a constitutive relation of reference does not vitiate the 'reality' of a term's meaning something, so too can propositional attitudes be said to be 'real' even though they are not identified with a physical brain state or atomistic representational state.

Normativity

The constraints of Charity, those of Coherence and Correspondence, are norms that on Davidson's view must govern a radical interpreter's intentional attributions if he is to find the agent intelligible as an intentional being at all. Thus the radical interpreter's theory ought to be disciplined by the requirements that the beliefs attributed to the agent are broadly coherent and largely true about the world: *"each interpretation and attribution of attitude is a move within a holistic theory, a theory necessarily governed by concern for consistency and general coherence with the truth"* (Davidson, 1974a, p.154). In Radical Interpretation then, the normative content of one's theory is that agents are interpreted as holding the beliefs that they ought to hold, according to the interpreter's own beliefs about what is true. For instance, an agent would be ascribed the belief that 'snow is white' if that is what he ought to believe, by the interpreter's lights. In this regard the demands of Charity look to supply a set of requirements for interpretation: if the interpreter is to attribute intentional content to observable behaviour, he imposes a structure of broad coherence and truth on his attributions.

Although Davidson's project is a theoretical abstraction of the conditions of possibility for interpretation in a radical scenario the constraints of Charity have frequently been taken to be directly applicable to ordinary interpretation, supplying an epistemological framework through which to build interpretation (Bortolotti, 2004a; Ludwig, 2004; Henderson, 1991). I will argue in the following chapter that this move of generating specific conditions on interpretation from Davidson's idealised methodology leads us to misconstrue the nature of the normative constraints on interpretation. However, as it has been a widespread presumption that Charity is applicable in this way, for the purposes of exegesis I shall assume this view is a tenable one, and term the normative requirements on interpretation derived from Charity the 'Rationality Constraint' (RC).

The Rationality Constraint is imposed on interpretation on the basis of a further assumption about the nature of intentionality that is central to the rational

interpretationist view. I shall term this the 'Rationality Assumption' (RA): a metaphysical thesis that intentionality is constitutively rational. That is to say, it is of the very nature of beliefs and other intentional states that they are rationally related to one another, to linguistic utterances and to intentional behaviour. In this respect the normative requirements of Charity that constrain Radical Interpretation are also descriptive requirements on what it is to be an intentional agent (Ludwig, 2004, p.346).

The assumption of Charity in RI (which supplies conditions for the epistemological Rationality Constraint) is intended to reflect the constitutive rationality and holism of the mental (the Rationality Assumption), by claiming that what it is to be an agent is to exhibit rationality in one's thoughts: *"…all thinking creatures subscribe to basic standards or norms of rationality…it is a condition of having thoughts, judgments, and intentions that the basic standards of rationality have application"* (Davidson, 1985a, p.195). Coherence and Correspondence therefore specify the kinds of relations that must obtain, not only for interpretation to be possible, but for the very identity of the propositional attitudes under scrutiny, and the close relations between propositional thought, language and the world entail that *"for someone to be a speaker he must be interpretable, and to be interpretable, he must be mostly right about his environment"* (LePore & Ludwig, 2006, p.16). Whilst we cannot 'mainline' on truth, fixing belief content through a reference relation to objects and entities out there in the world, the notion that it is a condition of being an intentional agent that one's beliefs are largely true nonetheless retains traction on the world. Hence, whilst Charity is the basis for a theory of interpretation its requirements escalate into conditions of interpretability for intentional agents. This is because the concepts of meaning and thought are applicable only in virtue of the conditions under which they can be attributed by an interpreter.

The normative constraints operating on interpretation thereby also apply to thoughts, meanings and intentional behaviour themselves, ensuring that the standards that apply to third-person interpretation mirror those that form constraints on the possession of

intentional, action-guiding states from the first-person perspective. Appealing to the beliefs held by interpreter's own lights as a way of initiating the project of interpretation is therefore not an unjustified imposition of one's own epistemic standards onto the behaviour of another, as the norms by which an interpreter makes intentional attributions are just those by which he himself is guided in forming beliefs, making inferences and acting intentionally.

One of the key criticisms levelled against defenders of the rational interpretationist conception of intentionality is that it is at best unjustified, and at worst, plainly false. In particular, it makes irrationality in a system impossible to accommodate: if propositional attitudes are constitutively bound by rationality, occurrences of irrationality ought logically to be impossible. Intuitively, it must be conceded that irrationality is possible and indeed that irrational behaviour can nonetheless be interpretable, but this entails making a concession that threatens to sever the connection between rationality and the intelligibility of behaviour. Furthermore, if behaviour can be intentionally characterised despite violating the RC, which is supposed to underpin standards of rationality, this strikes a *prima facie* blow to the claim that its standards are shared across intentional agents. This is a particularly pressing concern given the myriad evidence that we are frequently irrational and inconsistent in our thoughts. The problems this kind of scenario poses for interpretation were evident in chapter two where I discussed the apparent logicality and coherence of some delusional belief systems, in which the agent's grip on reality was questionable and generated problems for the possibility of interpretation and the comprehension of utterances and behaviour. In the next chapter I consider whether and how instances of inconsistency can be accommodated whilst retaining a commitment to the rational interpretationist thesis that interpretation and intentionality are structured by rationality.

4. IRRATIONALITY

4.1. THE CHALLENGE FROM IRRATIONALITY

Do We Need Rationality?

One of the most clear-cut challenges to the thesis of rational interpretationism emerges from the vast empirical literature on human reasoning, particularly studies on the consistency of people's beliefs[65]. Experimental psychological research on choice preferences, attributions and belief assertions is littered with examples of what Festinger (1957; cited in Elliot & Devine, 1994) first termed "cognitive dissonance": the common phenomenon of subjects holding beliefs that are inconsistent with one another, and acting in ways that are at odds with their professed beliefs, values and desires.

In this chapter I examine whether the claim that rationality is constitutive of intentionality can accommodate instances of intentional behaviour that seem to violate this central premise, given that *"the basic methodology of all interpretation tells us that inconsistency breeds unintelligibility"* (Davidson, 1982, p.184). Two objections to rational interpretationism will be outlined, each followed by a defensive move aimed at neutralising the objection posed. The main criticisms can be identified as challenges to the two components of the rational interpretationist view described previously, namely the epistemological Rationality Constraint, which sets normative parameters of truth and coherence for interpretation, and the ontological Rationality Assumption, which claims that rational norms are constitutive of the intentional realm.

The first argument from irrationality challenges the Rationality Constraint. The Principle of Charity, from which the RC is derived, stipulates that in order to make intentional

[65] I here take consistency to be a prime example of a principle that Davidson considers constitutive of rationality; others include the principle of total evidence and the principle of continence, which directs one to perform the action judged best on the basis of all available relevant reasons. It is this latter principle that appears to be problematic for the *akrates* (Davidson, 1985a).

ascriptions an interpreter must assume that *"by and large a speaker we do not yet understand is consistent and correct in his beliefs"* (Davidson, 1973a, p.238), and this injunction is of central importance to the whole rational interpretationist project (Bortolotti, 2003, p.121). The objection to rational interpretationism, deployed by Bortolotti (2004b), aims to undermine these third-person requirements on interpretation by demonstrating that when we face problems in interpretation, such difficulties are not created by the agent breaching principles of rationality. She uses this point to argue that interpretation therefore does not require the interpreter to attribute true and coherent beliefs, as an assumption that the agent is rational is not warranted. Bortolotti's contrasting account of interpretive difficulties is that they arise when an interpreter does not know enough about the context and environment the agent is in, and thus considerations of rationality are irrelevant to interpretation. She seeks to deny that rationality constrains interpretation by reducing it to a psychological heuristic that may serve a pragmatic function but that is not essential. The pragmatic use of rationality is particularly applicable in cases in which the agent's behaviour can be made more intelligible and predictable by interpreting his beliefs or actions as irrational, in much the same way that Humanity counsels. Taking rationality to be a heuristic allows that attributions of irrationality are sometimes more explanatorily appropriate: making sense of someone does not always involve maximising true belief or attributing consistency (Henderson, 1987, p.366).

If the RC is to serve its purpose in guiding interpretation, attributions of inconsistency need to be accommodated without thereby vitiating the claim that rationality is essential to the interpretive project. Although I have already outlined arguments rejecting the use of the norms of rationality as a pragmatic constraint on interpretation, here I expand upon the points previously raised by defending the role of rationality in the structure of reason explanations, arguing that rationality is in fact essential to interpretation: without a broad normative background of rationality, reason explanations would cease to function as explanations of intentional action.

The second and more fundamental argument from irrationality challenges the Rationality Assumption, that intentional states are at least partially constituted by the rational relations they bear to one another. The argument proceeds through an inference of *modus tollens*: one consequence of asserting the RA is that the logical possibility of a creature being an intentional agent and also being completely irrational (holding largely false and inconsistent beliefs) is ruled out. If rationality is a prerequisite for the very possibility of having beliefs and propositional attitudes, an intentional agent could not fail to be rational: *"to the extent that we fail to discover a coherent and plausible pattern in the actions and attitudes of others we simply forego the chance of treating them as persons"*[66] (Davidson, 1970a, pp.221-2). Hence critics (e.g., Bortolotti, 2004b; Henderson, 1987; Sesardic, 1986; Stich, 1983) cite obvious instances of the conditions of intentional agency being fulfilled in spite of inconsistencies in asserted or attributable beliefs as a refutation of the rational interpretationist's *a priori* commitments. Irrationality does pervade our belief systems, decision-making processes and actions: we frequently commit errors in reasoning, falling prey to logical fallacies and making intransitive choice preferences. In none of these cases would it be denied that the speaker is an intentional agent. Thus the *prima facie* objection to rational interpretationism arises in the form of a straightforward rejection of the RA. Against this challenge I suggest that taking the implications of the holism of the mental seriously relaxes these requirements for individual instances of intentional behaviour. This move enables instances of irrationality to be accommodated and understood as intentional against a background of broadly coherent, true beliefs. This is not a concession to the critic of the rational interpretationist view but rather a development of the idea that constitutive holism applies across the intentional realm.

[66] Consistent with Davidson's conception of intentionality, I take it that 'persons' here equates to 'intentional agents'.

Davidson's strategy for accommodating irrationality introduces us to a further problem for the Rationality Assumption, however. He suggests that intentional attributions that appear to breach the assumption of rationality can be tolerated by his view on the condition that agents seek to resolve and eliminate errors once they are made aware that, for example, they hold mutually inconsistent beliefs. I term the notion that agents ought to recover from epistemic errors the 'Rationality Requirement' (RR), and it is an explicitly normative prescription that we ought to be rational in our beliefs and actions. Davidson thus seeks to salvage his constitutive view of rationality from the threat posed by the existence of irrational beliefs by appealing to the inherent normativity of belief: the suggestion is that whilst it is possible to go wrong (for example, by holding false or inconsistent beliefs), in doing so one is deviating from what one ought to do or believe. This requirement, however, appears to impose a kind of rational imperialism on the beliefs and actions of others, setting the boundaries of intentionality according to what agents rationally ought to believe or do in a given situation. Yet the interpretability of actions and utterances does not seem to be beholden to their conforming to what an interpreter considers they rationally ought to be. Why are we entitled to believe that an agent has somehow gone wrong in his beliefs and judgements (and that he ought to believe or judge otherwise) if his behaviour does not appear to conform to the norms of truth and coherence?

The RR does sufficiently explain our tendency to recover from error in our beliefs and judgements when they are pointed out to us. But responding to the objection that it is a demanding and stringent constraint on intentionality draws attention to a tension in the rational interpretationist view I am advocating, between the claim that the intentional is constitutively rational yet agents are perfectly capable of breaching these constitutive principles without being condemned as not doing or believing what they ought. This brings us back to the question of how the normative force of rationality ought to be construed. In describing what would count as a failure of rationality, Davidson's articulation of the rational requirements operating on intentional agents creates an

overly prescriptive notion of rationality that is unnecessarily stringent for the purposes of advancing rational interpretationism about the intentional realm. It is a mistake to consider that the normative obligation we are under as intentional agents is one that can be described in principled terms. Indeed, I suggest that much of the criticism directed at the rational interpretationist view is based upon a misconception of the nature of the normative demands placed on intentionality and interpretation. There is, however, a way of conceiving this requirement that does not place intolerably strong demands on the rationality of an agent's behaviour. I consider whether a route can be navigated that avoids the charges of the arguments from irrationality whilst acknowledging the intrinsically rational normativity of the intentional realm, by rejecting the view that the norms of rationality take the form of prescriptive obligations towards abstract principles of truth and coherence. In the next chapter I will develop the argument against the codification of standards of rationality, arguing instead that the norms of rationality emerge from the shared practices of intentional behaviour.

4.2. AGAINST THE RATIONALITY CONSTRAINT

Attributing Irrationality in Interpretation

Agents frequently do act for bad reasons, make irrational plans and hold beliefs that they ought not to on the balance of evidence, yet we are still able to ascribe them such beliefs and explain their actions in intentional terms (Bortolotti, 2004). With this in mind the project of interpretation looks to be less constrained in what beliefs are potentially attributable to an agent than the demands of the Rationality Constraint would suggest: unlike the Quinean method of translation, contradictions are not completely precluded by strong *a priori* commitments to the logicality of intentional state ascriptions. If this is the case then the RC that is supposed, on the rational interpretationist view, to act as a normative framework for interpretation, appears unnecessarily stringent. Some philosophers have indeed suggested that conforming to constraints of rationality is necessary for the ascription of intentional states at all: *"when we are not [rational], the*

cases defy description in ordinary terms of belief and desire…when acts occur that make no sense, they cannot be straightforwardly interpreted in sense-making terms" (Dennett, 1987, p.87). Derived from the demands of Charity, the RC compels the interpreter to attribute to the agent beliefs that are coherent and true by his lights, but the potential occurrence of inconsistencies in an agent's belief set indicates that successful interpretation might not in fact hinge on the use of the RC.

This problem is particularly pressing for a Davidsonian theory of rational interpretationism because of the close connection Davidson forges between language and thought. For him, the RC is not a constraint on interpretation that can be loosened or abandoned, since to do so would threaten the entire basis of his theory of mind, language and mental content. Recall that from the standpoint of the Radical Interpreter, the attribution of beliefs and what a speaker means by his utterances are both, as it were, up for grabs. Neither can be assumed and the process of interpretation is a matter of gradually building up a coherent picture of meaning and belief at the same time, refining these attributions as behavioural evidence supports one interpretation or another. Suppose therefore, that in interpreting a speaker our best interpretation leads to the conclusion that he holds beliefs that are inconsistent with one another[67]. We have no reason to suspect the speaker is being insincere, speaking in metaphor or attempting to deceive us, or that he is being subject to perceptual error. Davidson suggests that in such an instance we should seek to revise our translations or intentional state attributions, adjusting our theory of meaning for that language accordingly: perhaps we have mistranslated the meanings of the utterances (1974b, p.196). Davidson's efforts to bind the development of theories of meaning to theories of intentional action (stipulated by his view of the interdependence of belief and meaning) entail that whatever our strategy for interpreting the actions and utterances of an agent is, it carries implications for our theory of meaning for the language of that agent. Such

[67] If the requirement of consistency is taken to mean that the set of our beliefs should not logically imply a contradiction, it is far too strong a constraint: none of us do or indeed could satisfy such a condition (Føllesdal, 1984)

a theory is supposed to enable us to make sense of behaviour in terms of the rationality of the attributed beliefs and desires of the agent but an unfortunate side-effect of this position, fully admitted by Davidson, is that if our account involves attributing inconsistency in that agent's beliefs, the theory of meaning is undermined (Henderson, 1987, p.361). The mechanics of the strategy for dealing with this problem need not detain us here: what is of interest is the suggestion that on a Davidsonian account of interpretation, which relies on the RC, attributions of inconsistency are barriers to successful interpretation and potentially breed unintelligibility.

Against this view is the claim that the RC is not a necessary constraint on interpretation, supported by the observation that the interpretability of intentional behaviour may in fact require attributions of inconsistency or error to be made. It is entirely plausible to argue that, as a matter of fact, speakers' beliefs generally do conform to the demands of rationality and that interpretation based on the RC is both possible and helpful in the attribution of intentional states, explanation and prediction of behaviour. Indeed, critics of rational interpretationism such as Bortolotti (2005; 2004a; 2004b; 2003) and Henderson (1991; 1987) embrace the notion that an assumption of rationality is a useful heuristic for the methodology of interpretation, but reject the view that the demands of the RC amount to necessary conditions. This particular line of criticism attacks the necessity claim by demonstrating that interpretation and the provision of reason explanations for intentional behaviour do not require any assumption of rationality, either for the interpreter or the agent whose behaviour is under observation.

Is Rationality Necessary to Interpretation?

Bortolotti (2004b) addresses the challenge posed to interpretation by instances of apparently irrational behaviour. She seeks to undermine the RC by demonstrating that more psychologically realistic explanations and predictions of such behaviour are available once the RC is abandoned. The specific target of Bortolotti's attack is the

notion that the predictive and explanatory success of our folk psychological attributions rests upon the interpretive constraint to attribute true and coherent beliefs. On her view the interpretation of behaviour depends only upon such factors as the contextual information available to the interpreter. Hence when irrational behaviour occurs, what renders interpretation difficult is the potential paucity of such information, not the fact that the behaviour exhibits an apparent violation of norms of rationality. Bortolotti aims to develop an account of interpretation that, whilst incorporating the normal concepts of folk psychology, eliminates what she considers an overly stringent and unnecessary constraint on the ascription of intentional states and explanation of behaviour. Here I sketch out Bortolotti's distinction between a demanding rationality constraint and what is necessary for intentional explanation. I then argue that the Rationality Constraint she attributes to Davidson is indeed too strong to be a necessary constraint on interpretability. However, I suggest that in appealing to the utility of folk psychological reason explanations Bortolotti does in fact rely on a thinner notion of rationality for the explanation and prediction of behaviour, because it is implicit in the very idea of actions being guided by reasons.

Reason Relations and Rationalising Explanations

When interpreting an agent, his utterances and actions may pose difficulties for our attempts to ascribe to him beliefs and desires. To use Bortolotti's own example, take the case of a man who asserts that there are flies in his head (2004a, p.359). We may be unsure whether he is using words correctly or if he genuinely believes that flies have got inside his brain, but either way such a case is puzzling, and we may suspend judgement as to what he means or believes. The fact that we find this interpretive situation difficult is uncontroversial, irrespective of the theory of interpretation employed. I have already argued that because rationality is a holistic and flexible constraint it has the resources to accommodate and render intelligible cases in which agents appear to have inconsistent or false beliefs. But according to Bortolotti (2004a,

p.361), for the rational interpretationist the reason that interpretation becomes difficult here is because the behaviour appears to violate norms of rationality.

Whilst advocates of rational interpretationism assert that an assumption of rationality is necessary for interpretation, none would wish to claim that it is a sufficient condition. There may be many other relevant requirements that are also needed in order for beliefs and desires to be successfully and usefully ascribed, such as an understanding of the specific context of the behaviour, the agent's history, some grasp of the functioning of human memory, awareness of the social and cultural background, and so forth. If we are committed to rationality being a necessary condition for interpretation, problems for or failures of interpretation need not be caused by violations of the norms of rationality. For instance, the interpreter could be unaware of limitations in the speaker's vocabulary, an unusual tendency to confuse words, speak in riddles or metaphor, be insincere or deceptive or that he is speaking in an unfamiliar cultural idiom. Bortolotti's own cited example of the man who claims he has flies in his head illustrates this point: whilst it is challenging for interpretation, it is not clear that such a belief does violate any norms of rationality. It is highly implausible but without further background information about the speaker's beliefs and perceptual experiences we cannot claim that the assertion of a particular belief entails that the speaker is departing from the canons of rationality. On the basis of the claim that rationality is a condition of possibility for interpretation, it is perfectly possible for failures or difficulties of interpretation to arise that have no bearing on the claims of the RC.

Bortolotti makes an intriguing suggestion as to the origin of the Rationality Constraint for intentional explanation, claiming that it rests on an equivocation between the idea that behaviour can be explained in folk psychological terms on the basis of the ascription of beliefs and desires, what she terms a *"schema for practical reasoning"* (Bortolotti, 2004a, p.362) and the normative demand that behaviour conforms to some rational standards, termed *"rationality as optimization"* (ibid.). Whilst intentional

explanations make sense of an agent's behaviour by rationalising it, that is, by articulating the reasons for his beliefs and actions, Bortolotti argues that evaluating such reasons in terms of their conformity to some normative rational standard is a separate and demanding move that is superfluous to determining the status of the behaviour as intentional. One's reasons need not be good reasons in order to qualify as intentional. In interpreting behaviour for the purposes of explanation, there need be no presupposition that the beliefs and desires being ascribed to the agent are conforming or aiming to conform to standards of rationality (ibid. p.363).

Drawing on a distinction made by Pettit and Smith (1990) helps elaborate upon the claims Bortolotti makes both about what the RC on interpretation consists in and why it is, on her view, false. Suppose that we can ascribe beliefs and desires to an agent without consideration for any of the underlying theoretical apparatus of interpretation. Given the set of beliefs and desires we can attribute to this agent, we can make sense of his action, that is, his behaviour can be seen as reasonable in light of the beliefs and desires he possesses. Pettit and Smith refer to this as the intentional conception of human beings (ibid. p.565) and it is a widely used picture of how we engage in folk psychological explanation and prediction of behaviour. They contrast this with the deliberative conception, according to which intentional behaviour is explained not only by a motivating reason, but also on the grounds that the action is appropriate, justified, or in some respect the correct option, given the agent's own values and standards (ibid. p.566). This is an explicitly normative conception of intentional behaviour, regarding what an agent ought to do in a given situation given his relevant beliefs and desires. Bortolotti draws a rough parallel between this deliberative conception and the demands of the RC as she portrays it.

If the deliberative conception of rationality is the one that constrains intentional attribution, this is indeed far too strong a constraint. For if the RC requires that in order to qualify as intentional, the reasons attributable to an agent for belief and action must

be characterisable as good, justified or appropriate, it is immediately obvious that it must be false: *"Rationality, as a responsiveness to norms governing the relations between propositional attitudes, is too demanding"* (Bortolotti, pers. comm.). If not rationality then, what does constrain the process of intentional ascription by an interpreter? On Bortolotti's view, all that is required is that an agent's behaviour is capable of being given an explanation in terms of the beliefs and desires that can be attributed to him. A capacity to adhere to a schema of practical reasoning, to act according to one's beliefs and desires, is all that is necessary for one's behaviour to be construed as intentional and amenable to reason explanation: *"[A]n intentional system does and says things for reasons…but this does not mean that their behaviour meets any normative standards of rationality"* (2004a, p.369). The ascription of beliefs and desires does not depend on their being justified, or of the agent behaving as he rationally ought given his set of intentional states. To use Bortolotti's own example (ibid. p.372), say that an interpreter is attempting to ascribe beliefs to an agent *X*. *X* asserts a belief that *p*, acts upon that belief and can provide reasons for his holding *p*: all behavioural evidence enables the interpreter to ascribe the belief that *p* to *X*, and this is a successful instance of interpretation. It is irrelevant to the ascription of the belief whether *X ought* to believe *p*. There may be good reasons why, all things considered, *X* ought not to hold that belief but this doesn't undermine the possibility of the interpreter ascribing that belief, nor of the fact that this is the most intelligible explanatory attribution to make.

Bortolotti's descriptive account of intentional explanation is intended to demonstrate that successful interpretation is not contingent on attributions of truth and coherence at all. She claims to have dispatched with the element of rationality in interpretation by distinguishing the possibility of providing intentional explanations for behaviour from the stringent normative requirement of adhering to principles of rationality, arguing that the former is sufficient for interpretability and, in consequence, intentionality. Thus whilst an agent may act in a way that ostensibly violates the norms of rationality, if an interpreter

can nonetheless rationalise this behaviour in light of the agent's beliefs and desires then he has all that is required for intentional explanation. It is possible to evaluate an agent's behaviour as rational or irrational, depending on whether he has good or bad, appropriate or inappropriate, justified or unjustified reasons for his beliefs and actions, but such evaluation is, on her view, independent of and subsequent to the process of ascribing such beliefs through the process of interpretation.

Interpretation without Rationality

In order to defend the RC from Bortolotti's attack I will now consider the cogency of her claim that the provision of reason explanations for behaviour is not necessarily underpinned by the normative requirements of rationality. I claim that reason explanations for behaviour cease to be possible if this is the case: without any normative constraints on interpretation and interpretability it is not clear how the ascription of intentional states enables behaviour to be explained at all. Arguing that interpretation is an intrinsically normative process I nonetheless also suggest that Bortolotti's strong conception of rationality as optimization ought to be rejected as a condition on intentional agency: an argument taken up again later in the chapter (4.4).

The fact that reason explanations can account for individual instances of interpretation without reference to their being good, justified or appropriate reasons leads Bortolotti to the assertion that there is no necessary general requirement on intentional states to broadly conform to principles of rationality, nor is the interpretive process necessarily guided by the Charitable concerns for coherence and truth. Although her main focus is on the epistemology of interpretation the ontological question of the nature of intentional states is pressing. Bortolotti's remarks on interpretation suggest she agrees, with the interpretationist, that interpretability is required for an agent to count as an intentional being, since what is relevant to characterising reasons depends at least in

part on their being attributable by a third person[68]. She thus accepts there is a connection between the possibility of forming reason explanations for behaviour and the conditions of intentionality, but rejects the view that rationality is what sets the boundaries on interpretability (2004a, p.365). Her conception of interpretation assumes that when an interpreter ascribes intentional states and generates reason explanations, a characterisation of the relevant intentional states can be given that does not itself rely on any notion of rationality.

Bortolotti does not address this question of the conditions for intentional ascription directly but hints that reason explanations can be wholly supported by a causal story about the relations between beliefs, desires and actions. She suggests that ascribed intentional states may be explanatory of an agent's behaviour in virtue of their bearing the *"right causal relations"* (ibid. p.365) to one another: belief and desire pairs motivate action through causal, as opposed to rational relations. Whilst the rational interpretationist can acknowledge that intentional states are at least partly constituted by such causal relations, what provides the explanatory force of reason explanations are their rational relations. These carry normative implications for the agent's other beliefs, desires and actions, and this normativity is reflected in the way reason explanations function: the intelligibility of behaviour is dependent on it being in some sense appropriate in light of attributable beliefs and desires. By contrast Bortolotti's conception of reason explanations as underpinned by causal relations and independent of rationality appears to class reasons and intentional state attributions as arational. On an account stripped of all notions of rationality then, the causal connections obtaining between attributed intentional states must carry the fully explanatory weight of the interpretation. Furthermore, an account is owed as to what differentiates reasons and intentional actions from causal chains resulting in mere vocalisations or physical

[68] Using the example of an apparent inconsistent believer, she states: "If Mark's adherence to a schema for practical reasoning is totally compromised, then Mark will not count as an intentional system" (2004b, p.365), implying that the provision of reason explanation does mark out a condition on intentionality.

movements: without the normative weight of rationality, what is intentional about intentionality?

Interpretation is, on this view, contingent on the interpreter's ability to ascribe intentional states to an agent without any normative guidance regarding what ought to follow from the holding of a given belief[69]. This raises the question of how the ascription of beliefs and desires could serve the purpose of providing reason explanations and enabling the prediction of behaviour: *"[a] belief theory with no rationality restrictions is without predictive content; using it, we can have virtually no expectations regarding a believer's behaviour"* (Cherniak, 1981, p.164). It is not clear that interpretation would be possible without the normative constraints articulated by the thesis of rational interpretationism. Consider Davidson's project of Radical Interpretation, starting from the basis of the behavioural evidence available to an interpreter regarding the propositional attitudes a speaker holds true. Without being able to assume what utterances mean in advance of ascribing beliefs and vice versa, there is an inevitable paucity of evidence from which to launch the interpretive project. Permit for a moment that intentional attribution is, as Bortolotti claims, possible without any normative rational constraints. Observing the behaviour of an agent *A*, an interpreter tentatively attributes to him the belief 'that-*p*', and this attribution can be explained in causal terms: some event or object in the world caused the belief 'that-*p*' in A. What follows from this attribution? A purely causal story might suggest that inductive inference suffices to account for what follows: something of the form 'in situations akin to *A*'s, agents holding the belief 'that-*p*' are disposed to believe '...', where the ellipsis is filled in by such propositions as 'not not-*p*' or whatever causally coincides with the belief 'that-*p*'. But even granting the availability of such inductive evidence about what tends to follow from a particular belief, the attribution of this isolated belief does not tell the interpreter much at all: he can go no further in his interpretation than causal chains of inference, to

[69] An assent theory of belief ascription is one such theory that involves no requirement of rationality at all, e.g., Russell's 'On Propositions' (Cherniak, 1981, p.163).

hypothesise what else the agent might believe or to say why an agent performed this or that action. The language of causal concepts is simply inadequate to explain how a reason has motivational (and explanatory) force. Say for example that an agent puts the kettle on, and I seek to explain this action via an attribution of reasons: he wants to make a cup of tea. Undoubtedly, casual mechanisms are at work in his action, but what furnishes the reason with explanatory power is not the fact that an attributed desire for tea caused him to raise him arm and flip the switch on the kettle. Whilst I would not wish to deny the causal role beliefs play in action-guiding and in relation to one another, I suggest an explanatory account of why agents behave as they do on the basis of the beliefs (and desires) they possess cannot be couched in the language of physical-causal concepts alone: the causal story does not do justice to the kind of inferences an interpreter draws when making intentional attributions[70].

Without any normative guidance as to what the agent ought to believe or do in light of this belief, for example, that he ought not to also hold the belief 'that not-p', the interpreter cannot formulate hypotheses and test them against the behavioural evidence available. In short, interpretation cannot progress:

> "[I]f the believer were not required to be at least more likely to undertake some of the apparently appropriate actions, then the attribution of a belief-desire set could never yield any predictions of behaviour, and would never be disconfirmable by observed behaviour. On the basis of such an attribution, no behaviour could be expected; every action would be equally probable" (Cherniak, 1981, p.166).

Bortolotti is committed to the claim that actions are guided by reasons, but in rejecting the necessity of rationality in reason explanations we are left questioning what disciplines the guiding of actions by reasons, if it is not a normative commitment of some kind, such as 'if I believe *x* then I ought not to believe not-*x*'? The implicit normativity of reason explanations is implied by Bortolotti's gesture towards seeking the "right" causal connections in interpretive attributions, indicating that successful explanations for behaviour are picked out by the normative appropriateness of the

[70] See Campbell (2009; 2007) for recent views on the role of causal processes in person-level psychological explanations.

relation between beliefs, desires and action. A causal account does not, however, do justice to the intuition that if the agent holds the belief 'not not-*p*' it is because that is *what he ought to believe*, given his belief 'that-*p*'. Whilst reasons may also turn out to be capable of description in causal terms, perhaps as physiological brain states, what characterises them as reasons just is what is revealed through the project of Radical Interpretation: a normative process that seeks to render behaviour intelligible through making intentional attributions on the basis of what it would be reasonable or correct for an agent to believe, say and do in light of his epistemic context:

> "What makes the task practicable at all is the structure the normative character of thought, desire, speech, and action imposes on correct attributions of attitudes to others"[71] (Davidson, 1990, p.325).

The description of beliefs, desires and actions in rational, reason-giving terms provides the explanatory power of folk psychological intentional explanations of behaviour. In seeking a theory of intentional explanation based upon the attribution of reasons to an agent, Bortolotti thereby inadvertently carries an intrinsic element of rationality into her account of intentional behaviour as essentially rationalisable. Belief-and-desire pairings sufficient to motivate an action can be attributed even in the case of behaviour that goes against the norms of rationality: the very fact that we can characterise the behaviour in this way ensures that it lies within the domain of rationality rather than brute causal explanation.

I have suggested that rationality is necessary for the characterisation of reasons but the account I have sketched thus far does not touch upon the objection that imposing the normative demands of rationality on conditions of intentionality and interpretation is an overly restrictive and unnecessary constraint. Bortolotti's argument for this view is that the Rationality Constraint on interpretation rests on an equivocation between the conditions for intentional agency and an adherence to principles of rationality. However, this conception of the RC is based on a misconstrual of the relationship between what it is to be rational and what it is to be irrational. The case that is not

[71] This conception of the mental also underpins Davidson's thesis of the anomalism of the mental (e.g., 1973a; 1970a).

amenable to reason explanation arises not when an agent violates norms of rationality in a particular instance, as in the case of irrationality, but when the behaviour- or belief-causing states are not capable of characterisation in terms of rational relations at all, that is, when they are arational. In irrational behaviour the rational link between belief attribution and action is maintained, thus enabling a reason explanation to be given even if it fails to be a good reason. Bortolotti's (2004a; 2004b) objection to the necessity of rationality commits the error of thinking that because irrationality consists in a failure to have good reasons or to act as one ought, the requirements of rationality demand that one does have good reasons for one's beliefs and actions. If the possibility of intentionally characterising an action depended on there being good reasons for it, there would be no sense in which behaviour could be deemed reasonable or unreasonable, sensible or foolish, appropriate or inappropriate, rational or irrational, for there would be no logical gap between full and complete rationality and complete unintelligibility. I turn now to consider how the claim that the intentional realm is constituted by norms of rationality, which I call the Rationality Assumption, can best be cashed out and defended.

4.3. AGAINST THE RATIONALITY ASSUMPTION

Incoherent Contradictory Beliefs

It is worth clarifying what possibilities for belief ascription are explicitly ruled out by the Rationality Assumption and what empirically feasible possibilities need to be accommodated if it is to hold as a plausible account of the structure of the intentional realm. The thesis rejects the logical possibility that an agent could sincerely and explicitly believe the conjunction of two inconsistent beliefs '*p*' and 'not-*p*'. That is to say, an agent cannot form a belief that endorses two contradictory statements and be cognisant of this inconsistency: *"no one can believe a proposition of the form (p, - p) while appreciating that the proposition is of this form"* (Davidson, 1985a, p.198). The very notion of an agent explicitly asserting '*p* and not-*p*' is unintelligible: an interpreter

would simply be unable to attribute a belief with propositional content to an agent making such assertions. Davidson terms this kind of irrationality 'synchronic inconsistency', and it is doubtful whether an assertion of such a manifest contradiction could possibly be attributed to an intentional being: *"[i]t is by no means easy to conceive how a single mind can be described in this way"* (ibid. p.197).

In a hypothetical case in which a speaker sincerely asserts a belief with the propositional content '*p* and not-*p*' (in a language that is understood, or in Davidson's terms, for which the interpreter has a reliable and accurate theory of meaning), it is not at all clear what the content of this assertion could be taken to be, or what was meant by the utterance. Such an event would be *"an occasion for sounding the epistemic alarm"* (Dennett, 1987, p.95). It is reasonable to suggest that without further behavioural evidence available for interpretation or the possibility of seeking clarification, the interpreter would lose traction on what it is the speaker meant by his utterances and what he believed. Few, however, would wish to challenge this point: such an assertion by a speaker would be incoherent and nonsensical irrespective of one's commitment to the relationship between rationality and intentionality supported by the rational interpretationist. Synchronic inconsistency falls below the threshold of what could count as intelligible, intentional behaviour (Davidson, 1985a) and does not therefore threaten the argument for the inherent rationality of intentionality.

Non-Obvious Inconsistencies

Whilst holding a belief of the form '*p* and not-*p*' is unintelligible the empirically robust fact that agents may hold beliefs that are inconsistent with one another, holding a belief '*p*' and also, perhaps inadvertently or at a different time, a belief 'not-*p*', poses a challenge for rational interpretationism. Such inconsistency is, for Davidson, at the heart of the problem of irrationality: *"we call a single attitude, belief, or action irrational only when we assume it conflicts with other beliefs or attitudes of the agent"* (1985a, p.193). Inconsistency between beliefs that can only be identified and asserted through

detailed questioning or inference, in the manner of those elicited from the slave boy in Plato's *Meno*, are undeniably common. This fact alone suggests it is probable that contained within one's network of intentional states one might hold beliefs that are inconsistent with one another. Given that inconsistency is often taken to be the paradigm of irrationality, how can rational interpretationism tolerate and account for the existence of such inconsistencies? Any explanatory theory of intentionality predicated on the presupposition of rationality faces this *prima facie* problem:

> "…it would not seem possible to have a propositional attitude that is not rationally related to other propositional attitudes. For the propositional attitude itself, like the proposition to which it is directed, is in part identified by its logical relations to other propositional attitudes" (Davidson, 1985a, pp.189-90).

Arguments seeking to explain this phenomenon whilst retaining a connection between intentionality and rationality often make use of the idea that the problem posed by inconsistencies of belief depends on their degree of salience to the agent (Føllesdal, 1984, p.305). Within the totality of an agent's beliefs there may be lurking inconsistencies of which he is by and large oblivious and if such beliefs are not made salient simultaneously the irrationality remains, as it were, undiscovered. In holding a particular belief it is implausible that one holds true each of its logical consequences. Omniscience and infinite cognitive resources would be required to fulfil such a demand. Hence whilst holding a belief with self-contradictory content of the form (*p*, - *p*) would be incoherent, tacit inconsistencies may arise, for example through a failure of inference in which one is not aware of one's logical error in holding a set of beliefs of the form (*p, p→ q, - q*). If mutually inconsistent beliefs are not attended to or 'activated' at the same time, then such beliefs may operate independently of one another in the agent's mental economy, even if they are directed towards the same propositional content[72]. The rational interpretationist can thus posit that the demands of rationality constrain only "activated" beliefs (Bortolotti, 2003, p.118), which are those beliefs that

[72] We can distinguish between beliefs being endorsed, activated and attended to, to discriminate between different levels of awareness at which beliefs may be held. These distinctions make more plausible the idea that non-obvious inconsistencies are an ordinary feature of intentional life: beliefs held at, as it were, different levels, may manifest inconsistency that the agent is oblivious to (Bortolotti, 2003, footnote i).

are being consciously considered by an agent at a given time, as these are the cognitive states that motivate action and are thus attributable by an interpreter: *"the activated belief subset is subject to a more stringent inference condition than the inactive belief set"* (Cherniak, 1981, p.178).

Although an intuitively plausible account of how inconsistencies may arise and persist, the level of an agent's awareness of his beliefs does not affect the logic of belief individuation. If we take it that the degree of salience or 'activation' of beliefs is extrinsic to the conditions of their individuation, then the claim that beliefs can be possessed at different strata of attention is orthogonal to our concern here with defending the assumption that beliefs are intrinsically rationally related. Thus the fact that one is ignorant of an inconsistency in one's beliefs does not in itself do anything to explain how inconsistencies are possible on a theory that is based upon the claim that propositional attitudes depend for their identity on the rational relations they bear to one another. What this unsuccessful strategy for defending the RA does suggest, however, is that if such compartmentalising of conflicting propositional attitudes could apply to the logic of their individuation, the logical tension inherent in the idea that an agent could hold inconsistent beliefs would dissipate. In other words, if the rational relations that constitute the identity of a belief 'that-*p*' are psychologically isolated from those that constitute the belief 'not-*p*', the threat to the RA from the existence of inconsistency would be neutralised.

Davidson's notion of mental partitioning is based on this idea of tolerating inconsistency by appealing to the relative isolation of incongruent beliefs from one another, and on his account it is not the agent's degree of conscious awareness that bears the explanatory weight of the division. Following a rich psychoanalytic tradition originating with Freud, he claims that the mind can be sub-divided into "psychic regions" (Heil, 1989, p.574) that, whilst fully rational and consistent within themselves, may contain beliefs that are inconsistent with those in another psychic region of the individual's

mentality. Thus local perturbations in rationality may emerge, such as inconsistencies that are identifiable as such only across the boundaries of different partitions. Davidson's partitioning strategy is intended to accommodate the possibility of holding inconsistent beliefs by circumscribing the extent to which the holistic identity conditions of intentional states are distributed across the whole network. Hence instances of psychological breakdown in which intentional attributions nonetheless apply can be accommodated by conceiving of psychic regions as largely autonomous islands of rationality, within the boundaries of which the identity of particular intentional states can be secured.

Such a compartmentalising strategy faces the problem that it undercuts one of the central features of the rational interpretationist conception of mentality: the holism of the psychological realm as a coherent body of relationally constituted intentional states. The partitioning argument proceeds on the basis of the assumption that the hypothesised holism of the mental is incompatible with the possibility of internal inconsistency, and therefore seeks to de-emphasise the holistic claim by narrowing its scope to separate sub-regions of an agent's mentality. Some cases of irrationality could be viewed as the fragmentation of this holism; for example, Radden proposes a conception of madness as a disintegration of the overall rationality of the mind (1985). This approach does, however, weaken the holism that was intended to be a key strength of the rational interpretationist thesis. I suggest then that rather than weaken the holistic claim, instances of inconsistency can be accommodated by taking the holistic aspect of the relational identity conditions for intentional states seriously.

To use a Quinean analogy, beliefs are identified by their location within a vast and elaborate web of interrelations with other intentional states, meanings and actions. No one belief is constituted by only one or two connections within the network but rather by many and varied relations. This entails that the identity conditions of particular beliefs may not overlap despite being in some way relevant to one another in terms of content.

Although there ought to be some rational relation obtaining between such beliefs in virtue of what they are about, constitutive holism allows that it is possible for this relation to fail whilst the different beliefs themselves are nonetheless maintained. It is then only by bringing the set of inconsistent beliefs to attention that this error is highlighted. Say for example that I believe that it is 1pm in England, where I am currently located, and I know that the clocks are set to British Summer Time. I also believe that Dubai is 3 hours ahead of BST and I desire to speak to my friend who lives in Dubai. I believe that she will be at home now because she finishes work at 5pm and thus form the intention to call her. Taken together, there is an inconsistency present: the consequence of one set of beliefs (regarding the time difference) is that I believe it is 4pm, whereas in forming the intention to make the phone call the belief that it is 5pm is also attributable to me. Nonetheless, if we conceive of beliefs as sets of interconnected nodes in a network, the propositional content of particular beliefs can be assured in virtue of their constitutive rational relations with other beliefs, even though they do not bear the appropriate rational relations to the beliefs with which they are inconsistent. Thus my belief about Dubai is still a belief about Dubai in spite of the failure of inference regarding its time zone, on account of numerous other beliefs I hold, such as the fact of it being located in the UAE, my friend's presence there, and so on. There are sufficiently many other beliefs and rational relations constitutively supporting the identity of each of the inconsistent beliefs that they can both be given an intentional description despite the logical tension between them. In this way the holism of the psychological realm ensures that beliefs that are inconsistent with one another can nonetheless be attributed to an agent (at least temporarily) without degenerating into incoherence, and without necessitating a compartmentalisation of the mind into discrete psychic regions.

Delusions and the Background Argument

The claim that identity conditions for beliefs need not overlap and can thus tolerate inconsistency explains how the assumption that intentionality is constitutively rational

can be retained in the face of evidence of reasoning mistakes and non-obvious errors in judgement. This account relies upon the idea that the inconsistent beliefs are the result of an error that is not obvious to the agent. However, a weakness in this defence of the RA is exposed by the existence of cases in which attributable beliefs are self-evident and persistent in spite of their falsity or lack of coherence with other beliefs. Considering the case of delusion exemplifies the problem faced by the RA here: such cases make it seem implausible to suggest that even broadly construed, rationality is constitutive of intentional states. Patients suffering from delusions often have a tendency to undertake a form of 'double book-keeping' through holding self-evidently inconsistent beliefs (Sass, 1994, p.86). Moreover, in cases such as Capgras syndrome, the delusional state itself is patently false but persistently retained, even if the patient acknowledges its implausibility and bizarreness. The belief is maintained in the face of overwhelming evidence to the contrary (DSM-IV-TR, 2000, p.275), the testimony of others he knows and trusts, the explanations of his doctors that he has suffered neurological damage, and he may even acknowledge the implausibility of the belief. The inconsistency cannot be written off as a rectifiable mistake or slip of the tongue and there are good reasons for characterising the delusion as an intentional state, as it plays a role in guiding action and rationalising at least some of the agent's actions: *"[Delusional subjects] behave as believers when they act on their delusion, when they offer tentative arguments for it and when they relate the content of their delusion to the other beliefs they hold"* (Bortolotti, 2005, p.206). Attributing the belief with the content 'that woman is an impostor' to the Capgras patient renders his behaviour (avoiding her, expressing distress at her physical presence and so forth) amenable to a reason explanation.

It therefore appears that the agent possesses a state that is action-guiding to a circumscribed extent, seemingly with determinate propositional content, available to conscious awareness and critical reflection, self-evidently bizarre and inconsistent with other beliefs, but that nonetheless does not satisfy the constitutive conditions of the RA

and thereby cannot fall under an intentional description. Two avenues are obvious from this point, neither of which is attractive from the perspective of the rational interpretationist. Either it follows that the norms of truth and coherence embodied by the RA cannot be conditions for the application of intentional concepts: a move that abandons the core of the rational interpretationist thesis by breaking the connection between interpretability and rationality, or the delusion is denied the status of a belief. The debate over the intentional status of delusional states is extensive (see e.g., Campbell, 2001), but even if such states are thought not to be intentional this creates further problems for the rational interpretationist, as it then follows that arational, non-intentional states are capable of entering into reason explanations and providing reasons for behaviour: an intolerable consequence for a theory that posits rationality as a condition of intentionality.

The most promising avenue for the rational interpretationist to defend the RA necessitates a shift of explanatory focus away from the isolated instance of delusional behaviour in order to view the irrationality instead as a localised disruption to a broad background of largely rational beliefs and behaviour. Thus rather than focusing on the individual constitutive conditions of intentional states, the rationality of intentionality is to be located at the level of the whole person, situated in an environmental and social context. This is essentially an extrapolation of the constitutive holism of the intentional previously discussed. The claim is known as the Background Argument (Bortolotti, 2005), and it aims to accommodate instances of irrational behaviour from within the conceptual sphere of rationality, rather than considering them as challenges to the RA. The Background Argument is that it is only against a background of broadly true and coherent beliefs that intentional descriptions can be given to behaviours and utterances that breach the norms of rationality described by the Rationality Assumption.

Davidson states that irrationality is *"a failure within the house of reason"* (1982, p.169), meaning that irrational behaviour and attitudes ought to be characterised as failures of

reason in a particular instance, not an absence of reason altogether[73]. Such failure can be understood only within the context of a background of intentional behaviour that does conform to the norms of rationality, that is, within a system in which the right logical and causal relations obtain between the world and the system, its beliefs and intentional actions:

> "To explain irrationality we must find a way to keep what is essential to the character of the mental – which requires preserving a background of rationality" (Davidson, 1985a, p.190).

> "[i]t is only by interpreting a creature as largely in accord with these principles [of rationality] that we can intelligibly attribute propositional attitudes to it, or that we can raise the question whether it is in some respect irrational" (ibid. p.196).

The rational background extends into the world, incorporating the agent's intentional actions, utterances and interactions with others. This argument is central to Davidson's theory of belief ascription and has been advocated as a strategy through which to immunise an assumption of the necessity of rationality for intentionality against instances of irrational behaviour (e.g., Heal, 1998).

What does it mean to say that agency requires a 'background' of rationality? Bortolotti (2005, p.190) suggests two conditions, drawn from Davidson. Firstly, that an intentional agent generally holds beliefs that conform to the norms of rationality, that is, they are mostly true and coherent; secondly that an agent will seek to restore rationality if he happens to be in violation of its norms. At present I am focused on the former condition, examining whether or not the Background Argument is capable of salvaging rationality as a feature of intentionality in cases where the propositional content of beliefs appears to be determinate but false and inconsistent with other intentional states. I will consider the requirements on belief revision and recovery from error in section 4.4.

[73] Considering delusions as involving an absence of reason would be consistent with Berrios' (1991) account of delusional utterances as "empty speech acts" devoid of meaning and not falling under an intentional description.

Aside from the difficulties of ascertaining how the condition of 'mostly' true is to be understood, it is not immediately clear how the Background Argument salvages the Rationality Assumption from the problem of delusions. Indeed, as Bortolotti argues, it appears to falsify the claim that agents are largely rational; there is a gaping hole in the rationality of the beliefs held by the delusional individual, who is an intentional agent nonetheless. Bortolotti concludes that considerations of rationality are thus irrelevant to determining the status of behaviour as intentional or not: it is not a constitutive aspect of intentional agency that one is broadly rational.

However, I suggest that Bortolotti's rejection of the Background Argument rests on a misunderstanding as to how the background of rationality ought to be construed. The notion that intentional agency requires a background of rationality has found expression in a variety of ways and not always in the language of truth and coherence. For instance, John Searle's "Background" and Wittgenstein's "Bedrock" refer to our pre-reflective dispositions and capacities (Rhodes & Gipps, 2008) rather than epistemic standards specifically. Nonetheless, arguments for a necessary background against which behaviour can be understood as intentional and action-oriented are all predicated upon the idea that our ability to get on in the world, to successfully act, survive and satisfy our goals, depends upon certain conditions. The Background Argument identifies these as being that our beliefs are largely true and broadly coherent with one another. Holding true beliefs generally enables one's actions to succeed in their intentions, and coherence is required in any situation in which values, beliefs and intentions motivate the pursuit of a goal. This assumption goes largely unnoticed, but *"what must be presupposed in any interpretative situation is the idea of the speaker's location in, and relationship to, a world"* (Malpas, 1992, p.89). Yet such presuppositions are imperfect and this is thrown into relief in cases such as delusions. Here, we face an admittedly baffling situation, and it is difficult to know how the delusional assertion is to be understood. But as Davidson observes:

> "The underlying paradox of irrationality, from which no theory can entirely escape, is this: if we explain it too well, we turn it into a concealed form of rationality; while if we assign incoherence too glibly, we merely compromise our ability to diagnose irrationality by withdrawing the background of rationality needed to justify any diagnosis at all" (Davidson, 1982, p.184).

The Background Argument thus reflects an *a priori* commitment to the overall cogency and groundedness of the attitudes and behaviour of intentional agents within which localised perturbations can be characterised intentionally. Irrationality is not the inverse of rationality but a domain within it. What is essential to this notion of the background is that it is not individualist in nature: it extends across the *"horizon"* (Malpas, 1992) that we share as intentional agents[74]. Without an assumption that an agent is by and large rational, we could not even apply intentional concepts to instances of irrational behaviour and single them out as being bizarre. It is only once this foundation has been laid that inconsistencies can be intelligibly attributed at all. Thus, it is precisely the idea that in the case of delusions there is some failure or breakdown in one's normal way of going in the world that makes the delusional belief stand out as indicative of there being something 'wrong' in the person's way of thinking:

> "We often, and justifiably, find others irrational and wrong; but such judgements are most firmly based when there is the most agreement. We understand someone best when we hold him to be rational and sage, and this understanding is what gives our disputes with him a keen edge" (Davidson, 2982, p.184)

I will return to the question of how best to characterise this notion of the background in chapter five, but here I wish to emphasise its holistic role as a prerequisite for identifying instances of irrationality in the first place. For instance, it is only because of the background that we can recognise the delusional assertion as a linguistic utterance rather than a meaningless vocal articulation, but also to identify it as somehow being out of place. It is not clear how Bortolotti's rejection of the Background Argument covers this possibility: if there is no background of broadly true and coherent belief, what makes the delusional assertion stand out as being odd, or indicative of there being something wrong in the patient's beliefs?

[74] Malpas (1992) and Evnine (1991, p.154) argue that Davidson's account of the background of rationality is not incompatible with a phenomenological approach. Although there are interesting parallels to be drawn here, I am interested in the implications of seeking epistemic conditions identifying the boundaries of intentional agency.

Permitting the possibility of inconsistently held, contentful beliefs does not therefore undermine the RA as a broad delimiter of the boundaries of intentionality: it is a feature of the intentional realm that admits of exceptions but a valid feature nonetheless. Rather than challenging the intrinsic connection between rationality and intentionality, the possibility of inconsistency and irrationality in attitudes and behaviour reflects the broad, messily holistic nature of the psychological realm. Content can be attributed intelligibly to beliefs that fail to cohere with others or are patently false, thus undermining the objection that such states are intolerable for the rational interpretationist.

4.4. OUGHT WE TO BE RATIONAL?

Recovery from Error

Having claimed that inconsistency can be accommodated by the rational interpretationist thesis, we now face a further problem. Since inconsistency in belief is possible and attributable by an interpreter, on the view I have developed thus far there is a danger of losing sight of the claim, central to the rational interpretationist thesis, that the boundaries on interpretability and sense-making in intentional behaviour are normative in nature. If inconsistency can be accounted for within a rational framework, in what sense is it a mistake to hold inconsistent beliefs? If we wish to retain the claim that inconsistency is indicative of a normative lapse, and that an inconsistent agent has gone wrong in his reasoning and epistemic processes of belief formation, a further requirement on intentional agency is needed.

The problem can be elucidated if we consider the intuitive differences between straightforward errors in reasoning, such as the mistake committed by the person who commits the conjunction fallacy or my unreflective inconsistency in beliefs about Dubai, and the case of the delusional individual who believes his wife is an impostor. Clearly,

the latter is irrational in a way that the former examples are not, but examined solely in terms of attributing intentional content to these agents, the RA and RC are incapable of distinguishing between them. Where they differ is in the processes by which these beliefs are maintained and subjected to revision in the face of evidence or awareness of the inconsistency. The defence I have given of the rational interpretationist thesis suggests that by and large, inconsistencies in belief are indicative of an error that the agent will seek to correct, given the appropriate cognitive resources. The rationality or irrationality of a particular belief thus lies not in the content that it has but rather the way in which the agent behaves in response to it. In the case of the conjunction fallacy, we would expect an agent to revise his beliefs once he realises the error in his probabilistic judgement. By contrast, the Capgras patient is characterised by the steadfast and stubborn conviction by which he maintains his delusional assertion, even whilst acknowledging its implausibility. These cases help identify the motivation for the intuition that the delusional patient is irrational whereas the reasoning task participant is simply mistaken. Irrationality arises not when attributed content is inconsistent or false, but when agents do not respond or react to such beliefs in the right way:

> "A person is irrational if he is not open to reason—if, on accepting a belief or attitude on the basis of which he ought to make accommodating changes in his other beliefs, desires, or intentions, he fails to make those changes" (Davidson, 1982, p.179-80).

In the instance of delusion, the agent does not recover from the false belief he has formed: he is failing to do something he ought when confronted with the bizarre nature of his beliefs. One way of describing what appears to be a normative obligation on intentional agents is to suggest that we ought to be rational in our beliefs, even though we might, and frequently do, fail to achieve truth and coherence. I will refer to this obligation as the Rationality Requirement (RR). Whilst it is clear that Davidson advocates the idea that intentional agents ought to be rational and there are clear explanatory advantages of stipulating such a requirement in this way, I will argue that it is a mistake to construe the norms of rationality that are constitutive of intentionality in terms of the prescriptive principles he suggests.

Ordinarily, if one is alerted to the fact that there is an inherent inconsistency in one's beliefs, or that one has acted in a way that is inconsistent with one's professed beliefs, desires and so forth, one would attempt to resolve what is perceived as a tension in one's intentional behaviour. The claim that agents usually seek to resolve inconsistencies is supported in principle by observations of psychological discomfort in studies of cognitive dissonance (Elliot & Devine, 1994). When subjects are aware of the conflict between their beliefs they feel psychologically uncomfortable and attempt to dissolve the dissonance either by explaining it away or by modifying one or other of the dissonant beliefs. A common example of cognitive dissonance occurs in smokers, who may hold the belief that smoking causes cancer but nonetheless acknowledge that they do continue to smoke[75] (Aronson, 1969). Here, the habitual smoker may be compelled either to stop smoking or to rationalise the dissonance by citing numerous other beliefs such as discrediting the scientific evidence, citing examples of family members who smoked for years without their habit leading to cancer, and so forth. Quite how the individual reduces the dissonance is beside the point: the fact is that he attempts to eliminate the inconsistency. Aronson observes that *"dissonance theory does not rest upon the assumption that man is a* rational *animal; rather, it suggests that man is a* rationalising *animal"* (ibid. p.3, emphasis in original). Thus whilst it is possible and indeed common to hold beliefs that fail to be true or coherent, it is an empirical fact that awareness of such inconsistencies is a source of psychological discomfort and that, time and cognitive resources permitting, we generally seek to eliminate them.

However, from empirical observation of the fact that agents tend to attempt resolution of manifest inconsistencies in their beliefs it is a further step to argue that they are normatively obliged to recover from such error, which is what the Rationality

[75] There is not a formal inconsistency in this case of the kind characterised by holding true both the propositions '*p*' and 'not-*p*'. Rather, together with the assumption that it is reasonable to suppose the agent would not wish to suffer cancer, the inconsistency between his beliefs is made manifest.

Requirement implies. Is the claim that we ought to be rational justified? Let us consider the implications of the idea that in holding inconsistent or false beliefs one has made an error that one ought to resolve. The normative prescription that one *ought* to fix such an error entails that the interpretation of one's behaviour by a third-person carries significant predictive power. An example used by Dennett (1987, ch.4) to defend his intentional systems theory will serve (in modified form) to illustrate how the notion that one ought to recover from one's errors equips interpreters with a powerful explanatory strategy. It might not be able to predict the occurrence of inconsistency or error, but it is nonetheless capable of explaining and predicting agents' responses to such mistakes in reasoning and action. Suppose that I see a boy selling glasses of lemonade at a stall by the side of the road. A sign on the stall reads 'Lemonade - £1.20 a glass'. I purchase a glass, handing over two pound coins to the boy. He then gives me a fifty pence piece and a twenty pence piece in change and thanks me for my custom. The boy has made a mistake in handing over my change: he ought to have given me eighty pence. The Rationality Constraint looks to be capable of generating intentional attributions in this scenario: we could attribute to the boy the belief he had given me the correct change[76]. However, it is also capable of allowing further attributions and predictions: we would expect that once his mistake has been pointed out to him, the boy would exhibit surprise, possibly embarrassment (ibid. p.85), he would apologise and hand over an additional ten pence. How can we predict this? It is in virtue of the idea that in giving me the incorrect change the boy has failed to act as he ought, and that he would seek to correct this error upon realising what he had done.

Going Against One's Own Norms

What content can be given to the idea that there are constraints on what one ought to do when faced with an error in one's beliefs? Davidson's strategy for explaining the fact that we do tend to attempt to recover from error is to claim that as intentional agents we are under an obligation to adhere to our own norms of belief and action. He argues

[76] Dennett himself argues that we cannot make definite explanatory attributions, but here I use the example only to highlight the scope and possible predictive power of attributing false beliefs.

that *"if someone does on occasion think or act or feel in ways that offend against those norms, he must have departed from his own standards, that is, from his usual and best modes of thought and behaviour"* (1985a, p.197). The normative dimension to intentional behaviour can be construed as the idea that we can fail to do and think what we have best reason for. Thus, what is wrong with being inconsistent in one's beliefs is that one is failing to adhere to one's own conception of what it would be right to believe. Recovery from normative epistemic breaches is instigated by the recognition that in holding inconsistent beliefs one is violating one's own standards of belief, and is compelled to resolve it in order to restore coherence (Davidson, 1982). According to the Rationality Requirement, the standards of belief in play here are those of rationality.

I may hold one belief for good reasons at one time, and another inconsistent belief at another, but what I cannot do (according to Davidson) is assert both beliefs simultaneously and be content to accept this tension, because to do so would constitute a breach of my own epistemic standards. This claim needs unpacking if we are to be clear about what obligations the Rationality Requirement are supposed to place on agents. The normative content of the RR is that one's beliefs ought to be true and coherent, both with one another and with one's intentional behaviour and utterances. Going against these standards would be indicative of irrationality (Davidson, 1985a, p.192). To clarify this point and the significance of the argument that irrationality is a matter of breaching one's own norms, let us consider the idea in logical form. An agent holds two beliefs that imply the negation of one another, one with the propositional content (*p*), the other, (-*p*). On the holistic construal of the RA and the RC I have developed, it could be both possible and rational for an interpreter to attribute these beliefs to the agent. But if we add the stipulation of the RR, the irrationality of such inconsistency is noticed: the rational norm of consistency derived from the law of non-contradiction, (if *p*, then not -*p*), has been breached. The principle sets out what ought to follow from the holding of a belief, in this instance proscribing the belief (-p) from being held. With that principle in place, the inconsistency indicates that the agent

has committed an error, since one cannot simultaneously hold (*p*, if *p* then not -*p*, -*p*). Agents whose behaviour does not obviously violate this principle can be described as conforming to it, even if not intentionally seeking to maintain consistency: *"It is true that believers conform to consistency in the sense that they do not engage in an obvious inconsistency"* (Bortolotti, 2003, p.118).

Introducing the notion of the agent's own standards has the advantage of providing a way of distinguishing the case of mistaken reasoning from that of delusion. In the former case, if conscious awareness of the inconsistency prompts one to revise one's beliefs or correct one's response, one is eliminating the error and seeking to maintain conformity with one's own epistemic standards. In the case of delusion, by contrast, the problematic belief is maintained and thus represents a continued and persistent breach of these standards, and it is this failure to modify one's beliefs in accord with the requirements of rationality that makes the belief an irrational one. According to this view then, irrationality arises *"only when beliefs are inconsistent with other beliefs according to principles held by the agent himself"* (Davidson, 1985a, p.192). This application of the Rationality Requirement extends beyond strictly epistemic concerns, for example, a set of beliefs or an action would be irrational if an agent acts against his own conception of what would be in his best interests, or against what he justifiably values. Here I am concerned only with the characterisation of irrationality as a breach of the norms operating on one's standards of belief formation and maintenance:

> "The possibility of (objective) inconsistency depends on nothing more than this, that an agent...must show much consistency in his thought and action, and in this sense have the fundamental values of rationality; yet he may depart from these, his own, norms" (ibid. p.197).

However, as Davidson acknowledges, construing irrationality as a failure to adhere to one's own standards leads to the following question: *"why must* inconsistency *be considered irrational? (Alternatively, or perhaps equivalently, one could ask: who is to decide what consistency demands?) Isn't this just one more evaluative judgment, and one that an agent might reject?"* (ibid. p.194, emphasis in original). In other words, do

rational norms such as consistency exert any necessary normative demand or are they merely contingent values that one might or might not adopt? The norms of rationality are not relevantly similar to contingent values, because breaching them exemplifies what Davidson terms *"objective irrationality"* (ibid. p.189): objective because, on his view, all agents have these fundamental rational values.

The compulsion to recover from error is crucial for reconciling the apparent dissonance within rational interpretationism, and appealing to the RR ostensibly resolves this problem: whilst inconsistencies in belief possession and attribution are possible, if made salient then the agent's awareness that he has gone against his own norms will provide motivation to remove the tension in his belief system. If an agent fails to resolve a manifest inconsistency in his beliefs he has violated a fundamental norm of rationality. However, if we examine how the claim that agents ought to be rational (but yet can fail to be) is supposed to function in practice then an intractable problem for rational interpretationism emerges, the source of which, I will suggest, is in the idea that norms of rationality can be codified in principles.

Conformity and Subscription

The rational requirement of recovery from the error of going against one's own principles introduces something akin to the competence/performance distinction previously discussed with regard to psychological accounts of human reasoning. Here descriptive accounts of behaviour are framed around the question of whether or not the person performing a reasoning task successfully conforms to the principle under scrutiny. Defenders of the view that humans are rational have argued that committing a logical fallacy does not entail that a person is fundamentally irrational (Stein, 1996). An agent's performance is not necessarily indicative of his underlying reasoning competence, as there may be numerous reasons for errors that arise. However, once the fallacy is pointed out or explained, we would expect him to acknowledge he had committed an error, thus demonstrating that despite the evidence of his performance,

his reasoning competence was intact. The disposition of agents to recover from inconsistency indicates that the occurrence of irrational 'performance' errors does not undermine the claim that agents possess a necessary underlying rational competence, conceived by the Rationality Requirement as a commitment to fundamental norms of rationality. Without the normative conception of competence entailed by the claim that we *ought* to be rational it is difficult to explain why agents are in fact motivated to recover from errors in their reasoning and beliefs (given the constraints of time and cognitive resources). Bortolotti (2003) casts this distinction in terms of an agent's conformity to standards of rationality versus his subscription to them. Rather than rely on the idea that individuals' intentional behaviour does in fact exhibit conformity to standards of rationality, we can argue instead that they subscribe to standards of rationality but may on occasion behaviourally deviate from them. This gives substance to the idea that rationality is a normative standard, since it is possible to fail to act as one rationally ought.

Davidson states that the question of whether or not an agent subscribes to principles of rationality *"is not an empirical question"* (1985a, p.196), but rather a condition of possibility for ascribing intentional states to a creature:

> "These are principles shared by all creatures that have propositional attitudes or act intentionally…it comes to no more than this, that it is a condition of having thoughts, judgments, and intentions that the basic standards of rationality have application" (ibid. p.195).

Subscription is, on his view, an *a priori* requirement of intentional agency. An agent could not explicitly violate a norm of rationality or fail to recover from the error once it is made obvious to him and still be ascribed intentional status by an interpreter (Davidson, 1982). Bortolotti (2003, p.118) observes that Davidson's account of subscription is never clearly articulated but rather hinted at in numerous places throughout his writings (especially 1986b; 1985a). When referring to the fundamental constraints of rationality Davidson suggests that as intentional creatures we are *"largely in accord"* (1985, p.196) with such principles, but being in accord does not entail

conforming to them in all cases. Nor for that matter does being in accord with a principle entail one is aiming to be in accord with it. It is not necessary (and indeed unlikely) that these rational standards are available to conscious awareness and the agent need not explicitly endorse them as constraining his processes of belief formation and revision of his intentional actions. The notion of subscription is perhaps misleading, since Davidson acknowledges that one need not be able to articulate or even recognise such a constraint: *"I think everyone does subscribe to those principles, where he knows it or not"* (ibid. p.186). The behavioural manifestation of subscription consists only in a speaker not engaging in obvious inconsistencies and in being motivated or disposed to eliminate them when they arise (Bortolotti, 2003, p.118-9). Hence the only evidence for the claim that one subscribes to such standards is the capacity to recover from error: in the case of cognitive dissonance for instance, it is thanks to one's subscription to the norm of consistency that one sees the tension and seeks to remedy it.

Føllesdal (1984, p.316) considers this capacity to be a second order disposition, comparing subscription to rational principles to the grasping of grammatical rules. One intrinsically knows when one has committed a grammatical error even if one is unable to articulate the rules of grammar for one's language. In the same way, he suggests that the disposition to be rational enables one to acknowledge when one's own behaviour is somehow in error, or that one's beliefs are mistaken. This is all that is necessary for motivating one to revise and change the deviant behaviour even if one is not cognisant of the nature of the normative lapse[77]. Hence subscription allows us broadly to observe the principles of rationality without an awareness of the nature of their prescriptions or the fact that one's thoughts and behaviour are rationally constrained.

[77] Frequent or large-scale failure to conform to such standards would, however, put the possibility of attributing intentional states in jeopardy.

In ordinary cases conforming to a norm, that is, behaving in a way deemed appropriate or correct by it, and subscribing to it go hand in hand. For the rational interpretationist the very possibility of interpretation hinges on the notion that a speaker's observable patterns of behaviour enable an interpreter to ascribe intentional states to him, on the basis of the assumption that this behaviour conforms to broad rational requirements. However, distinguishing between conformity and subscription ensures that if particular instances of behaviour, beliefs and utterances are irrational, this does not threaten the assertion that there are normative obligations imposed by a commitment to rationality, on the condition that agents exercise their capacity to eliminate obvious, acknowledged inconsistencies in their beliefs[78]. The distinction thus allows the freedom to permit instances of irrationality in an agent and to recognise them as being deviations from what he ought to do or believe, whilst continuing to assert that his behaviour can be described in intentional terms and is amenable to reason explanation.

Subscription and the Project of Interpretation

Treating subscription to principles of rationality as constitutive of agency and distinguishing this from mere conformity provides a Davidsonian account with the resources to defend against the charge that evidence of reasoning errors and inconsistency in beliefs would undermine the interpretability of an agent. However, employing this strategy has significant implications for his overall interpretive project (Bortolotti, 2003). If what is constitutive of rationality is an agent's capacity to recover from inconsistency (or indeed any other breach of rational standards), a gap opens up between the constraints guiding interpretation and the behavioural evidence upon which interpretation must be based. The rational interpretationist account is intended to provide a strategy for ascribing beliefs on the basis of the evidence supplied by observable behaviour. The very appeal of rational interpretationism is the assertion that an insight into the nature of propositional attitudes can be gained by understanding how we go about interpreting behaviour (Child, 1996a), and that by reflecting on and

[78] Subscription is not necessary for conformity: one's behaviour may indicate conformity to a norm contingently and not be in any sense bound by a normative constraint on future actions.

analysing our practices of interpretation we can specify the conditions of intentional agency (Bortolotti, 2004a, p.371).

In order to accommodate instances of irrationality the Rationality Requirement must be pitched at the level of subscription rather than behavioural conformity. Subscription may be manifested by the speaker's recovery from an obvious inconsistency, but he need not display conformity to these principles in every instance of behaviour (Bortolotti, 2003, p.119). Bortolotti argues this entails that for any particular utterance or action that is being interpreted, the demands of the Rationality Constraint on interpretation need not necessarily apply. As a framework for guiding the interpretation of utterances and behaviour, the RC then falters since the constraint dictating that an interpreter ought to attribute beliefs that are coherent and largely true does not permit attributions that necessarily reflect the structure of the agent's intentional behaviour. The RC cannot therefore be necessary and fundamental to individual interpretive encounters. That is to say, the ascription of beliefs to an agent cannot be underpinned by the assumption that they are rational, as it is plausible that he is not conforming to the demands of rationality in that particular instance.

There are two issues here. Firstly, Davidson's solution for accounting for error in beliefs and reasoning has inadvertently generated a practical issue for his commitment to the claim that interpretation is normatively structured: on the modified view standards of truth and coherence do not necessarily constrain individual instances of interpretation. This is because the interpreter cannot rely on the necessity of a connection between a speaker's behaviour, on the basis of which he ascribes beliefs, and the standards of rationality to which he ought to adhere. It is not, on Bortolotti's reading, therefore clear what role the notion of rationality is playing in the interpretive project. Secondly, the link between a speaker's behaviour and the standards to which he ought to adhere now appears to be one of contingency rather than necessity. In creating this gap the account has driven a wedge between what is behaviourally manifested, and thus

evidence for interpretation, and what is necessary to rationality and hence intentional agency. Bortolotti articulates the worry this distinction between subscription and conformity generates for the Rationality Constraint:

> "The relation between patterns of behaviour and norms could not be weaker than conformity, since it is by hearing a creature's utterance and observing its behaviour that the interpreter *can find* a large background of rationality and legitimately ascribe intentional states and action to that creature" (Bortolotti, 2003, p.119, emphasis added).

If behaviour is supposed to provide evidence in support of the attribution of rationality to an individual, nothing less than complete conformity to principles of rationality will suffice to ensure that rational interpretationism is a true reflection of the structure of the intentional realm. If subscription to principles of rationality is manifest merely as a disposition to conform but there is no substantive requirement for one to do so in individual instances of behaviour, then the utility of the Rationality Constraint as a necessary starting point for interpretation is undermined. On this reading, the demands of rationality do not necessarily guide an interpreter's options in ascribing beliefs based on individual instances of behaviour, and this calls into question the need to suppose that the speaker being interpreted subscribes to principles of rationality in the first place.

Although the ambiguity of Davidson's own wording makes it difficult to ascertain whether this was a view he actually held, I suggest that this understanding of the relation between conformity and subscription is a misconstrual of the nature of the relation between Rationality Requirement and the constraints on third-person interpretation captured by the Rationality Constraint. Bortolotti takes it that if an interpreter fails to find conformity to principles of rationality in an agent's behaviour then he has no justification for asserting that the agent subscribes to these principles. This assumes the relation to be an evidential one: a view exemplified by her claim that in observing behaviour and utterances an interpreter "can find" a background of rationality. On this view it follows that if the interpreter does not find rationality in an agent's beliefs, desires and actions, he cannot interpret his behaviour and cannot

thereby treat him as an intentional agent at all. An analogy will serve to illustrate the error in conceiving of the relation in this way. Let us imagine that interpreting behaviour is, in the relevant respects, like watching a game of football without any prior knowledge of the rules. If we take the purpose of the game to be that of winning, achieved through scoring more goals by kicking the ball into the opposing team's net, this makes sense of and explains many of the moves that the players make in kicking the ball up the pitch, tackling and so forth: in making these moves, players are doing what they ought to do (whether or not they do actually score). However, in particular instances behaviour might not conform to the rules of the game, say in giving away a free kick, going offside, knocking the ball onto the sidelines and so forth. On Bortolotti's construal of the relation between conformity and subscription, such instances threaten to undermine the idea that the players are aiming to score goals and win the game.

If an observer knows that the players are playing a game, the degree to which they conform to the rules of the game does not play a justificatory role for the observer's judgement about the aim of their behaviour. Subscription to the rules is presupposed by the idea that the players are engaged in a game of football. This is not a claim that is either supported or undermined by particular instances of behavioural evidence, unless the behaviour is such a widespread radical departure from the rules that it is no longer comprehensible as the same game (if the footballers picked up the ball, formed rucks, touched the ball down behind the back line and intentionally kicked it over the crossbar, for example). If players commit errors this does not undermine the observer's commitment to the idea that they are playing football and hence that they subscribe to the rules constituting that game. In fact, it is only because we conceive of such behaviour *as* a failure to behave in such a way as to satisfy the aim of the game that it can be understood as an error, an accidental handball for instance. By the same token then, evidence that agents fail to conform to principles of rationality does not vitiate the

claim that agents ought to be broadly rational in their beliefs because the claim for subscription is not one that is justified by the available behavioural evidence[79].

Bortolotti's insight is not, however, entirely misplaced because it leads us to question what supports the claim of subscription, if not the evidence from behaviour. Recall that the components of rational interpretationism form a holistic circle reflecting the constitutive inter-relatedness of the possession of intentional states and the interpretability of actions and utterances through third-person ascriptions. There are no resources that can serve to justify intentional attributions beyond those available to interpretation. If the claim that agents subscribe to principles of rationality lacks empirical support from the behavioural evidence available to an interpreter, the thesis of rational interpretationism is under threat. However, for Davidson the argument for subscription is a transcendental one: subscription to norms of rationality sets the boundaries of intentionality itself. This is not to deny that behavioural evidence is what drives the project of interpretation. It cannot be known *a priori* what norms of rationality are guiding an agent's behaviour in a particular instance, but the evidence is not used to justify the claim that the agent is constrained by rationality more generally. The claim that we are committed to finding rationality in agents is therefore not a hypothesis or empirical postulate that could turn out to be false. Our primary commitment is to treating and interpreting people as intentional agents and this is not a prescription we can possibly go against: seeking rationality in beliefs and actions as far as possible is not something we could choose to abandon whilst continuing to treat people as agents capable of intentional action and linguistic behaviour.

Failure To Recover from Error

Throughout this chapter I have attempted to immunise the thesis of rational interpretationism against arguments from irrationality and objections that its requirements undermine the attribution of intentionality to perfectly explicable

[79] If behaviour persistently exhibits such violation of the rules that it breaches the bounds of intelligibility, we can question whether it is intentional at all.

behaviour. Rational interpretationism has the resources both to tolerate the existence of sets of beliefs with inconsistent content and to allow such intentional attributions in interpretation. I have emphasised the holistic nature of the normative constraints of rationality, its flexibility in accommodating apparent irrationality and the idea that it is only against a broad background of rationality that behaviour is even identifiable as intentional in the first place. Thus, even though a delusional belief does not cohere well with the totality of the agent's other beliefs and actions, and is a bizarre departure from his ordinary epistemic standards, it can nonetheless be attributed content and enter into reason explanations for a circumscribed set of actions. What marks a delusion out as being irrational is that the delusional agent is not responding appropriately to the bizarre belief, i.e., not seeking to resolve the epistemic mistake once he is made aware of it. In persistently asserting a delusional conviction, the agent is going against his own conception of what he rationally ought to believe and do. This is an important point as it highlights the pre-philosophical puzzlement we face in attempting explain delusions: they are bizarre and baffling. Davidson attempts to cash out this normative requirement on what agents ought to believe and do by positing principles to which intentional agents necessarily subscribe. However, whilst his account allows that we can on occasion fail to conform to such principles, it seems to carry an unfortunate consequence: agents who fail to recover from obvious error, such as the delusional individual, cannot be understood as intentional beings. As empirically robust cases of persistent failure to conform to principles of rationality, the very existence of delusions casts doubt on the transcendental claim that agents subscribe to principles of rationality, forcing us to question why we should be committed to such a claim in the first place.

I consider the implication that persistently irrational individuals have questionable intentional status to be a mistaken step, and I will address why this is the case in the next chapter. I suggest, however, that a rational interpretationist account can deny the claim that failure to recover from error undermines intentional status, without sacrificing

the claim that agents are constitutively rational. Although I have set Bortolotti's account up as a foil to the thesis of rational interpretationism, the conflict between the two positions does, I submit, amount only to a difference in conceptions of what the normativity of rationality consists in. It is only if the demands of rationality are thought of as imposing strict conditions on the attribution of intentional states and propositional content that the possibility of providing reason explanations for all but the most logically watertight actions is in jeopardy. A conception of rationality as highly prescriptive and demanding is familiar in empirical psychology literature on reasoning and, as suggested in previous chapters, has formed the basis of what Stein refers to as the 'standard picture' of rationality (1996, ch.7). There is textual evidence that Davidson subscribed to such a view, for instance advocating an idealised consistency constraint on interpretability and implying that violations of transitivity in preferences would be unintelligible (e.g.,1973a, p.237). This commitment to a strong principled conception of rationality reflects Davidson's Quinean heritage, since the notion of ideal rationality and the necessity of behavioural adherence to logical laws underpin Quine's conception of Charity for the purposes of Radical Translation. Davidson also cites the principles of continence, total evidence, sentential calculus and those derived from Decision Theory as requirements for the possibility of ascribing intentional states (1985a; 1973a).

I agree with Bortolotti that this conception of rationality is indeed far too demanding. It cannot be a condition of possibility for interpreting an agent's behaviour in a given instance that the intentional states ascribed ought to adhere to abstract norms of truth and logical coherence. Most proponents of rational interpretationism concur that if the requirements of rationality are too strong, requiring deductive closure and logical consistency, our theory of interpretation would be *"embarrassed by absurdities"* (Dennett, 1987, p. 94). However, the problem with this view lies not in the thesis of rational interpretationism itself but rather with the claim that the normative content of the thesis is that agents ought to be rational, formulated in terms of subscription to principles of rationality. A principled conception of the demands of rationality

exemplified by the standard picture and the notion of procedural rationality (Bermúdez, 2001), combined with the epistemic standard of true belief, has been predominant in literature on reasoning and decision-making. This narrow view need not, however, be the conception of rationality required by the thesis of rational interpretationism. I suggest that this particular commitment to principles of rationality can be abandoned whilst retaining a claim to the essential normativity of the intentional realm. This view is not without precedent; Cherniak, for instance, conceives of the minimal conditions for rationality as a cluster concept, employed probabilistically (1981, p.175), rather than the rigidly defined set of rules implied by the notion of subscription.

In the next chapter I elaborate on this proposal, arguing that standards of rationality ought not to be construed as a set of abstract principles that can be applied to evaluate whether or not a reason or action is the one the agent ought to commit to in a particular instance. The absence of such principles nonetheless does not threaten the status of rationality as a universal and necessary constraint on interpretation and interpretability. I will argue that judgements of rationality are not characterised by contingent agreements between interpreter and speaker over the truth-value of beliefs, convergence on particular values or over standards of evidence evaluation, but are instead underpinned by the norms that are constituted by and emerge from the social practices we are embedded in as intentional agents.

5. PRINCIPLES & PRACTICE

5.1. NORMATIVE STANDARDS OF RATIONALITY

Reconsidering the Rational 'Ought'

I take as a departure point in this chapter the assumption that there are normative standards that constrain interpretation and are constitutive of intentionality. I wish to address what the status of such norms is, and how they can be characterised. Here I set out the two horns of a dilemma faced by the rational interpretationist. The Rationality Requirement attempts to cash out the normativity of belief with the claim that we subscribe to principles of rationality: codifying the demands of rationality appears to enable prescriptions to be made about what it is rational to intend or think in a given situation, and also to provide explanations as to why certain beliefs or behaviours are rational or irrational. However, I have suggested that it is a mistake to attempt to construe rationality in this way. If rationality is to be conceived in terms of principles to which we subscribe, the question arises as to why we should be obliged to adhere to these principles and what justifications we have for thinking they are constitutive of intentionality. The first aim of this chapter is to seek to explain why the principled view of rationality cannot fulfil the role of specifying the norms of rationality that are essential to intentionality and interpretation. On the other hand, without codification it is unclear whether there is any sense of rationality that universally and necessarily constrains intentional behaviour and disciplines our interpretive practices, since being unable to prescribe correct belief in advance implies the norms of rationality are not force-makers that guide us in how particular intentional states and actions ought to be attributed[80]. If rationality is not strictly prescriptive then it is an open question as to how it exerts any normative constraint at all.

[80] Schroeder (2003) argues that a Davidsonian interpretationist theory is non-normative, because the normative force-maker of 'rationality' is extrinsic to the machinery of the descriptive categorisation of intentional states, and is thus superfluous. It should be clear from the argument developed thus far that I consider this argument to be misplaced: the norms of rationality are not incidental to the categorisation of intentional states but rather intrinsic to their description *as* intentional.

I suggest that this is a false picture of the challenge to rational interpretationism. By considering insights from Wittgenstein's rule-following considerations, I move to undermine the picture of normativity that forced us onto the horns of this dilemma in the first place: the normativity of rationality need not and indeed cannot correctly be thought of in this way. Furthermore, attempting to seek universal principles of rationality is a misguided step, generating the illusion that the normative force of rationality derives from abstract principles themselves. I diagnose the source of this misconception and suggest that committing to rational interpretationism does not entail a commitment to a conception of rationality as a set of context-free principles operating to prescribe what one ought to do or think.

Rationality as a Set of Explanatory Principles

If we conceive of rationality as imposing a normative constraint on intentional behaviour and interpretations of that behaviour, it is natural to think that this constraint could usefully be characterised in terms of principles or rules to which we subscribe. A rule is like a function that, given a relevant set of inputs, identifies one option as being the correct or most appropriate output (Pettit, 1990, p.3). Thus, rules of rationality would specify what one ought to do or believe in light of the particular beliefs and desires one has. Appealing to abstract principles such as that of consistency or of non-contradiction makes it appear as though the demands of rationality can be articulated and imposed as constraints on interpretation that reflect the normative nature of belief:

> "It seems that any theory of belief which is to satisfy the fundamental constraints of having significant empirical content...must include the basic principle that a believer has some, but not ideal, logical ability" (Cherniak, 1981, p.182).

A creature whose belief system and actions adhered to all of these principles would represent the picture of "perfect rationality" (Heal, 2008), taken by many researchers in the field of human reasoning and logic to be the normative ideal to which we, as rational agents, should aspire in our beliefs and actions. Even if we do not always

conform to these demands, our intentional behaviour can be defined and evaluated in reference to them (ibid. p.53).

Whilst this picture of rationality may appear to belong in the domain of abstract logical calculi, Davidson goes some way towards encouraging the view that subscription to these kinds of principles is a condition of intentional agency: *"these are principles shared by all creatures that have propositional attitudes or act intentionally"* (Davidson, 1985a, p.195). Commitment to rational principles may not always be manifested in an agent's behaviour but nonetheless they are principles an agent holds and aims to conform to. Recall that this distinction between actual behaviour and higher-order reflection on one's thoughts and behaviour is the basis for the distinction between conformity with and subscription to principles, which Davidson needs to retain in order to accommodate instances of irrationality within a rational framework. Principles are therefore taken to be explanatorily basic, because the rationality or irrationality of an agent's beliefs or behaviour can be explained in terms of his conformity or failure to conform to these principles. On this view it is *because* an agent ought to be consistent that he seeks to resolve obvious tensions in his beliefs, and generally intends to act consistently with what he believes.

Appealing to principles can generate useful explanations for the irrationality of, for example, *akrasia* in the case of action. It is reasonable to have competing desires and part of the process of forming an intention and executing an action involves weighing these up against each other, but *akratic* action occurs when an agent acts contrary to his own best wishes, behaving, as it were, in spite of himself. What makes this action irrational? It is capable of being explained in intentional terms, perhaps by reference to conflicting desires the agent possesses. But if we appeal to the principle of continence which prescribes that an agent ought to perform the action that he does, all things considered, judge best, the irrationality becomes manifest: the agent has violated his own standard in failing to act according to his best judgement (Davidson, 1985a, p.193;

1982, p.174). I will argue below that understanding rationality in this way rests on a flawed assumption about the nature of the normative commitment entailed by being a rational agent. However, even taken at face value, any account of interpretation that is underpinned by an assumption that agents are broadly rational runs into particular difficulties if it is based on a principled conception of what rationality demands.

What do Principles Demand?

It is worth attempting to establish what the relevant principles of rationality are thought to be, and how they are supposed to function to impose constraints on interpretation and intentional ascription. I have already mentioned the conjunction rule, which follows from the extension rule of probability. Davidson himself refers to the principle of consistency, which derives from the law of non-contradiction, throughout his writing on rationality. He also considers the principles of Decision Theory to be necessary and adds *"the basic principles of logic, the principle of total evidence for inductive reasoning, or the analogous principle of continence"* (1985a, p.189) to the non-exhaustive but demonstrative list.

Compiling such a list of principles does not, however, help clarify what demands are imposed upon us as agents. For a start, it is not obvious precisely which principles of rationality can or should be applied in a given situation. Decision Theory relies upon the assumption that one may possess degrees of belief, ranging from confident assertion to scepticism or suspension of judgement; hence this perhaps implies that the canons of rationality should be expanded to include Bayesian notions about the degrees of probability assigned to beliefs. Furthermore, given that choice preferences are rarely made in isolation but instead might be strategic with respect to other individuals, perhaps the complex canons of game theory ought also to be included in the specification. Heal (2008) suggests there are numerous principles of logic that could be incorporated into the demands of rationality in addition to propositional and predicate calculi. But if such principles as these are accepted, we see the beginnings of a

proliferation: a vast array of canons of rationality could be argued for, without any clear distinction between what is necessary for intentional ascription and explanation and what is not.

A further problem for the view that takes rationality to be codified in principles that can be straightforwardly applied in interpretation concerns the way that logical functions map transitions from a given set of premises to a conclusion. I have talked loosely of an agent's set of beliefs and desires, considering them as determinable starting premises upon which the dictates of rationality operate. If these dictates are to have traction on our behaviour, a determinate specification is required of what is referred to by the phrase 'the set of an agent's beliefs and desires' in a particular instance (Heal, ibid.). However, I query whether it is possible to circumscribe this set of beliefs and desires for the purposes of seeking to establish what an agent rationally ought to believe or do. I will argue below that we are motivated to think that such a notion is meaningful by a mistaken view of the nature of our psychological makeup. For now, I wish to explore the implications this vaguely characterised obstacle has for a principled conception of rationality. Even if we could specify a set of starting premises there is no way of determining in advance what other factors will be relevant to the processes of forming and revising one's beliefs and intentions: the concerns we bring to a decision-making process do not form a complete and closed system (Wiggins, 1975). Whatever the constraints of rationality are, they cannot be specified in a way that allows us to ascertain what they demand in a given situation: *"the broad notion of rational coherence…does not seem to admit of precise conditions of application"* (McLaughlin, 1985, p.356).

Thus the question of what one rationally ought to do or think in a particular set of circumstances cannot be settled by appeal to principles of rationality (e.g., Child, 1993, p.219). It is commonly presumed (see Stein, 1996) that any characterisation of the normative demands of rationality would have to provide a specific prescription about

what an agent rationally ought to do or believe on a particular occasion: a view of rationality as prescribing correct belief or action that is prevalent in the philosophical and psychological literature. I have already rehearsed the arguments for Bortolotti's claim that rationality, conceived of as this kind of prescription on what one ought to do or think, imposes an implausibly strong demand on interpretation and intentional explanation. Whilst I agree that codification in terms of principles fails, I do not consider that such codification is necessary in order to ascertain whether a given decision, belief or action is the one that ought to have been made, held or performed. In other words, codified principles are not what underpin the claim that there are normative constraints operating on intentional behaviour and interpretation. I will return to this question below when considering the nature of the normativity of rationality and the kinds of obligations entailed by agents in virtue of being intentional creatures.

The Uncodifiability of Rationality

In highly circumscribed domains such as in experimental tests of rational choice theory (developed from decision-theoretic approaches), the psychological variables implicated in decision-making or probability judgement are artificially controlled (Elster, 1984). In such cases it is clear what conformity to canons of rationality entails, for example through expressing transitive preferences or obeying the conjunction rule. However, in the context of ordinary interpretation, belief-desire pairs do not form closed systems devoid of relations to a whole network of other relevant psychological elements: a sentiment echoed by Heal's argument that *"our thinking and desiring life does not go on in a form which allows the demands of deductive logic, decision theory and so on to get a direct and unproblematic grip on it"* (2008, p.56). It may be that such principles enable intentional behaviour to be explained or accounted for in certain circumstances: there is no doubt that, for instance, the principle of consistency captures an important inductive generalisation about our beliefs and behaviour. But an appeal to principles of rationality cannot prescribe in advance what an agent rationally ought to do or think: *"there are no a priori criteria of rationality"* (Malpas, 1992, p.81).

These considerations gesture towards the view that the demands of rationality are uncodifiable[81]. This is the claim that it is not possible to derive a prescription of what it would be rational to do or think in a given set of circumstances from an appeal to principles of rationality. McDowell makes the parallel case in the domain of morality for denying the requirements of virtue are susceptible to codification:

> "[T]he best generalizations for how one should behave hold only for the most part. If one attempted to reduce one's conception of what virtue required to a set of rules, then…cases would inevitably turn up in which a mechanical application of the rules would strike one as wrong…one's mind on the matter was not susceptible to capture in any universal formula" (McDowell, 1979, p.336).

If we wish to remain committed to the rational interpretationist thesis that the demands rationality do act as a necessary and constitutive constraint on interpretation, then the uncodifiability of rationality translates into the claim that there can be no principled determination of what intentional states an interpreter ought to attribute to an agent in order to make his behaviour and utterances intelligible (Child, 1993). In the language of the earlier discussion of Davidson's theory of interpretation, it looks as though there can be no specification of how Charity should be applied in practice.

What is interesting about the uncodifiability thesis for rationality is that it does not appear to be susceptible to positive proof[82]. Rather, as a thesis about the constraints on interpretation and intentionality it is supported by negative claims that codification cannot provide an account of what rationality demands. In spite of his frequent reference to the necessity of principles of rationality, Davidson acknowledges the futility of attempting to formulate a formal theory of human reasoning in principled terms, citing his failure to do so as a reason for abandoning his career as an experimental

[81] Davidson exploits this claim in his account of Anomalous Monism, arguing that the uncodifiability of rationality entails there can be no psychophysical laws: the norms of rationality *"have no echo in physical theory"* (Davidson, 1973a, p.230). I will assume here that the normativity of rationality does preclude its reduction to principles given in physical-causal language, but it is not necessary for a defence of this position to embrace Davidson's contentious theory of psychophysical relations.

[82] Although see McDowell (1984a) for a view that it might be.

psychologist (Davidson, 1973a, p.232). He thus agrees that what rationality requires is not amenable to complete codification:

> "I have greatly oversimplified by making it seem that there is a definite, and short, list of "basic principles of rationality". There is no such list" (1985a, p.196).
>
> "There are no rules in any strict sense, as opposed to rough maxims and methodological generalizations" (1986a, p.446).
>
> "Rationality is…a normative notion which by its nature resists regimentation in accord with a single public standard" (1985b, p.245).
>
> "In neither case [theoretical or practical rationality] is there a fixed weighting or ordering of the competing considerations, or any definite rule for comparing them" (Child, 1993, p.222).

Child (1993) draws an analogy with theory choice in science to clarify the thesis of uncodifiability, discussed most prominently by Kuhn (1970a). In comparing scientific theories, numerous values are brought to bear on the consideration of which theory it would be best or correct to adopt. A good theory is one that is parsimonious, accurate, consistent, fruitful in its explanations and broad in scope, to name but a few principles. Each of these features takes the form of criteria for assessing theories, the fulfilment of which count in favour of the theory. But judgements about the relative merits of different theories are not a matter of mechanically checking off these criteria, as they make different demands that need to be balanced and weighed up against one another and that may apply differently in different theories.

An analogy from aesthetic judgement helps to clarify this point. Aesthetic principles such as beauty, elegance, simplicity and so forth may all be relevant considerations that enter into a judgement of the aesthetic worth of an object, but this value is not determined by deductively ascertaining how these principles apply to a situation[83]. In making an aesthetic judgement every detail of the situation is relevant, and incapable of being reduced to an evaluation of a mere subset of its features (Child, 1993). There are always numerous *ceteris paribus* clauses that could defeat the prescription of one particular principle in favour of another, hence there can be no account of what, all

[83] Kuhn makes a similar point regarding the role of aesthetic values in forming the basis for comparison between scientific theories (1977).

things considered, the correct judgement derived from these principles could possibly be. We could extract a general principle of aesthetic taste by building in all the characteristics of a particular case but this would simply be a summary of the judgement in that case, incapable of being used to derive prescriptive guidance about each new case. We could not therefore provide the means of codifying general rules about aesthetic judgement or specify in virtue of what something could be deemed to be aesthetically good. Child draws out the analogy with rationality thus: *"principles like that are not the materials for a codification of rationality; they are the results of applying the uncodifiable norms of rationality to a particular case"* (ibid. p.224). Although such analogies are epistemological and therefore not directly comparable to the norms of rationality implicated in rational interpretationism, they lend credence to the idea, to be developed in this chapter, that constraints need not be codified in order to guide judgements normatively.

Uncodifability and Objective Standards

The principled conception of rationality purported to provide abstract, universal standards, conformity with which could be determined in any given situation. One of the key motivations for seeking principles is that considerations of what rationality demands may appear to be less objective unless the canons of rationality are codifiable (Child, ibid. p.221). What implications does an acknowledgement of the uncodifiability of rationality therefore have for the claim that there are objective normative constraints on interpretation and intentionality? If uncodifiability entails that it is not possible to generate a normative prescription specifying what an agent rationally ought to believe or do in a given situations, it is not clear that there is anything that could be called a normative constraint exerting an effect either on intentional behaviour or interpretation. Moreover, if we cannot derive a statement about how one rationally ought to behave, the normative force of rationality loses its impetus. Bortolotti (2004b) exploits this fact to great effect by pointing out an apparent flaw in the interpretationist theory of the conditions of belief ascription: failure to articulate what rationality

demands in a given case means that there is no normative ideal in place against which an agent's behaviour can be judged or evaluated. Lacking any clear specification of what the correct or appropriate behaviour in the given situation would be, how can we say whether a piece of behaviour is rational or irrational?

Following this line of thought, if the norms of rationality are not codifiable and this uncodifiability threatens their objectivity, then whatever standards do govern evaluations of rationality could in principle differ between people, communities and populations, being relative to each particular group. I have suggested that rational interpretationism undermines this conceptual possibility, but I have not yet offered a substantial argument to support this view of the normative structure of intentionality and interpretation. On a relativist conception of the demands of rationality, it is empirically possible that the reasoning processes, conditions governing the formation of beliefs and intentions, the weighing up of evidence and the processes of decision-making could all potentially diverge between different groups of people. In this respect rationality would be akin to a set of values that guide what one would think of as reasonable or appropriate in a given situation. However, I will argue that the plausibility of such relativism about rational standards rests upon a misconception about the normativity of rationality, because there are limits to the intelligible hanging together of beliefs, utterances and actions that render idea of alternative forms of rationality an empty notion.

5.2. RULE-FOLLOWING

Language Use and Belief

To address the question of whether or not normative standards of rationality are objective and to ascertain how they could be characterised, I turn now to a discussion

of Wittgenstein's famous rule-following considerations[84]. In the *Philosophical Investigations* (1953) Wittgenstein queries what it is to mean something by a word and how this is connected with the way it is used by a speaker. In doing so he suggests that using a word correctly is akin to following a rule that specifies the correct conditions of that word's application[85]. The focus of these initial remarks and the vast secondary literature they have spawned, in particular Kripke's (1982) influential treatment, is on linguistic usage and meaning which, although relevant to interpretation, is not directly applicable to the discussion at hand[86]. Nonetheless, here I aim to provide a brief exposition of the problematic regarding linguistic meaning and to consider the parallels between grasping the conditions for the correct application of a rule and the conditions on the attribution of intentional states. I shall extract some general insights about rule-following from the original context in which they were made, in order to apply them to considerations of how rationality might exert a normative constraint on our behaviour and interpretive practices[87].

Much of the discussion about what it is to follow a rule can be brought to bear on the question of whether there is any essential normative obligation entailed by the possession and ascription of intentional states. In elucidating the way in which the meanings of words impose conditions of correctness or appropriateness on linguistic moves, I argue that it is a mistake to consider that the only way a normative constraint can function is via a codification of its demands in terms of rules or principles. This insight carries implications for our understanding of what rationality requires and how it ought to be construed, and in the course of this discussion, I attempt to correct the

[84] It is misleading to suggest that there is a definitive interpretation of Wittgenstein's remarks on this subject. At best, we can point to a set of philosophical problems raised without claiming to provide an exegesis of Wittgenstein's own view (Schulte, 2008).
[85] McDowell and Pettit jointly argue that Wittgenstein should be understood as advocating the normativity of mental content: *"mental activity is undertaken under the aspect of allegiance to norms"* (Kusch, 2006, p.51).
[86] Pettit argues that the capacity to follow rules is not only a condition of possibility for speech, it is also necessary for thought (1990, p.5).
[87] Wittgenstein draws attention to the topic of interpretation in §§205-207 and refers back his earlier remarks on rule-following. He also employs something akin to the notion of a radical interpreter to cast light on meaning and intentionality.

mistaken assumption about the normative force of principles that underpins the apparent forced choice between the codification of rationality and relativism about rational norms. The crucial question here is whether there are non-principled standards of rationality that form an intrinsic condition of possibility for interpretation, or whether interpretation and intentional ascription are contingent only on standards that are constructed and constituted by consensus within a particular community. I will claim that freed of the misconception exposed by the rule-following considerations, the rational interpretationist thesis is in a far stronger position as an account of intentionality and interpretation.

Wittgenstein's Rule-Following Considerations

There are normative constraints on what we ought to say if we are to use a word or deploy a concept appropriately or correctly. In §§138-242 of the *Investigations* Wittgenstein draws attention to a philosophical tension between the way that we use words and our ability to understand their meaning[88]. When we understand words we appear to grasp them *"in a flash"* (§139), as though something is instantly presented to the mind, but at the same time we think of meaning as being determined by use, which is extended in time (§138). In this latter respect, the use one makes of a word is what McGinn refers to as a "criterion" for what one means by it (1997, p.74). On the one hand therefore, grasping linguistic meaning looks like having a kind of mental state but on the other, such understanding seems to consist in an ongoing obligation towards picking out the right conditions of application for the term. The problem Wittgenstein identifies is that of how to reconcile these two intuitive aspects of what it is to mean something by a word: how can one's sudden grasp of the meaning of a term also provide a normative constraint as to how one should go on in the future?[89] What is

[88] All citations of sections (§) will refer to the *Philosophical Investigations* (1953) unless otherwise indicated.

[89] Many commentators (Thornton, 2005; Luntley, 2003, p.115-9; Pettit, 1990; McDowell, 1984b) have taken it that this normative obligation is not merely contingent on a desire to be understood by others, but rather is fundamental to the meaning of the term used. This obligation is therefore made *"on pain of failure to obey the dictates of the meaning we have grasped"* (McDowell, 1984b, p.325), suggesting that meaning itself is intrinsically normative. Critics of the semantic normativity thesis such as Kusch (2006) and Hattiangadi (2006) argue nothing intrinsic

perhaps most puzzling about this philosophically charged difficulty is that we do grasp meanings and know how to apply terms correctly, often without difficulty or even conscious effort[90]. Therefore the question arises as to how this is possible: what kind of entity could satisfy the condition of being both graspable and applying to an indefinitely large set of cases?

Wittgenstein develops this question by considering understanding meaning as an instance of rule-following. It is natural to think of the constraints on using words correctly as rules that guide our actions and utterances, in the sense that it is possible to specify what action, utterance or behaviour would be in accord or fail to be in accord with that rule. There are numerous and detailed technical accounts exploring what a rule consists in (e.g., Pettit, 1990) but for present purposes it will suffice to pick out a couple of features of an intuitive lay concept of a rule that will bear relevance to my focus on the normativity of rationality. Firstly, a rule prescribes conditions of its correct application. When we apply the concept 'green', for example, we can be said to be following a rule regarding the use of the concept: if we apply the concept only to things that are green and not to those that are not green, we have succeeded in using the concept correctly. Similarly if I use the word 'green' there is something I mean by that; I am communicating an intention. McDowell suggests that it is entirely plausible to *"think of meaning…in…contractual terms"* (1984b, p.221), highlighting the sense of obligation one feels when attempting to use a word in the right way. In a semantic context this means that following a rule entails knowing how to apply the term correctly, for example by picking out the correct referent. In a decision-making context rule-following amounts to identifying what options or subset of options it would be correct or

---

to meaning renders it normative. However, the implications of the view of rationality and the constitutive interdependence of belief and meaning I will adopt position my views in the former camp, committing to the thesis that meaning is indeed normative.

[90] We can reflect on Wittgenstein's use of philosophy as a non-revisionary and therapeutic exercise: *"Philosophy…leaves everything as it is"* (§124). In everyday interaction we are confident in our use of words and our ability to understand another's intended communication. It is only philosophical theorising about meaning and rule-following that is being queried here. Indeed, it takes a degree of philosophical training to even find difficulties such as Wittgenstein's puzzlement in §138 and §139 compelling (Kusch, 2006, p.4).

appropriate to choose[91]. This feature of rules is reflected in the principled conception of rationality: obeying the dictates of certain principles entails believing or intending what one ought to in a given situation.

The obligation implicated in the idea of following a rule projects beyond actual examples or instances of its application: it seems to provide *"rails invisibly laid to infinity"* (§218) that show us how to go on in future cases. A rule determines a potentially infinite number of moves made in accord with it. In the case of meaning, the rule provides a measure against which to judge one's linguistic behaviour as correct or incorrect. However, the extension of these correctness conditions may not be specifiable in advance of a particular instance of application[92]. For example, it is not possible to specify the complete set of circumstances in which the normal use of a particular word such as 'chair' may apply. Certain rules may be circumscribed to a specific situation such as within the context of a game[93]. Even here however, for example in a game of chess, whilst the number of possible moves that satisfy the rule for a rook is potentially calculable, it is in effect indefinite for the purposes of human comprehension. The rule therefore acts as an abstract function that operates on certain inputs such as a set of premises or a particular word or concept, to produce as an output a specification of the correct or appropriate decision, or application of the term or concept. Referring to this as the "rule-in-intension", Pettit (1990, p.3) suggests that this concept of a rule is more pre-philosophically familiar to us as a graspable abstract object: a concept, property or universal that somehow identifies conditions of its own correct application.

---

[91] "Appropriateness" may be constituted by norms of etiquette, prudence, pragmatism, justice or any other normative standard to which behaviour could be said to conform, such that the constraint *"should* tell *me what I ought to do"* (Kripke, 1982, p.24).

[92] Pettit refers to the exhaustive set of correctness conditions as the "rule-in-extension" (1990, p.3). It would obviously not be possible to grasp any such infinitely large set identifying the meaning of a word.

[93] The analogy with games should not be taken literally as it implies that rules of language are fixed and definite. In §81 Wittgenstein suggests that the source of misunderstanding about language and meaning that he goes on to consider derives from thinking of using a language in terms of playing a game operating with a calculus according to definite rules that form an 'ideal' standard. His view of language as a calculus was prominent in the *Tractatus*, but here he is suggesting that such a view is mistaken.

This sense of indefinite applicability can be drawn out by considering the case of the stubborn pupil being taught to continue a mathematical series, which Wittgenstein revisits in §185 and elsewhere (e.g., *Remarks on the Foundations of Mathematics*, 1956, hereafter *RFM*). Having been given the instruction to "add 2", the correct application of this rule requires expanding the series by adding 2 at each step (0,2,4,6,8…n). In this sense, to say one is following the rule is to say that one is committed to a certain pattern of continuation (Wright, 1980, p.21). Obeying the rule seems to be a matter of tracing out what is, in a sense, already there (Bloor, 1973, p.181). Understanding the rule requires grasping how one ought to go on, and it is the aim of Wittgenstein's dialectic on rule-following to consider how it is possible to do this and to get the pattern of usage correct[94].

This brings us to the important second feature of the concept of a rule: that it is possible to be in error about its correct application. No matter how certain that one has grasped the rule correctly, this does not provide an epistemic guarantee that one has got its obligations right (Pettit, 1990, p.3). One can be mistaken about the meaning of a word and apply it inappropriately: if one could not fail to follow a rule there is no sense in which one's behaviour could be deemed correct or appropriate. It is therefore only because it is possible to fail to conform to a rule that it makes sense to talk about behaviour and utterances in normative terms. In being part of a linguistic community there are ways in which one is obliged to speak or act in order to be understood, and one's linguistic behaviour may be in accordance with or against certain rules. Violating these rules entails one's utterances may not be understood[95]. There are many rules of language use that are matters of social convention, prudence, pragmatism, justice and

[94] Although Wright suggests we do this by a process of "cottoning on" (Wright, 1980, p.216) to the pattern a teacher is attempting to convey, Wittgenstein seems to oppose this as an explanation when his interlocutor (typically set up as articulating a position to which Wittgenstein is opposed) suggests this understanding is a matter of guessing the essential drift of explanations (§210).

[95] One may of course violate these rules intentionally in order to deceive, to provide a humorous context, to be insincere, etc. The point is that in such cases it is only because we are using words incorrectly that these intentions can succeed.

so forth, but what the rule-following considerations expose is that grasping the meanings of words can itself entail some normative obligation towards their correct use.

The most forceful objection to the claim that we can and do follow rules in using a term correctly comes from Kripke's (1982) influential sceptical challenge, which has framed a substantial debate around the normativity of meaning. I do not wish to provide an exegesis of Kripke's interpretation of Wittgenstein's remarks nor of his attempts to recover the notion of meaning from a position of scepticism. Rather, I am concerned here to use the challenge Kripke poses to shed light on the nature of the normative obligations entailed by using a term and on how one is justified in claiming that one remains faithful to its correct use. The challenge elaborated by Kripke aims to undermine the pre-philosophically intuitive fact that one can mean something by a word one uses. He asks what kind of fact could constitute a person's following a rule (ibid. p.11). In line with Pettit (1990) I will shift the emphasis from facts about a person to a consideration of what kind of thing could constitute a rule, given that my focus here is on the nature of normative constraints of meaning and how these could be characterised. Taking the usage of the addition function '+' to be typical of a rule, Kripke queries whether there is anything about one's past behaviour and intentions to use '+' when performing mathematical additions that determines the correct answer, the answer that one ought to give, to a novel addition calculation. What makes it the case that by using the '+' function one means to perform an addition, such that the correct response to the question '68 + 57 =?' is the value '125' (Kripke, 1982, p.8)? Presuming that one has never performed that particular calculation in the past, how can one tell that what one is doing now (using the '+' function) accords with what one was doing in the past?

Wittgenstein points out that trying to determine how a rule exerts a constraint on one's behaviour is a matter of logical as opposed to causal determination: *"How am I able to*

obey a rule?"- if this is not a question about causes, then it is about the justification for my following the rule in the way I do" (§217). Any account of one's behaviour that cites one's dispositions, perhaps inculcated through education, cannot account for why a certain response is correct and others are incorrect (Kripke, ibid. p.24). Appealing to the fact that one intends to go on in the same way as one has been disposed to in the past also begs the question against what 'same' means in this context, given that it is a novel calculation[96].

The sceptical conclusion Kripke derives from this argument is pitched at both epistemological and metaphysical levels[97]. The epistemological issue concerns how one can be certain or confident about the answers one gives when one is purporting to use the addition function (ibid. p.8; Kusch, 2006, p.14). Kripke argues that there are no facts about one's previous actions, intentions or utterances that one can appeal to in order to justify the claim that one is following a particular rule: *"…it seems that no matter what is in my mind at a given time, I am free in the future to interpret it in different ways"* (Kripke, ibid. p.107). No matter how consistent one's use of the '+' function has been in the past, there is no epistemic guarantee that one is following the rule of addition as opposed to some deviant function. Crucially however, there is a metaphysical consequence to this sceptical argument: any cited fact cannot possibly constitute my meaning something by the function '+' such that it could guide or dictate my future use of the term. No fact constitutes my having attached one meaning to this term '+' rather than another (McDowell, 1984b, p.329) and if there is no fact of the matter at stake the sentence does not express a proposition: it is not truth conditional[98].

[96] Kripke presses this point by suggesting that one may actually be following a different function of 'quus', which deviates from the plus function for (arbitrarily) numbers greater than 57, thus issuing the challenge to explain in virtue of what one can claim to be using 'plus' rather than 'quus' or any other deviant function in performing the calculation.

[97] Kripke argues for scepticism about meaning based on epistemological grounds, somewhat like Quine's argument for the indeterminacy of meaning (Miller, 1998, p.154).

[98] Commentators such as Dummett have argued that Wittgenstein does indeed reject a truth-conditional conception of meaning in the *Investigations*, marking a significant shift away from his previous views in the *Tractatus* (Dummett, 1959).

Therefore, the sceptical argument concludes *"[t]here can be no such thing as meaning anything by any word"* (Kripke, ibid. p.55).

Although it is beyond the scope of this discussion to examine the implications of Kripke's sceptical conclusion, the form of his argument is relevant to our current concerns. In attempting to specify what a normative constraint on one's linguistic behaviour might look like, Kripke searches for a fact about a person that could constitute his obligation to go on in a certain way. Bearing in mind that what is at stake is the determination of the very meanings of words, a speaker cannot justify why he goes on in one way rather than another (using 'plus' rather than 'quus') by adverting to his intention to 'add', or to say that he is performing a 'counting' function, since both appeals are subject to the original sceptical attack. There is nothing to justify the claim that one is performing the functions of 'adding' or 'counting' rather than some other similar-looking function (ibid. p.21). Without a way to specify what the correct interpretation of these words is that is independent of a rule prescribing their application (and hence invulnerable to a sceptical challenge as to what they mean), one's search for justification falls prey to a vicious regress of interpretations. Whatever *"mental furniture"*, to use McDowell's phrase (1998, p.226), one cites to claim that one is using '+' correctly, the sceptic can point to other consistent interpretations of one's pattern of use, thereby requiring one to appeal to a further interpretation to fix what one understood the rule to entail. If a rule stands like a signpost (§85) indicating the direction in which one ought to go, understanding the direction itself requires an interpretation to explain how one ought to read it: *"any interpretation hangs in the air along with what it interprets, and cannot give it any support"* (§198).

If we are to follow the sceptical line of reasoning, the only way the regress could be halted would be by an interpretation that ensures the connection between the instruction given by a rule and its being successfully applied in practice. Wittgenstein refers to the kind of entity that could fulfil this role as a "superlative fact" (§192), which

is a *"self-standing source of significance"* (Finkelstein, 2000, p.54), and he acknowledges that we have no model for what kind of thing this might be. McDowell considers that an account resting on the notion of a "super-rigid" self-interpreting abstract entity makes our ability to adhere to rules look mysterious and supernatural. Succumbing to this "rampant platonism" (1994, p.92), we cannot account for how the normative structure of meaning impacts on use and constrains our linguistic behaviour, since there is no way of specifying how such an entity, laying down rails in a Platonic heaven (McGinn, 1997, p.107), could engage with our finite minds. Furthermore, on a sceptical reading, the connection between a rule and its application cannot be grounded in the idea that some normative compulsion takes hold of us. Supplying conditions in virtue of which a term is correctly applied does not fulfil the role of showing us how we ought to use that term in practice: stating the rule does not itself do any normative work. Any attempts to locate this normativity through seeking to justify the connection between the rule and its correct use by appeal to a further interpretation lead to a regress that leaves our capacities to follow rules and conform to their normative prescriptions looking mysterious (Wright, 2002, p.151). Thus the epistemology of a Platonic conception of rule-following (and by implication, of meaning, logic and mathematics, which are prime candidates for being rule-governed) is circular (Bloor, 1973).

Without a model for this kind of fact, on Kripke's account we are forced into scepticism about the notion of meaning something by a word. The implication for our present concerns is this: if a rule does not itself compel its correct conditions of application, then there appears to be nothing constraining our use of the rule at all: *"if everything can be made out to accord with the rule, then it can also be made out to conflict with it. And so there would be neither accord nor conflict here"* (§201). For Kripke this entails that any normative standard that does guide or constrain the correct use of a term is not something that is intrinsic to using that word. Whilst there are myriad implications of the nuanced and complex view of meaning following from this claim, his resultant

account contains the central idea that the correct use of a term is determined only by a kind of communal agreement within a linguistic community[99]. The justification conditions for saying that someone has used a word correctly or uttered a meaningful sentence are a matter of the acceptance by and consensus with others, because such conditions cannot be intrinsic to an individual's grasp of the word's meaning.

In one respect accounts of meaning derived from a sceptical standpoint do identify something intuitively correct about language use: that the meanings of words are flexible and subject to change over time depending on how they are used by members of a community. However, scepticism about meaning inevitably opens the door to the kind of relativist worry mentioned previously: if the rules determining the correct use of words have no objective justification, there is no obstacle to the idea that different communities might have no mutual ground or common co-ordinate system of reference from which to interpret the utterances of the speaker of an unfamiliar language.

5.3. RULES AND RATIONALITY

A Mistaken Dilemma

Cast in this light, the dichotomy between the codification of rational standards of belief and the suggestion that they are relative to a community can be given philosophical substance by understanding it as an instance of the apparent dilemma about linguistic meaning raised by the rule-following considerations. Whilst Wittgenstein asks after what it is to grasp the meaning of a word and use it correctly, I am here interested in establishing what it is to possess a belief and know what follows from it. In seeking to ascertain what it is to use a word correctly according to its meaning, Wittgenstein invokes the notion that one follows a rule specifying its correct application. But in attempting to explain the connection between the rule and a prescription of these

[99] There are numerous variations of this idea. Kripke retains scepticism about meaning while salvaging correctness conditions for a term's application from the idea of communal assent, while Wright (1984) attempts to recover the notion of meaning itself by arguing that it is constituted by shared agreement.

conditions, there is no foundational justification to be found and, on a Kripkean reading, an infinite regress of interpretations looms. This results in a forced dilemma between mysterious rampant platonism and a Kripkean scepticism that there is anything at all fixing the correct conditions of application for the rule. There are clear parallels here with the problem of characterising what rationality demands if it is construed in terms of subscription to rational principles. If we take it that to be rational is a matter of subscribing to principles of rationality, then the challenge lies in explaining how subscription connects the principle with moves made in accord with it, the moves that would enable the beliefs attributed to the agent to be deemed rational. Take, for instance, the claim that one subscribes to the principle of consistency. The principle can be described as a rule, prescribing conditions under which it would be correctly applied in a given situation. What connects the rule (consistency) with the conditions of its correct application? How would we know that a set of beliefs conform to this rule and how would we know it had been applied correctly[100]?

Drawing on the analogy with Wittgenstein's insights into rules of linguistic meaning, any explanation of this connection seems to require a further interpretation to support it: if we appeal to the idea that by 'consistency' we mean having beliefs that do not logically contradict, a sceptical argument would seize upon this interpretation and query what justifies our intended application of this concept of 'contradiction', and so on. Unless we can posit some self-interpreting Platonic "superlative fact" to explain the connection between the principles of rationality and how they are correctly applied, the normative obligations they place on our beliefs, behaviour and utterances seem mysterious and inexplicable[101]: *"in rampant Platonism, the rational structure within which meaning*

[100] Wright develops a strong conventionalism in response to this question with respect to mathematical inferences, arguing that one's present judgements that one sincerely meant the same thing ('plus') in the past determine that one did in fact mean the same thing: one's present judgements are stipulations, and these fill the role played by interpretations for Kripke (Wright, 1980). As shall become clear, I consider this move to be a sceptical solution to a dilemma that need not arise in the first place

[101] In commenting on McDowell's articulation of the Platonist supposition, Charles Taylor argues that such a view paves the way for much of our thinking about the mind even now, citing

comes into view is independent of anything merely human, so that the capacity of our minds to resonate to it looks occult or magical" (McDowell, 1994, p.92).

If the sceptical reading of rule-following is right, attempts to understand rational interpretationism in terms of subscription to rational principles are doomed to failure, either leaving the normative constraints exerted by those principles unexplained, or conceding that there is nothing necessary about adherence to such principles. This brings us to the alternative conception of normative standards on the sceptical view. If the rule or principle itself cannot compel a particular (correct) application, our use of it appears unconstrained and there is no intrinsic normative prescription governing what it is to act in accordance with it. This implies that whatever standards are in play when we interpret an agent, they are not disciplined by any universally shared criteria. Just as communitarian responses to a Kripkean reading of Wittgenstein (e.g., Williams, 1991; Wright, 1984) cite the consensus of a linguistic community as providing the standard of correctness by which word use should be judged, so too would standards of rationality be contingent on the shared agreement of a community.

Recall that the relativist concern mooted the possibility of radically different standards of rationality: that what counts as a reason might be idiosyncratic to one individual or group and not intelligible to others as being a reason. There might be differences between the normative structure of intentional behaviour and language of an agent and an interpreter, in which case the interpreter would not be in a position to make normative judgements about the agent's behaviour. The motivation for adopting relativism about rational standards arises if one takes seriously the implications of the kind of scepticism developed by Kripke about meaning. Without any justification for the connection between holding a particular belief and the behaviour that ought to follow from this belief, it appears that whatever standard there might be is one imposed only

---

cognitive psychology as resting on both ontological and methodological assumptions that the mind is a mechanical part of nature constituted by atomistic elements (Taylor, 2002, p.110).

by a contingent consensus with others. If this is the correct way to understand interpretation, then it does indeed invite the question as to what guarantees that the norms of the interpreter match up with the constitutive norms governing the agent's own processes of belief formation and possession. There is thus a potential for radical divergence between interpreter and agent in standards that govern the application of intentional concepts. This entails that there may simply be different ways of going on in the world, which are not susceptible to normative judgement from outside that particular community. By this I mean that there may be such fundamentally different ways of going on, of individuating objects and events, using words and speaking a language, and of relating beliefs to actions, utterances and other intentional states, that there is no possibility that one set or group of people could form judgements about the intentional behaviour and utterances of agents outside of their own linguistic community. On this view, normativity is extrinsic to the concepts of meaning and belief, imposed only by the contingencies of shared agreement. Thus a community could be thought to possess language and exhibit intentional behaviour, but the standards governing the correct or appropriate moves are particular to that community.

We therefore appear to be forced onto the second horn of the sceptical dilemma in attempting to explain how normative standards exert a constraint on intentional behaviour and interpretation. In denying that interpretation is underpinned by a necessary assumption of subscription to objective, codified principles or standards of rationality, we seem to have no choice but to accept that what normatively constrains the application of intentional concepts is contingent upon the agreed standards of a particular community. Furthermore, two potential intractable problems arise for interpretation, depending on the extent of the differences between communities. If, as is empirically plausible on this view, two communities differ radically in their systems of concepts and the kinds of connections that obtain between their utterances and actions, their ways of going on in the world will be incommensurable, and the behaviour and language used by members of one community would be unintelligible to the

other[102]. If, on the other hand, there is some overlap between the agreed judgements that constitute the communities' rational standards, the behaviour of each may be intelligible as being intentional, but nonetheless problematic for interpretation. Interpretation of individuals outside one's own community would necessarily involve imposing one's own standards onto behaviour that may in fact be normatively governed by very different constraints. Such an outcome clearly has implications for the possibilities of reliably and accurately attributing intentional states and interpreting utterances without cultural bias or rational prejudice about what it would be correct or appropriate to say, believe, intend or do.

The Master Thesis

We do not, however, need to embark upon the route of scepticism about meaning, and the diagnosis of misconception here bears relevance to the particular view of mentality and the constitution of mental states that I have been advocating. McDowell argues that the motivation behind thinking that we are forced onto the horns of the dilemma between Platonism and scepticism in the first place derives from a Cartesian dogma that the mind is populated with items that stand like sign-posts without need of interpretation (McDowell, 1998, p.264). If the identity of mental states does not depend on their normative relatedness to the world and to each other, it is mysterious how they could possess intentionality at all, and indeed it is this vulnerability that the sceptic exploits in formulating the regress of interpretations.

Beliefs, desires and so forth impose a standard by which the world can be judged: a wish, for instance, is something that certain states of affairs in the world would satisfy, and its content can be identified by describing the conditions under which the wish

[102] A previously mentioned analogy in the philosophy of science helps clarify the more radical implications of this view. If there is no common, theory-neutral vocabulary and system of reference through which to co-ordinate the comparison of successive scientific theories then incommensurability between the theories results (Kuhn, 1970b, p.267). Without such a vocabulary we have no way of grounding or co-ordinating different schemes. This appears to lend support to the conclusion that there may be radically different and incommensurable ways of conceptualising and going on in the world, which cannot be normatively judged or even perceived as intelligible from outside.

would be fulfilled. McDowell argues that this intentionality of mental states cannot be explained or accounted for if there is nothing intrinsically world-involving about them, because if these states are only contingently related to the world this connection would need to be made via an interpretation, thus precipitating the sceptical regress. A conception of the mental realm as consisting in free-standing entities makes their intentionality mysterious, and requiring of a substantial philosophical theory to bridge the gap: *"within the Cartesian picture there is a serious question about how it can be that experience…is not blank or blind"* (McDowell, 1986, p.152). The close parallel with Wittgenstein's charting of the apparent tension between a rule and its conditions of application is clear: such mental entities are akin to sign-posts that stand in need of interpretation (Thornton, 2005, p.41). But this puzzlement is an artefact of a self-inflicted philosophical anxiety (Horwich, 2005) through which we are tempted, illegitimately, to postulate some kind of inner mental mechanism to underpin our ability to practically grasp a rule (Lear & Stroud, 1984, p.226). McDowell attributes the temptation to conceive of mentality in this way to the success of explanation in the natural sciences and the superficially plausible attempt to incorporate our understanding of psychology within a naturalistic framework, using the vocabulary and concepts of the physical-causal sciences. He refers to this conception of the mind as the "master thesis", and its rejection underpins his broader project of attempting to ease philosophical anxieties about the world-directedness of empirical content[103] (passim., but especially McDowell, 1994).

Rejecting the master thesis points the way towards resolving the apparent tension between grasping a rule and correctly applying it, which finds its clearest expression in Wittgenstein's assertion that *"there is a way of grasping a rule that is* not *an* interpretation*, but which is exhibited in what we call "obeying the rule" and "going*

[103] This project forms the central aim of McDowell's *Mind and World* (1994), in which he argues for a "partially re-enchanted" conception of nature that non-reductively incorporates normative relations. Whilst it is beyond the scope of this discussion to chart the complexities of this view, McDowell's point that despite the intelligibility of rational relations being *sui generis* they can nonetheless form part of a naturalistic framework is well taken.

against it" in actual cases" (§201, emphasis in original). The tension is generated by a philosophical anxiety about the justification we have for interpreting a rule in one 'correct' way rather than another, but in rejecting the idea that the rule can only exert a constraint on behaviour via an interpretation, we avoid being swayed by the sceptical concern. It is only if we think of intentional states as atomistic, free-standing entities that a puzzle arises as to how they could bear normative relations to one another and to the world. If, however, we recognise the essential normative relatedness of intentional states, this need for a bridge between the content-bearing entities and the normative standards to which they conform dissolves.

What implication does this diagnosis of the error motivating the sceptical dilemma have for my primary concern with establishing how normative constraints may operate on our interpretive practices and intentional behaviour? The purpose of this foray into rule-following has been to diagnose misconceptions about the demands of rationality, which create the anxiety that it is only through a set of principles that any objective normative constraint on interpretation and intentionality could be exerted. The negative lessons of the rule-following considerations support the claim that rationality is not a contingent feature of interpretation that can be laid like a grid over intentional behaviour as a heuristic to aid our understanding and explanations of an agent's actions and utterances, but is rather intrinsic to the concepts of intentionality.

Before considering whether any positive account of rationality could be given once we are cured of the sceptical illusion, I wish to examine why a principled notion of rationality might be compelling in the first place. Certainly the appeal to principles has been central to the way rationality has often been articulated, as well as being a focus for attacks on the claim that rationality is necessary for interpretation. Given both the uncodifiability of rationality and the analogous arguments from rule-following I have outlined here, it is perhaps surprising that a principled conception of rationality has held sway as a model of human thought for so long. I now turn to examine the reasons for

this predominance and aim to undermine its tenability by dissolving a misunderstanding about the source of normative force upon which it is based.

The Normative Force of Principles

Conceptualising rationality in terms of principles to which agents subscribe is not entirely without merit. Indeed, powerful insights into human reasoning psychology have only been possible through experimental setups in which tests of conformity to or violation of a principle of rationality have framed the research methodology. I am not seeking to undermine the practical or theoretical utility of capturing generalisations about how we reason intentionally act. I am instead claiming that to take the view that agents are rational *in virtue of* their subscription to such principles is to misunderstand where the normative force of rationality arises. The principled conception of rationality as a necessary constraint on interpretation and interpretability implies that the principles themselves are supposed to provide justification for a determination of the intentional descriptions under which an agent's behaviour falls. But if the insights gleaned from Wittgenstein's examination of rule-following are correct, the principles do not themselves carry any normative force outside the context in which they are used. If construed as abstract force-makers, they cannot serve the explanatory role ascribed to them: either the obligation to conform to them is a mysterious Platonic fact about our psychology, or the standards are just contingent on their adoption by a particular community.

Davidson recognises the challenge that arises from appealing to principles as normatively compelling, citing the fact that such an appeal begs the question against why we ought to conform to them (Davidson, 1985a). This is a question about the justification of the Rationality Requirement that one ought to be rational in one's beliefs, utterances and actions. Lewis Carroll's (1895) tale of Achilles and the tortoise provides a lesson about the appeal to principles of logic for both the justification and explanation

of the correct course of action issuing from the observance of a rule[104]. The example serves to demonstrate why this appeal is invalid and points the way towards an alternative way of conceiving of the demands of rationality that does not rest on the requirement of self-interpreting principles for justification.

The scenario is set up thus: Achilles and the tortoise, characters reminiscent of one of Zeno's paradoxes, are discussing the relations between a set of propositions. Following Carroll, I call these 'A' to denote a universal generalisation, equivalent to a principle; 'B' to denote an antecedent premise; and 'Z' to denote the consequent, which follows from the conjunction of A and B. The inference can be described thus:

A:	'if p → q'	(the generalisation)
B:	'p'	(the antecedent)
Z:	'therefore q'	(the consequent)

The tortoise facetiously queries whether there might exist a person who accepts the truth of A and B but denies Z. Having acknowledged that such a person might exist, the challenge set for Achilles by the tortoise is to *"force [him] logically, to accept Z as true"* (ibid. p.278). Why must it be the case that the consequent is true? Achilles' strategy is to introduce a conditional 'C' specifying the relation between A, B and Z, namely that if A and B are true then Z is true. Having written this down in addition to the other premises, Achilles falls prey to an infinite regress as a further conditional is then needed to ascertain the relation between A, B, C and Z. He attempts to stop this regress by claiming that *"logic would take you by the throat and* force *you to do it!"* (ibid. p.279) but the tortoise insists that another conditional is always needed to force him to accept the truth of Z. From this we are led to draw our own conclusions about the impossibility of meeting the tortoise's challenge.

The tortoise is demanding some justification of how Z follows from A and B, in much the same way that the sceptic demands a justification for how a rule connects with its

[104] Lear's (1982) critique of Dummett's attack on the validity of the law of the excluded middle follows the same general structure as the argument put forth here.

conditions of application. The most usual strategy posited by in answer to Carroll's challenge is to deny that the conditional C should be added as a premise to the argument in order to strengthen it (Thomson, 1960, cited by Stroud, 1979). The parable of Achilles demonstrates not only the futility of appealing to a further conditional connecting the rule to its usage (leading to a regress of interpretations), but also the fact that no additional hypothetical is actually needed in order to augment the argument that Z follows from the conjunction of A and B. Rather, inferring the consequent is just what we should do, without the need for additional inferential apparatus. In a powerful critique of the picture of rationality that appears to drive Achilles' problem, Searle (2001) argues that the very sophistication of our formal syntactic models of reasoning, epitomised by proof-theoretic and computational models, creates an illusion that the logical principles of reasoning are justificatory. He attacks the notion that rationality is a matter of obeying the rules of logic, arguing that the tortoise's challenge serves to demonstrate that logic itself cannot serve a role in justifying our practice. Searle's strategy for fending off the threat of regress is to deny that the universal generalisation (A) plays any role in establishing the validity of the inference from B to Z. Rather, *"the inference is perfectly valid as it stands without any outside help"* (ibid. p.19). Thus if the generalisation connecting '*p*' to '*q'* holds, it is not in virtue of one's acceptance of the rule that if one believes '*p*' that one ought to believe '*q*': the rule does not in any way sanction the inference or guarantee its validity.

The conclusion to be drawn from these remarks is a Kantian one, that the connection between a rule and its application is not one that can take the form of an empirical explanation[105] (Lear & Stroud, 1984, p.227). We can better understand the connection by gaining insight into what we do, not through explanation. This is not a point about the stringency of a principled conception of rationality, but rather an insight into the "grammar" of the concept of a principle (McGinn, 1997, p.105), when considered as

[105] This stands opposed to a sceptical conclusion that the relationship is an inexplicable fiction, analogous to a Humean view of causation.

being akin to a rule in the way it is supposed to provide normative constraints on rational agency. At this stage, our explanations as to why we feel compelled to go on in a certain way come to an end: *"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""* (§217). All we can do is point to our behaviour and acknowledge that the search for further justifications is fruitless[106] (Wittgenstein, 1969, *On Certainty*, §130; §189; §192, hereafter *OC*). Thus in claiming that someone has grasped and successfully applied a rule, *"the rule he has grasped does not explain his activity; his activity gives substance to the claim he has grasped the rule"* (Lear, 1986, p.274). The point is not that we have failed to provide an account of the logical compulsion pressing us towards making certain moves that we consider to be correct or appropriate. Rather, Wittgenstein's insight is that there is no justification to be had in the first place from outside the practice in which those moves are made. Seeking interpretations to explain and justify the connection between a rule and its correct applications is futile because to think that we need to provide such justifications is to take a mis-step in determining how normative obligations operate and in understanding why we go on as we do. Thus even in the paradigmatic instance of logical reasoning, the assertion that Z 'must' follow is not grounded in a superlative fact that grabs us by the throat, and to think that it must is to fall prey to a fundamental misunderstanding about the way in which normative constraints on behaviour operate.

This is a crucial point: far from presenting a kind of semantic nihilism, Wittgenstein aimed to uncover and dismantle a misunderstanding about the way we use words that comes to prominence when we attempt to seek justification for our linguistic practices (see esp. McDowell, 1994). He demonstrates that being led down the path of regress indicates we have gone wrong in our analysis of what the norms governing the correct

[106] Wittgenstein's remarks here refer to the grounds we have for taking certain beliefs to be true about the world and thus address the traditional sceptical problem of justifying one's claims to knowledge about the external world. Nonetheless, the metaphor of hitting bedrock in one's explanations applies equally to the notion of rational standards I am charting here.

use of words must look like. The idea that we are compelled by the force of abstract logical principles to go on in a certain way is a chimera generated from a philosophical misunderstanding of what governs our linguistic interpretive practices (McGinn, 1997, p.105). From this picture emerges the idea that going on correctly is not a matter of conforming or aiming to conform to normative principles or rules.

This critique of the normative force of logical principles extends to the supposed principles of rationality. Whilst there may be rules and maxims that aid decision-making and interpretation, *"rationality is not constituted as a set of rules…the structure of intentional states and the constitutive rules of speech acts already contain constraints of rationality"* (Searle, 2001, p.22). The consequence of adopting the mistaken principled view of rationality is an overly prescriptive, implausibly stringent conception of the necessary conditions on intentionality: the idea that one can be deemed a rational agent only to the extent that one's beliefs and intentional behaviour are what they rationally ought to be. This is the position Bortolotti (2004a; 2004b) attacks, and rightly so. However, once we take seriously the idea that principles are explanatorily impotent this objection can be seen as misplaced. Principles cannot themselves serve a role in justifying why certain moves are correct or appropriate and others incorrect or inappropriate in their given context.

5.4. THE NOTION OF PRACTICE

A Therapeutic Resolution

Once we are disabused of the assumptions of the master thesis, conceptual space is made for a constructive answer to the question of how rationality exerts a normative constraint on intentional behaviour and interpretation, which doesn't rest on the requirement of justification via an interpretation. McDowell (1984b) takes Wittgenstein's remark that *"'obeying a rule' is a practice"* (§202) to constitute the seed of an answer to this question: following a rule is a matter of acting as one has been trained to do

(McDowell, ibid. p.339). Appealing to such training does not provide a brute causal explanation for one's behaviour, but rather provides the requisite normative context in which it makes sense to judge the behaviour as correct or incorrect in light of one's *"initiation into a custom"* (ibid). A word's meaning what is does depends on there being a *"use and custom among us"* (Wittgenstein, RFM, I §63), hence it is an error to strip away this context in search of the foundations of rule-following. This is an important point as it demonstrates Wittgenstein's rejection of the idea that fundamentally normative behaviour could be characterised in non-normative terms. Even at the "bedrock" of explanation, the point at which justifications run out, normative notions such as correctness, accord and rule-following have application, and these notions are essential to language (McDowell, ibid. p.341). The norms that permeate language use cannot be derived from a physical-causal disposition to act in a certain way in response to a rule (§§193-195). Thus, according to McDowell at least, Wittgenstein's mention of customs (§198) practice (§202) and institutions (RFM, VI §31) is an attempt to retain the fundamental normativity of meaning and interpretation.

The idea of social practice occupies a central role in Wittgenstein's account of rule-following and reflects, in Lear's phrase, Wittgenstein's *"anthropological stance"* (1986): his concern with understanding language within the context of its use within a community. The very notion of meaning something by a word is only intelligible if it is used within a community sharing certain practices in the way they go on. These practices form *"an indeterminate and unspoken horizon"* (McGinn, 1997, p.96), inculcation into which allows an agent to master the appropriate use of words as he is educated. Furthermore, in answer to the epistemological question of how we can know that going on in a certain way is correct or appropriate, our linguistic usage is legitimated by our shared mindedness with others[107].

[107] Lear and Stroud (1984) refer to this as the synthetic unity of representations, analogous to the Kantian notion of the analytic unity of apperception.

The concept of a rule is intelligible only within the embedded context of the practice of using it, not as some Platonic ideal telling one how to apply it correctly independently of this surrounding (Finkelstein, 2000). The crucial steps in this characterisation therefore appear to be sociological: we mean certain things by a certain word because this is *"the way we are taught to use it"* (RFM, I §2), which suggests that following a rule correctly is the culmination of a social process of learning (Bloor, 1973, p.184). I have used the term 'practice' previously in describing rational relativism, motivating the concern that if the demands of rationality are framed only by social customs, techniques, institutions and so forth, there are no universal standards underpinning linguistic usage or intentionality. In this respect, appealing to social practices appears at first glance to be an admission that standards of rationality are relative to a community. If the correctness of, for example, particular intentional attributions is fixed only by the contingencies of consensus within a community, there is space for the possibility of there being different ways of going on.

Several passages in the *Investigations* can be read as concessions towards the possibility of relativism about practices. As a simple example, the idea of going on in a different way by reacting to one's training in a manner that is not in accord with one's peers, strikes us as empirically plausible (§185). Given that justifications for saying what constitutes the correct moves have run out, viewed as an empirical claim about how we go on in the world there is nothing to rule out the possibility of being other-minded. This view is perpetuated by Wittgenstein's positing of a hypothetical tribe who appear to behave and speak in ways radically different to our own (§§205-207) but to whom we wish to attribute language and intentional states.

There is nonetheless an essential difference between the kind of relativism that results from communitarian readings of Wittgenstein, developed in response to the sceptical dilemma, and what is suggested by grounding the correct application of a rule in the practices and customs into which one is initiated. The sceptic finds no epistemic

foundation for using a rule in a certain way and therefore goes about reconstructing an ersatz notion of meaning by referring to the negative role of the community in ruling out incorrect applications of a rule (Kripke, 1982), or by characterising standards of correctness in terms of patterns constituted by ongoing communal judgement (Wright, 1984). As McDowell argues, these accounts only generate the illusion of being guided by norms and rules (Thornton, 2005, p.38). By contrast, freed from the grip of the demand for justification via an interpretation, the notion of practice accommodates the constraints operating on behaviour and utterances without attempting to eschew their intrinsic normativity. Grasping a rule is a matter of *""obeying* [it]*"* and *"going against it" in actual cases"* (§201), and what it is to comply with or violate a rule can only be identified against a background of the practice of using the rule. The invocation of the community therefore serves a different role from that required by the sceptic.

The Transcendental Argument

To construe the notion of practice as a contingent sociological construct would be to overlook the transcendental thread that permeates Wittgenstein's discussion and the lessons about the relations between language, thought and the world that emerge from the rule-following considerations. Despite the anthropological orientation towards understanding what it is to use words according to their meaning implied by Wittgenstein's reference to training, practices and customs, there is a non-empirical, non-explanatory insight at work about how we go on in the world[108] (Lear, 1986).

Wittgenstein considers the circumstances under which we would deem an alien tribe to be performing intentional actions, speaking a language, obeying orders and so forth (§206). This thought experiment is akin to Davidson's use of the notion of a Radical Interpreter, but the aim of this exploration is different. Whereas Davidson is concerned

[108] Lear argues that this kind of inquiry deserves to be called "transcendental", as the concern with establishing how rules relate to their instances of application parallels Kant's *a priori* investigation into how concepts apply to objects. The Kantian argument that concepts have no meaning beyond the contexts of the judgements in which they are applied reflects Wittgenstein's own conception of the intelligibility of rules as manifested in their use within a practice (Lear, 1986, p.269).

with uncovering the conditions of possibility for languagehood, Wittgenstein is seeking to make a grammatical observation about our concept of a language; namely that it describes characteristic regularities in the use of signs, such as words, and performing of actions (McGinn, 1997, p.110). Within the activity and practice of using words and attributing intentional states, certain patterns emerge that fix the correct applications of those words, and this is fundamental to the idea of a *"form of life"* (e.g., §241). Wittgenstein thus resists the temptation to mythologize meaning by showing that the distinction between correct and incorrect responses is grounded in the context of the practice in which the response occurs (McGinn, 1997, p.102). The *"feelings of naturalness"* (Lear & Stroud, 1984; Lear, 1982) that accompany our progression in continuing a series in a certain way, or in applying a particular term, reflect the effects of our being trained and initiated into a custom. There is a multiplicity of language-games that make up these patterns and the breadth, complexity and richness of these serve to emphasise that language can only be identified as part of an activity or form of life (Finkelstein, 2000). Furthermore, we are brought into awareness of these connections and the correct ways of going on through our development and education: *"the demands of reason are essentially such that a human upbringing can open a human being's eyes to them"* (McDowell, 1994, p.92).

Wittgenstein's employment of the notion of a "form of life" signifies the way we are minded to go on in the world, reflecting our perceptions of salience, feelings of naturalness, shared epistemic interests and so forth (Lear, 1982, p.385). Any moves to understand what this form of life is and how it is constituted will necessarily be made from within, as we cannot step outside of our experience and concepts to evaluate our mindedness from outside, from a Nagelian "view from nowhere" or a position of Quinean "cosmic exile". We cannot consider our mindedness to be one possibility amongst others as we would not be able to make sense of a form of life that was not in large part similar to our own. The examples of differing tribal practices should therefore not be taken as describing instances of genuine other-mindedness (ibid. p.389). They

are a device to reflectively probe our understanding of our mindedness, to enable us to see both how our own ways of going on are constituted by our interests and practices (Lear, 1986, p.276) and that the possibility of being other-minded lapses into incoherence. Beings that do not exhibit patterns in their vocal articulations and actions that we take to be characteristic of using a language would not be intelligible as intentional agents[109]. Lear suggests that in considering what the notion of "agreement" consists in and querying the possibility of being other-minded we are confronted with a modal duck-rabbit. The way we happen to use words is not guaranteed (by some superlative fact) and seems only to be contingent on a shared form of life with our fellow man: a grasp on the world that is precarious at best. At the same time, however, there is no genuine empirical possibility of going on in a different way, as beyond the boundaries of our mindedness we can say nothing at all: *"however tenuous a fact our being minded as we are may at times appear, it is not a fact that could genuinely have been otherwise"* (Lear, 1982, p.387).

When observing an agent's behaviour and listening to his utterances, we cannot help but see intention and hear meaning. No matter how alien the language or unfamiliar the tribe, we see characteristically human patterns and regularities in the relations between verbal behaviour and action (Hopkins, 2004, p.10) that provide us with a frame of reference through which we can perceive what going on in the right way is, in practice. Furthermore, intentional explanation relies on being able to identify and co-ordinate speech behaviour with an agents' actions (ibid. p.4). Speaking a language involves a kind of order that is essentially found in human behaviour: *"the common behaviour of mankind is the system of reference by means of which we interpret an unknown language"* (§206). Consider our ordinary practices of interpretation, in which we take linguistic utterances to express intentional states, and make sense of an agent's

[109] This assertion is not meant as a rejection of the idea that we can explain and perhaps predict the behaviour of non-linguistic creatures. The point is rather that the concept of language, as a communicative tool, describes a form of life that exhibits the characteristic patterns of activity that constitute the following of rules (McGinn, 1997, p.110).

actions on the basis of such attributions. What allows us to take this interpretive step with any degree of confidence is that this attribution allows us to co-ordinate interpretation of the agent's other actions and utterances that are related to the attributed belief or desire: we have a common system of reference.

Without such regularities, the gestures and vocalisations of a group of speakers would be unintelligible as intentional action and language. Their behaviour would seem confusing and illogical. If their way of going on is so radically different from our own that we cannot observe any patterns of regularity in the connections between utterances and movements, then we cannot apply the concepts of belief and action to them. To use Wittgenstein's own example from early in the *Investigations,* if there was not a regular connection between a tribesperson's utterance of the term 'slab' and his gesturing towards a building-stone of a particular shape (§2), not only would our ability to grasp the meaning of the term falter but we would question whether the sounds being made constitute an (albeit primitive) language (§207). Hence it is only through identifying connections between utterances and actions within the context of an agent's behaviour, and his location within a community, that interpretation is possible at all.

An important aspect of this transcendental insight is borne out in the idea of shared agreement. If part of what it is to use a language is to exhibit patterns and regularities in one's utterances and actions, and for these to be understood as such by others, then there must be a degree of agreement between members of the community as to what constitutes the right way to go on:

> "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" (§241, emphasis in original).
>
> "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 judgements" (§242).

Agreement about judgements form part of the framework through which language can function: if we are to succeed in communicating at all we must agree on certain judgements about the world and the way our words and concepts relate to the world

and to our actions. For instance, agreement about whether or not a rule or order has been correctly followed (§458) in using a term or deploying a concept appropriately is necessary if we are to communicate and understand one another's use of that term or concept. In perceiving movements and noises made as manifestations of intentionally directed behaviour and speech we find regularities and patterns of activity, which emerge when we examine a whole range of behaviour, utterances, actions and interactions between members of the community going about their activities in the world. Naturally there is much potential for variation in practices within different cultural and linguistic communities; indeed we may find the customs and ways of using language exhibited by an unfamiliar tribe difficult to comprehend. Such practices are nonetheless intelligible as intentional and linguistic behaviour, and it is only because we can understand them as such that the difference in details and emphasis of their practices come to light: *"we could not chart these differences, not justify a claim to have done so correctly, unless we could rely on Other Minds to be basically like us; so the key is to make these similarities our bridgehead"* (Hollis, 1994, p.247).

There are strong parallels between this Wittgensteinian insight into language use and the thesis of rational interpretationism, derived from Davidson's argument that the methodology of interpretation, constrained by the demands of Charity, reflects the relational structure of intentionality. We must generally agree on judgements about what is true if we are to use language to communicate, co-ordinating our beliefs about the world with those of others[110]. In particular, there must be agreement among those who use a language as to the circumstances under which a proposition would be held true[111], thus emphasising the situatedness of language not only within a community but also within the world. This is not an authoritarian or charitable empirical assumption

[110] Davidson (1974b) considers intertranslatability to be a criterion of languagehood. He seeks to undermine the possibility of there being radically different ways of going on in the world by denying that it is possible for languages to fail this intertranslatability criterion, arguing that any vocal articulations that cannot be translatable in principles should not count as instances of language-use at all (p.186), and hence do not represent intentional behaviour.

[111] Kripke (1982) famously rejects a truth-conditional theory of meaning in favour of constructing an ersatz notion of meaning based on the conditions under which one would assert a proposition. The connection of language to the world is thus lost on this view.

about the possibilities of two agents happening to converge in their judgements and attitudes towards events and objects in a shared environment (Joseph, 2004, p.67). The only way we can initiate interpretation of another language is to start with general agreements about what is true: we assign truth-conditions to sentences held true by another, based on what we ourselves consider to be true in our language: *if we want to understand others, we must count them right* [by our own lights] *in most matters"*[112] (Davidson, 1974b, p.197).

The requirement of agreement in judgements also suggests that we must largely agree on such things as the appropriate or correct moves that ought to be made in light of using a word, possessing a belief or forming an intention. An intention to use the word 'red' carries with it a normative commitment towards using it in the right way, applying to the correct objects and not to others. It thus constrains the pattern of one's actions and further utterances in a certain characteristic way. From the perspective of an interpreter attempting to understand an agent's linguistic utterances, he can attribute to the agent a grasp of the meaning of a word insofar as the agent uses the word correctly in the appropriate circumstances. By the same token, a belief is attributable to an agent insofar as he behaves in ways that follow from that belief. Thus, possessing a belief entails a commitment towards patterns of action and utterances that would be in accord with that belief, the absence or obvious violation of which would undermine the identity of that particular belief.

It is of course quite possible for me to use words incorrectly, to assert open contradictions and perform actions that thwart my own avowed intentions: if we are unable to make mistakes or violate constraints on language use, belief and action there is no sense in which intentional behaviour could be said to be normative. Whilst the logic of belief precludes the possibility of sincerely holding a belief with the content '*p*

[112] Luntley casts a similar argument in providing a transcendental account of selfhood, citing the self as the ground for the possibility of keeping track of things: an epistemic requirement of contact and interaction with the world (Luntley, 2006).

and not-*p*', the idea that the normative constraints on intentionality and interpretation are necessary does not rule out the empirical possibility of going against them on occasion. Just as I am free (in the sense of being physically able) to intentionally move my rook diagonally across the chess board, so I can also intentionally utter the term 'red' to apply to square things. Such gestures take on an air of stubborn defiance of the constraints of our ordinary practices, and may identify the behaviour as, for instance, humorous or deceptive. But in order for an interpreter to recognise such linguistic behaviour as intentional it must first be clear that the agent has, in the instance of language use, gained mastery of the words he is using. Otherwise, his mis-use will look like error or ignorance[113]. Similarly, I may not act on the basis of a belief I assert: there may be numerous reasons for this but the possibility of, for example, intentionally going against my belief (again, humour and deception are good examples) only arises if sufficient of my behaviour enables that belief to be attributed to me in the first place. Otherwise, it's questionable whether I do possess that belief: I am perhaps deluding myself. The point to be drawn out here is that it is only if one has established a path of the right way to go on that one can potentially stray from it. Thus the empirical possibility of not using words correctly or not respecting the normative commitments entailed by one's beliefs does not undermine the existence of general, holistic constraints on one's intentional behaviour but rather emphasises that our language-games and customary practices are normative in nature.

There is no cut-off point beyond which a specific belief could be said to violate a norm of reasoning, precisely because a specific belief cannot be considered in isolation from a whole set of other intentional states. Yet there is nonetheless a limit to the level of internal inconsistency that can be tolerated before abandoning the project of attributing meaning and propositional thought altogether: at this point a creature could not be said

[113] There may of course be myriad interpretive explanations to accommodate this kind of deviant language use and render it perfectly intelligible. The point here is rather that whilst one can violate the rules, doing so threatens to push one outside the boundaries of interpretability in that specific instance.

to be participating in our shared form of life and his behaviour would thus be beyond the boundaries of intelligibility altogether. The normative standards constituting the shared background of agreement therefore set the limits to sense, as one could not fail largely to conform to them whilst still being intelligible as an intentional, linguistic agent. These standards are not, however, observer-independent. The point is a familiar one, echoing the Background Argument that it is only against a background of largely correct, intelligible behaviour that certain moves can be identified as being in error: *"In order to make a mistake, a man must already judge in conformity with mankind"* (OC, §156). Construed as a general, contextualised description of the assumptions that comprise a shared agreement in forms of life, this requirement of 'conformity with mankind' reflects a feature of shared agreement (Davidson, 1970a, p.222).

Wittgenstein's remarks on agreement in forms of life therefore bear a resemblance to Davidson's conception of the background of rationality, pointing towards similar general conclusions about the conceptual limits on interpretation and intentionality. If we are able to attribute beliefs, desires, linguistic utterances and intentional actions to an agent, we are necessarily doing so within a broad context of shared agreement: *"finding the common ground is not subsequent to understanding, but a condition of it…If we understand [an agent's] words, a common ground exists, a shared 'way of life'"* (Davidson, 1995b, p.51). Wedgwood (2007, p.277) draws on a metaphor from Plato's Republic to characterise this type of position: the Form of the Good is to the understanding what the sun is to vision. Just as we count as being sighted creatures because we are appropriately sensitive to light the source of which is the sun, so too we count as rational, intentional agents because we are appropriately sensitive to the normative requirements of reason, the source of which is our form of life, systems of communication and practice, and ability to get on in the world.

This reading of Wittgenstein's rule-following remarks demonstrates that despite the fact our linguistic and interpretive practices cannot be explained or justified in a way that

guarantees their correctness, this does not diminish their normative force as standards that guide our behaviour and constrain intentional attributions. The assertion “this is simply what I do” is a recognition of the fact that at bedrock, we are minded in a particular way that is exhibited in characteristic patterns of activity that constitute what we can call ‘following a rule’ and using a language. This idea of mindedness is not something that we can grasp and evaluate from the outside since it is not a genuine possibility that we could be radically other-minded. It is just a feature of language that there exist these patterns of use, and these are common to any community to whom we’d attribute the speaking of a language. Thus we gain an important insight from a transcendental consideration of how we go on in the world: using language, intentionally acting and interpreting the behaviour of others all rely on a common system of reference, manifested in the rich and complex context of our activities and practices. This system is not law-like as it is essentially heavily contextual, capable of being characterised only in terms of regularities and patterns in the relations between utterances and action. Indeed, the transcendental understanding of rules and rule-following reveals that even basic laws of logic and arithmetic cannot be understood as platonic entities (Lear, 1986), independent of the human activity in which they are used: *“The criteria of logic are not a direct gift from God but arise out of and are only intelligible in the context of ways of living and modes of social life”* (Winch (1958) cited by Hollis, 1994, p.239). Yet from an empirical perspective the requirements of agreement are genuinely objective and necessary (Lear, 1986, p.271). In the final section of this chapter I will turn to address how these insights from Wittgenstein impact on the relativist concern previously raised, and the question of whether we can go some way towards describing the norms of rationality, which, I argue, are constitutive of our being minded as we are.

5.5. SHARED AGREEMENT

The Unintelligibility of Radical Difference

In chapter two I raised a potential worry that the standards of rationality by which one judges the decision-making process of another might admit of differences between individuals or communities, thus rendering invalid any attempt to evaluate normatively another person's reasons and decisions. This view is supported by sceptical readings of Wittgenstein's remarks on rule-following that suggest the correctness of attributions of meaning and intentionality are only fixed by the contingencies of the community's agreement: a move that opens up the possibility of alternative ways of going on if different communities reach different kinds of consensus.

This concern can now be dissolved. On the rational interpretationist view I have developed, the conjunction of two claims serves to undermine the possibility of relativism about rational standards underpinning interpretation and intentionality. Firstly, the intentional realm is essentially intersubjective, and thereby constitutively bound up with the structure and constraints of interpretation by a third-person. This conception of intentionality as beholden to the possibilities of third-person interpretation finds echoes in Wittgenstein's notion of shared practice and is consistent with a rejection of the "master thesis": a Cartesian view of mentality as a private and inner realm accessed with privileged authority in the first-person. Thus, a person's intentional states are constituted in part by third-person attributions: *"we can see from a person's actions that they believe certain things definitely whether they express this belief or not"* (OC, §284).

Drawing out the parallel with using words is informative in grasping how this constitutive claim works. To reiterate, knowing the meaning of a word is a matter of being able to use it correctly. The pattern that is manifest by this usage can, if we wish to reflect philosophically on it, be construed as being in accord with a rule. But conceptualising the correct use of a word in terms of following a rule leads us into a

seductive misconception about the nature of meaning and the grounds for saying we have gone on in the right way when using a word correctly. The meaning of a word is not some platonic, atomistic entity that can be identified independently of its context of use. Nor does the rule specifying the correct application of a word constitute what it is to mean something by that word. To think that an explanation is needed here of the way the meaning of a word connects to its conditions of correct use is to fall foul of a platonistic reification of meaning that leaves language (and intentional behaviour) unconnected to anything and in need of justification (Finkelstein, 2000, p.67).

If we are led to conceive of meanings as atomistic entities constitutively devoid of any intrinsic normativity, as the sceptical dilemma suggests, we are attempting to adopt a perspective on meaning that is external to the practices and customs in which language is used. Taking this position as a genuine point of view entails that the puzzlement about how standards of correctness for linguistic meaning impose normative constraints on our utterances is inevitable and intractable. If we succumb to this misconception then it appears as though all that is holding our standards of correctness in place are the agreements of the community one happens to be a member of. If this is indeed the case, there is scope for consensus to differ within different communities, and the possibility of relativism about meaning arises. A therapeutic reading of Wittgenstein shows us why this is a flawed view, as it represents a misguided attempt to step outside of our own linguistic and social practices to get a better grasp of their structure:

> "[Wittgenstein] hopes to get us to see that when we envision ourselves occupying an external point of view on language we don't succeed in articulating *any* thoughts - and that he sees our difficulty as one of coming to recognize that the idea of such a point of view creates the illusion of understanding" (Crary, 2000, p.6, emphasis in original).

What this view shows is that it is an error to seek to grasp meaning from outside the perspective of our social and linguistic practices. The positive aspect to this therapeutic view is that meaning is in fact constituted by using words in the right way within the context of our linguistic practices, which I shall come to presently.

The analogy with belief is clear. Having a belief (or other intentional state) is a matter of speaking and acting in a way that is appropriate or reasonable in light of this content, such that that belief is attributable by an interpreter observing one's behaviour. Possessing a belief or forming an intention is much like following a rule: certain normative commitments are entailed by, for example, the sincere assertion that I believe it is raining, such as not simultaneously asserting that it is not raining, or when coupled with a desire to remain dry, acting in such a way as to avoid getting wet when going outside. It is an error to conceptualise beliefs as standalone psychological objects that can be identified independently of the normative relations they bear to other intentional states, utterances and the actions that are intended or performed in light of them. If we think of beliefs in this reified way, from outside the practices in which they occur, then all that fixes judgements about beliefs are reconstructed ersatz notions of correctness constructed by one's community. Without anything intrinsically normative to belief, the possibility of alternative notions of rationality and practices of reasoning look entirely plausible. However, to continue the analogy with meaning, it is an error to mythologize the nature of belief in this way, abstracting it away from its context and seeking an external perspective on the way beliefs are structured. The normativity of belief is intrinsic and constitutive, but it emerges only within the context of a form of life and cannot be grasped from outside. We are embedded within our form of life, which entails that all intentional behaviour is understood from within this framework. The conceptual possibility that gives rise to the relativist concern is therefore an illusion, because there is no conceptual space open for behaviour to be perceived as intentional and yet not to operate within the bounds of intelligibility shaping our form of life. The very idea of alternative ways of going on and radically different structures to the relationship between reasons, thoughts, actions and linguistic utterances is therefore devoid of content.

This brings us to the second claim: that the boundaries of intentionality are set by the standards of rationality shared by all intentional agents. This is what the constitutive rationality of the intentional domain I have described as the Rationality Assumption amounts to. Because what constitutes belief and meaning is partially determined by these standards, the very fact of being able to apply these concepts to the movements and utterances of a creature ensures he is participating in a shared form of life and is thus subject to these very standards. The norms governing the process of intentional attribution (encapsulated by the Rationality Constraint) converge in all creatures that can be interpreted as intentional beings, and are thus universal in virtue of our nature as social, action-guided and linguistic creatures[114].

The Rational Norms of Practice

With this notion of non-principled normativity in place, the thesis of rational interpretationism becomes more cogent and less indicative of an empirically questionable thesis that renders most human behaviour irrational. The principled conception of rationality is a misguided attempt to decontextualise and codify the patterns, connections and ways of going on that emerge from our ordinary intentional and linguistic practice: *"To say that there exists rationality is to say that perspectives blend…there is behind it no unknown quantity which has to be determined by deduction, or…demonstrated inductively"* (Merleau-Ponty, 1962, p.xi). This does not, however, mean that rationality is a mysterious and coincidental norm that ranges over our intentional behaviour:

[114] I am not committing to giving an account of rationality independently of the requirements of intelligibility. Epistemologically, we seamlessly perceive intentionality in movements and utterances but this understanding cannot be explained by an ontological account of the bounds of intentionality from the outside, in terms that reduce the constitutive norms of rationality to something else, such as a set of procedural criteria. To think that a reductive account of the normative constraints on interpretation is necessary is to fall prey to the philosophical temptation to step outside of one's mindedness in order to grasp its structure. This is a line of thought akin to McDowell's critique of a sideways-on view of meaning and the description of intentional terms (e.g., 1994, Lecture II). The relation between the claim that there are epistemic normative constraints on interpretation and the ontological thesis that the intentional realm is rationally structured is therefore not an explanatory or reductive one, but rather one of interdependence. There is content to the idea that rational constraints operate on interpretation, since there the relations between beliefs are matters of what one ought to believe, but these cannot be captured in terms that are reduced beneath the bedrock level of norms.

> "our responsiveness to reasons is not supernatural, we should dwell on the thought that it is our lives that are shaped by spontaneity, patterned in ways that come into view only within an enquiry framed by what Davidson calls "the constitutive ideal of rationality"" (McDowell, 1994, p.78).

Contrary to the implicit assumptions of the predominant view of rationality, an agent is not rational in virtue of his subscription to rational principles. But it is difficult to grasp how interpretation and intentionality could be thought to be normative if it is not possible to provide clear rules for the general application of normative standards in judging behaviour and attributing intentional states. This inability to prescribe at the abstract, generalised level does not, however, entail that in individual circumstances it is not possible to provide a judgement as to what one ought to believe or do.

Dodd (1999) provides a minimalist account of truth to demonstrate that the accumulation of instances in which we can make normative prescriptions about what one ought to assert, all of which prescribe that one ought to make assertions that are true, does not add up to prescribing a norm of truth. The analogy with the norms of belief is as follows. In particular circumstances, it might be the case that one ought to hold a belief that is true. For instance, if I am sitting in a pub with a pint of bitter on the table in front of me, I ought to believe that there is a pint of bitter on the table in front of me. Similarly, if last orders at the pub are called at 11pm, I ought to believe that last orders will be called at 11pm. Both of these facts are empirical truths, but enumerating all the instances of what I ought to believe in these particular situations ought not to result in their common denominator of truth being abstracted away from the distinct circumstances, aims, interests and specific contents of each of these situations to provide a context-free, generalised normative obligation. What makes it the case that I ought to hold a particular belief is not adherence to a norm of truth; truth is not an abstract force-maker in prescribing what I ought to believe, but rather in that particular context, I ought to hold a belief to which the truth predicate applies.

Principles of rationality are derived from the complex and varied relations that exist between our utterances, actions and interactions with the world and other people, but

have no life or platonic existence outside the context in which they are characteristically used. On the view I have described rationality is manifested in our ordinary practice and any attempt we make to codify its normative demands as universal generalisations or principles is necessarily an abstraction away from this practice. Descriptions of the kinds of relations that obtain between behaviours, codified as principles, arise from the patterns of use and action that make up the broad scope of our ordinary activities and practices but do not themselves carry any normative weight. They are abstractions that emerge when we philosophically reflect on the form of life and human activities in which language and intention find their meaning. This does not mean, however, that rationality exists entirely in the eye of the beholder: *"It exists in the ability of the agent to govern her behavior using those norms in context"* (Gerrans, 2004, p.44). The source of normativity is thus located within the scope of human life and behaviour, on account of our ability to reflect on our reasons (Korsgaard, 1996, p.xii).

The implications of this view for our conception of rationality are striking. Much of the empirical and philosophical literature on human reasoning and rationality has focused on the question of whether or not we conform to principles derived from logical laws, on the assumption that in order to be rational, one's beliefs and reasoning processes ought to display conformity with these principles. Abstracted from specific contexts, logical principles have been taken to constitute a normative ideal to which we should aspire, whilst the processes of reasoning, choice preferences and decision-making have been evaluated in isolation from the context in which they occur, in terms of whether or not they conform to the schematised structure of a logical inference. In light of the rule-following considerations, we can see that this view is mistaken and understand why this is so. The very idea that principles provide a specification of ideal rationality reflects the misconception that normative standards somehow exist independently of us, laying rails to infinity that prescribe the correct action for every instance of intentional behaviour. As is clear from the threat of the regress of interpretations that follows from this view of rules, it is a mistake to consider that such

principles themselves play the role of *"normative force-maker"* (Schroeder, 2003) here, compelling us to believe or intend one thing or another in a given situation.

The interpretive process applies to individual speakers and our endeavours naturally concern the attribution of intentional states to that individual at any one time. Put this way, the constraints operating on interpretation may be thought to be individualistic and situation specific, to the exclusion of the fact that the individual is integrated into a wider linguistic, social, cultural community. On the account of rationality I have sought to defend, the individual cannot be understood in isolation from his relation to the world and to the community of which he is a part. I am therefore largely in agreement with Bortolotti's (2004) sentiment that interpretation depends on acquiring knowledge of the person's environment, his actions in relation to it and his linguistic interactions with others: in short, the rich and varied context in which behaviour ordinarily occurs.

Take the example of a person who deliberately asserts a contradiction or paradox for comic effect. Within the context in which the remarks are made, and knowing the speaker's humorous intention, his utterances are perfectly intelligible. It is only if, in an attempt to identify whether the assertion itself (perhaps given schematically in terms of propositions held true) conforms to principles of rationality that we uncover a problem for interpretation, since considering the assertion in isolation entails that we lack the essential background resources to make sense of the contradiction. This is what I mean by the claim that it is not principles doing the work in constraining our interpretations: the intelligibility of behaviour arises from its being embedded in a rich context from which we identify normative constraints on how we ought to use words and how we ought to act if we are to be understood. This view has parallels with much thinking in the hermeneutical and phenomenological traditions (particularly in the work of Gadamer and Husserl respectively), and has gained increasing currency in Anglo-

American philosophy[115]: *"We can no longer regard the social and physical environment as simply surrounding the psychological subject…contextual factors inextricably permeate the field of psychological investigation"* (Pettit & McDowell, 1986, p.14).

Given that I have argued that efforts to abstract general normative principles of rationality from the context of the activities in which they are collectively manifest are fundamentally flawed, there is an obvious bar to making any broad but informative generalisations about the standards of rationality that do constrain our linguistic and interpretive practices. If we are looking to ascertain how our interpretive practices are constrained in specific, individual instances of interpretation we lose sight of the broader patterns of practice that enable the behaviour to be intelligible as intentional action. Whatever standards are in play, they are always necessarily heavily contextual and subject to wide variation depending on the circumstances of the particular interpretive situation at hand: they are not absolute and indefeasible.

Reflecting back on the principles that emerged from Davidson's account of Radical Interpretation, we can construe the demands of Charity as informative but not strictly prescriptive abstractions of the boundaries of intentionality and interpretability. The principles of truth and coherence provide a description of what these constraints look like at an abstracted and generalised level ranging over the entire intentional domain: one must get the world mostly right, and one's beliefs, desires and other intentional states must generally hang together as a coherent whole, in order for one to be interpretable and to count as an intentional agent. However, Charity does not provide us with the resources to make specific *a priori* prescriptions about what one rationally ought to intend or believe in particular situation[116] (Cherniak, 1981, p.165). *"Charity can offer no precise interpretive prescriptions"* (Malpas, 1992, p.152) if it is assumed to

[115] A precedent for this view within the philosophy of psychology was set by Dreyfus and Dreyfus (1986), who considered the view of the mind as a formal, symbolic information-processing system to be a fundamental misconception driving artificial intelligence research.

[116] Heil refers to Charity as *"parsimony applied in the mental realm"* (1989, p.574), implying it is not an ideal or strict constraint but is rather about the optimisation of standards of rationality in particular instances.

consist in a web of context-free principles that can be filled out with the contents and antecedent beliefs relevant to the circumstances of interpretation. At its root the process of intentional attribution is a matter of degree and intrinsically contextual judgement by an interpreter, and this reflects the nature of the subject matter at hand: the richly contextual practices and customs that constitute our form of life.

6. IMPLICATIONS FOR CAPACITY?

6.1. FROM RULE-FOLLOWING TO DECISION-MAKING

In this final chapter I attempt to draw out the main themes and arguments that have been presented, to consider their implications for clinical judgements about capacity. Whilst it is beyond the scope of the discussion to propose any direct applications of these ideas to clinical policy, training and practice, I argue that such insights may provide a useful first step in framing practical questions about the concept of capacity and how it ought to be assessed.

I showed in the first chapter that the cognitive conception of capacity underpinning the test of capacity set out in the MCA implies the process of a person's decision-making can be judged independently of the decision outcome. What matters for capacity, on this view, is that the person is able to use or weigh information in coming to a decision, irrespective of the perceived wisdom of the decision itself. Nonetheless, I suggested that procedural standards alone do not map on to what clinicians would consider to be indicative of capacity and incapacity. If values that seem distorted or patently false beliefs influence the decision-making process, capacity may be lacking even if there is a recognisable and logical connection between the input factors and the decision outcome. This indicates there is some kind of epistemic standard operating on capacity judgements disciplining what legitimately counts as a recognisable reason for a decision, and the question was raised as to whether such standards could vary between groups. The discussion here was premised on the idea that whilst epistemic standards potentially looked relative, the norms governing the process of decision-making itself were fixed and objective. In other words, whatever the content of the particular beliefs and values entering into decision-making, the process from inputs to outputs was presumed to be governed by the abstract norms characterised by procedural rationality. The rule-following considerations have, however, both given us reason to question this assumption and provided us with the resources to explain why

and how epistemic considerations enter into judgements about a person's decision-making process. I will take these points in turn, dealing first with the negative implications of the rule-following analogy and then considering how capacity judgements can proceed reliably and accurately in spite of the problems for assessment I have raised.

Decision-Making Isn't a Psychological Process

What does it mean to say that a person is (or is not) using or weighing information in coming to a decision? The capacity criteria require more than an understanding of the information given about a potential treatment or course of action; they require an indication that this information has been used appropriately in some unspecified way to influence the decision outcome. The metaphor that springs to mind in attempting to describe how this process might be envisaged is that of an information processing black box: a visual metaphor common to cognitive psychology, which is concerned with explaining mental functioning primarily through sophisticated flow diagrams of cognitive mechanisms. On this model, various factors serve as inputs to decision-making, including the information given along with a person's known beliefs, values, desires, fears and so forth. These feed into a process, whereby various cognitive mechanisms operate upon the information received, and subsequently an output emerges in the form of a decision about what the person wishes to do. A person thus uses or weighs information to the extent that these psychological mechanisms operate on the information, beliefs and desires relevant to the decision being made, acting like a computational information-processing function to produce an appropriate output. How, though, could the operation of such a psychological mechanism be judged? It is plausible that one could take on board the information, deliberate with it, weigh it up against one's beliefs, values and so forth and make a decision on the basis of that information, but equally plausible that one could take on board (and thus use an identical input to the former case) but discard the information and make a decision in spite of it.

It is essential to the test of capacity that examining the process of decision-making can enable an observer to distinguish between these two possibilities, since they differ in ways that are significant for an assessment of capacity: the former indicates fulfilment of the using or weighing information criterion, whereas the latter might not. Yet all that a clinician has to go on in judging whether a person is using or weighing information is the outcome of the decision, and some awareness of the input factors (many of which will be unknown to the clinician). Judgements about the decision-making process are therefore based upon the perceived connection between the input factors and the output: on whether or not the decision is one that ought to have been made in light of the information given. This is what gives rise to the thought that a person must have recognisable reasons for his decision. The decision ought to be one that in some respect follows from the person's beliefs, including the relevant information that has been conveyed. There ought therefore to be a reasonable connection between the inputs and the decision outcome if the person is to be judged to have used or weighed the information in coming to that decision.

Now the relevance of the insights from rule-following emerges, if we think of the process of decision-making as akin to the application of a rule operating like a function on a given set of inputs to produce an outcome. The question arises: what connects successful using or weighing of information with the decision outcomes that ought to follow from it? We saw previously that it is misleading to seek an explanation of the relationship between a rule and the conditions under which it correctly applies by appealing to a platonic mechanism, because positing such a mechanism cannot justify the claim that the process followed is the right one (such as 'plus' in the case of the rule of addition), as opposed to some deviant function (such as 'quus'). Similarly, is it a mistake to seek the connection between sets of inputs and the right kind of output that follows from them by appealing to a psychological mechanism of using or weighing information, because such a mechanism is of no help in differentiating between going

on in the right way (successfully using or weighing information) and the wrong way (following a deviant function that is indistinguishable in most cases). Positing a mysterious psychological mechanism does not enable one to distinguish capacity from incapacity solely in terms of the process of decision-making. Rather, what it is to successfully use or weigh information, and thus fulfil the essential criterion for capacity, is just that one reliably gets the right kind of output, that is, the right kind of decision outcomes that reasonably follow from the information one is given[117].

A necessary caveat must be added here, in extrapolating lessons from Wittgenstein on rule-following in language use and applying them to judgements about decision-making processes. Despite the fact that the correct meaning of a word is something that emerges from its use and is not a platonic, atomistic entity, it is nonetheless natural to think that there are prescriptions on what counts as the correct or incorrect use of a word. It is not so clear that using or weighing information and having a recognisable reason for one's decision admits of such tightly bounded prescriptions. We might say that a particular decision reasonably follows from the information given to a decision-maker, and either condone or criticise it, but we would not necessarily consider the decision to be correct or incorrect. This is in part because of the vast number of factors that could potentially enter into and influence the decision-making process: a point to which I shall return below, but it is mentioned here in order to justify the vague terms with which I am describing the normative appropriateness of decision-outcomes, in sharp contrast with the definitive terminology of 'correctness' and 'incorrectness' employed in describing linguistic usage.

The analogy with Wittgensteinian insights into rule-following points to a conclusion that has two significant negative implications for the conception of capacity upon which capacity legislation is based and its assessment by clinicians. The conclusion is that

[117] To use terms originating with Reichenbach, the argument here is analogous to Kuhn's argument in the philosophy of science that the context of justification is not distinct from the context of discovery (Kuhn, 1970a).

any attempt to appeal to a psychological mechanism or process rests upon a misconception about the nature of the intentional realm, which is what McDowell calls the “master thesis”. This misconception fuels the perceived need to seek an explanation of the way in which norms and rules constrain intentional behaviour through a psychological mechanism or process. But if we are freed from this illusion and understand the intentional realm correctly, we see that no such explanation could succeed. Identifying the process that links a set of inputs with the outcome that ought to follow from them via the application of a psychological mechanism is explanatorily redundant.

This conclusion supports the claim that it is futile to attempt to define what it is to possess capacity according to purely procedural criteria. Rather, a person’s ability to use or weigh information emerges through his general reliability in making decisions that reasonably follow from the information given. It is a mistake to consider that a specification of mechanisms or processes abstracted away from all of the everyday instances of decision-making and interpersonal communication that comprise our ordinary practice provides an essentialist prescription for what constitutes reasonable decision-making: a prescription that is subsequently applicable in a particular instance to determine whether a decision is made on the basis of recognisable reasons. To be judged as having reasons for one’s decisions, whether these are good or bad reasons, entails that one is participating in a shared normative practice. It is this practice that supplies the norms by which one’s intentional behaviour is judged and constrains the kinds of outcomes that would be deemed to follow reasonably from the information one is given on a particular occasion of decision-making.

Decision-Specificity

The first implication of this conclusion is that it raises a question about just how decision-specific an assessment of capacity can be. Whilst there are good reasons to avoid taking a status approach to the possession or lack of capacity, the conclusion

that one's ability to use or weigh information arises from a general ability to act for reasons and be interpreted as behaving intentionally potentially undermines the idea that a capacity judgement ought to be specific to an individual instance of decision-making. The issue here turns on how we are to understand the term "decision-specific". If what is meant is that the information and input factors (such as known beliefs and values) a clinician considers to be relevant to the decision at hand are isolated from any contextual factors that are not obviously related to the process of making that decision, then I do not think a capacity judgement can be decision-specific. This is because the very identity of beliefs can only be established in light of their relatedness to innumerable other beliefs, actions, interactions and utterances all of which form a rich and complex pattern of practice. Conceptually isolating beliefs, values, intentions and so forth from their constitutive relations and construing them as atomistic, determinate mental entities entails that we lose sight of the normative structure of reasons that forms the very basis of judgements about whether or not a decision outcome reasonably follows from its inputs.

To clarify this point, consider the problems faced by a purely procedural conception of capacity identified in the first chapter. Based on procedural criteria alone, the anorexic who can set out his reasons for refusing to eat, premised on his high valuation of thinness, or the schizophrenic who refuses a blood transfusion on the grounds that he believes the blood will be poisoned by MI5, look as though they possess decision-making capacity. This is because the values and beliefs attributed to them have been taken out of the context in which they occur and used as starting premises in an inferential reasoning process, and logically the process from premises to conclusion is intact[118]. Shorn of the epistemic and rational relations they bear to other beliefs, utterances and actions, the potentially capacity-undermining nature of these intentional states cannot be identified. If, on the other hand, we take specificity to pertain to a

[118] It might be protested that such beliefs are patently false, but content alone does not enable capacity to be distinguished from incapacity here: we all hold false beliefs, some of them stubbornly.

particular decision whilst fully acknowledging the contextual embeddedness of the factors entering into that decision within the person's own life, presupposing his status as an intentional agent who is responsive to reasons, then we are in a far better position to assess whether the decision outcome is one that the person ought to make. This assessment necessarily includes a judgement about the reasonableness of the input factors, construed not as starting premises for an inference but rather as features of the person's intentional and motivational structure that are subject to a range of epistemic normative constraints themselves. I will expand upon this point when considering the role of context in more depth below, but for now I suggest that we can reject the idea that decision-specificity means isolating the process leading to a particular decision from the rich context that gives the relationship between its inputs and output a normative structure.

Substantive Epistemic Standards

The second negative implication of the conclusion that decision-making is not underpinned by a psychological mechanism is that capacity judgements cannot be content-neutral. The decision outcome will in fact impact upon the judgement as to whether or not the patient has successfully used or weighed information in coming to a decision. This is because assessing whether or not a person has used or weighed information just is a matter of determining whether the decision outcome is one that reasonably follows in light of (among other things) the information given: there is, as we have seen, nothing intrinsic to the process itself that distinguishes success from failure on this criterion of capacity. In spite of the wording of capacity legislation and its insistence that unwise decisions do not undermine capacity, it appears that evaluating the outcome of a decision-making process is inevitable if the process itself is going to be assessed. This is a problematic conclusion as it potentially undercuts the main motivation for seeking a process-based conception of capacity, which was the avoidance of undue medical paternalism. Paternalism could result from capacity being denied on the basis of patients making choices that are considered unwise or against

medical advice, hence the understandable attempt to ensure that no substantive criteria formed part of the test for capacity. The suggestion I am making is that there are indeed substantive standards underpinning capacity judgements, but that this lack of content-neutrality need not be considered to lead to judgements about capacity being based on the perceived wisdom of decisions.

This claim requires unpacking if it is to avoid the risk of promoting paternalism, and the important distinction here turns on a subtle point about the nature of the normative commitments entailed by possessing beliefs and having reasons for one's decisions. Suggesting that there are indeed substantive standards underpinning judgements of capacity implies that in order to be deemed to possess capacity with respect to a decision, there are certain things one ought to believe or do (and want or value, although I am focusing here on epistemic commitments specifically). Failure to hold beliefs that accord with the requisite epistemic standards or holding beliefs that violate such standards would entail that such a person would be deemed to lack capacity. The question of whether or not the inclusion of substantive criteria would accord with clinical intuitions about capacity would depend on how these standards are defined. If they are construed as requiring agreement with what a clinician takes to be reasonable, or at least not significantly deviating from prevailing medical opinion, then the risk of paternalism arises. Decisions risk being undermined by paternalistic intentions if they are made in part on the basis of beliefs that do not accord with these epistemic standards. This is clearly an undesirable consequence. So how ought substantive epistemic standards to be characterised?

In earlier chapters I used the theoretical apparatus of Davidson's project of Radical Interpretation to examine where the limits of interpretability and possibilities for intentional attribution lay. His Principle of Charity revealed two interdependent constraints of Correspondence and Coherence which were, he argued, necessary normative presuppositions for an interpreter seeking to attribute intentional states and

linguistic meaning to a people without any prior knowledge of their beliefs or language. From this theoretical reconstruction of the boundaries of interpretation emerged the epistemic standards of truth and coherence, taken broadly to be descriptive criteria of intentionality. At first glance, it therefore looks plausible to suggest that truth and coherence could be taken to form the epistemic standards underpinning judgements about a person's reasons. Indeed, these criteria are elevated to the status of normative prescriptions by the assumption of the Rationality Requirement that one ought to be rational. However, I diagnosed this assumption as resting upon a misconception. It would be a mistake to translate the normative constraints exposed by Radical Interpretation into an argument that truth and coherence supply normative epistemic standards circumscribing judgements about people's decision-making processes. This is because, as I argued previously, abstract generalised principles do not themselves normatively constrain our shared practice: the normativity of such principles emerges only within the context of the innumerable examples of practice from which they were derived. Seeking principled epistemic standards stripped of their context throws the baby, the source of normativity, out with the bathwater.

The epistemic criteria of truth and coherence do not, therefore, normatively constrain intentional behaviour, although they may be useful descriptive generalisations about what beliefs look like. It thus appears that we have made no progress in identifying any criteria to fill out the claim that there are substantive epistemic standards constraining interpretation and judgements of decision-making, and there are good reasons for resisting any attempt to abstract and codify these standards. But some positive argument is surely needed to carry the weight of the claim that the epistemic standards intrinsic to capacity judgements exist and do not admit of radical variation between communities, or between individuals. I therefore turn now to consider the positive implications of the view of capacity I have argued for.

We can say only the following about the norms shaping our practice: that in individual instances of intentional behaviour it is possible to identify whether a particular belief or action is one that reasonably follows from a given belief, or in the case of decision-making, whether a decision-outcome is one that reasonably follows from the information relevant to the decision at hand. In the first instance, this is not because the belief is subject to particular normative obligations of truth or coherence, but rather because what it is to have a belief is just that one reliably acts in a normatively appropriate way in light of it. By the same token, the input information is not constrained by specific epistemic standards in the process of being used or weighed, but rather what it is to use or weigh information is just that the resultant outcome is one that ought to have been reached. This normative obligation ought not to be construed in terms of adherence to principles but rather as a broad and general responsiveness to reasons that is manifest in our shared form of life.

It is possible to reconcile the claim that capacity judgements cannot proceed through using content-neutral procedural criteria alone but involve substantive epistemic standards, with the need to avoid paternalistic denials of capacity based on the perceived lack of wisdom of decision outcomes. The key to understanding how this reconciliation can proceed lies in how we are to construe the content-ladenness of decision outcomes and the standards by which these are judged. I have argued that isolating the process involved in making a specific decision from the context in which it occurs leads us down the wrong path and breeds misconceptions about what it means to use or weigh information. The same is true of attempting to judge the determinate content of a decision as an outcome that one ought or ought not to have reached on the basis of the information available to the decision-maker. The fact that the decision outcome is evaluated in a capacity assessment does not mean that its content, in isolation, ought to be considered wise or good in order to indicate the presence of capacity. On its own, the decision outcome can neither be good nor bad, wise nor

unwise, appropriate nor inappropriate[119]. What makes the decision indicative of capacity is that it reasonably follows from the information given, the person's beliefs and so forth, all of which is normatively disciplined by the form of life we collectively inhabit as intentional agents. The decision content and the process by which it is formed are not conceptually separable elements of decision-making.

On the rational interpretationist conception of intentionality I have adopted, what makes the decision an instance of intentional, reason-guided behaviour is just that it does bear normative epistemic and rational relations to the world and to the person's beliefs and other intentional states. Thus an evaluation of the decision outcome is not independent of its preceding process or the normative appropriateness of that decision within the context of the person's life and values. In light of this, the epistemic standards underpinning capacity judgements do not prescribe what one ought to do or believe and thereby constrain what decision outcomes one ought to reach, which would give rise to paternalistic judgements where the decision is considered to be unwise, but instead circumscribe the general normative structure of decisions made on the basis of reasons.

6.2. THE NATURE AND SCOPE OF CAPACITY JUDGEMENTS

Holism and Context

The view of capacity judgements I am defending suffers from the disadvantage that it is largely defined negatively by reference to what it rejects, rather than permitting a constructive positive argument to be developed about what capacity is and how it ought to be judged. Whilst a Wittgensteinian form of philosophical therapy is supposed to free us from the need to seek such positive explanations and justifications of our practice in the first place, I consider that adopting a form of quietism about capacity judgements is not a plausible option for a philosophical endeavour aiming to gain traction on the conceptual problems of assessing capacity in clinical practice. Hence although what

[119] Unless of course that decision leads to an action that breaches some legal or moral standard, for example, by causing harm to others.

follows gestures towards a constructive account of the framework in which capacity judgements take place, I recognise the inherent limitations of attempting to construct positive philosophical theses from insights largely derived from Wittgenstein's remarks on rules and Davidson's explicitly theoretical reconstruction of the grounds of interpretive judgement.

An ability to be responsive to and act for reasons is a constitutive hallmark of intentional agency. The thesis of rational interpretationism advocated here claims that reasons, and the beliefs and desires that comprise them, ought not to be construed as private mental entities bearing contingent relationships to utterances and actions but instead as features of intentionality that are by their very nature interpretable by observers and disciplined both by the world and the norms of social practice. Having reasons for ones actions and decisions is therefore a matter of being able to respond to information gleaned from the world and from others in a way that is normatively appropriate and is interpreted as such by others. Different cultures and communities will undoubtedly diverge in many aspects of social practice, for example by placing high value on individual or community welfare, holding medical opinion or religious doctrine in high regard, considering scientific evidence or spiritual guidance to weigh in the balance of decision-making, and so forth. However, all such divergences in practice fall under the rubric of intentional behaviour and as such possess the intrinsic similarities captured by what I have called the normative structure of reasons. There is no abstract form to the structuring of reasons but we can still go some way towards charting their boundaries, or what Malpas (1992) refers to as the horizon of rationality, through cases where this structure breaks down.

I suggest that to understand how this holistic normative account of the structure of reasons impacts upon individual judgements about a person's reasons, the insights of the Background Argument for rationality can be pressed into service. The norms of rationality are broad and messy, but they provide a framework within which intentional

behaviour, whether it is normatively appropriate or not, can be interpreted and understood. It is only in virtue of the implicit background structure of interconnected reasons, actions, intentional states and utterances that individual cases are able to be picked out as appearing irrational: a note of discord in an otherwise fairly coherent and harmonious symphony of intentional behaviour.

Consider how acknowledging the essential role of this background affects judgements about mental capacity. Capacity is potentially undermined if a person makes a decision which does not reasonably follow from the information he has been given about the choice to be made. His behaviour can be understood as intentional if there is a clear reason explanation available for his decision, but if there are grounds to suspect the reason is caused by a mental impairment it will not count as being recognisable in the sense required for capacity. The difficult question facing a clinician is how to determine whether beliefs and other intentional states or values that appear out of the ordinary are impairing the decision-making process. It might be the case that there are elements of the person's belief and value system or motivational structure of which the clinician is unaware but in light of which the decision would reasonably follow, or it might be the case that it is indeed a mental impairment influencing the decision and thus the decision ought not to be respected as the expression of an autonomous choice.

Presume for a moment that we can understand the person's assertions about his beliefs and decision in isolation, without questioning the meaning of these utterances. If there is any logically coherent reason connection between his asserted beliefs and the decision outcome no resources are available for the interpreter to distinguish between decisions that are the result of a pathology of belief from those that are legitimately made. As I have argued, examining the procedural criteria alone is of no justificatory help, and if the content of the beliefs entering into the process are judged on the basis of their truth or internal coherence with other beliefs, then many of our ordinarily held beliefs would fail to lead to reasonable decision outcomes. However, if the rich

framework of rationality within which reasons operate is taken seriously, then the conceptual resources are available to recognise blips in this rational landscape in a far more refined manner than simply identifying bizarre content and attempting to grasp whether or not it is indicative of pathology. For instance, in the case of the anorexic patient his overvaluation of thinness will likely be intrinsically bound up with numerous beliefs about how he appears, issues of self-esteem and self-worth, positive aspects of self-control, perceptions of aesthetic ideals, and so on. All of these factors and many more will modulate this valuation of thinness and each subtly influence what that valuation means and how it impacts upon his decision-making. Far from being so broad and general as to paralyse normative judgements about individual instances of decision-making, the global, holistic structure of reasons provides the backdrop of understanding against which potentially capacity-undermining divergences can be identified.

In chapter two (section 2.4) I challenged the view that capacity could be determined either by appealing to the procedural logical coherence of the process of decision-making or by evaluating the truth of or shared agreement about the particular beliefs that influence the decision outcome. Such standards fail to provide a criterion to distinguish between two cases in which different judgements of capacity would intuitively be made. A Jehovah's Witness has reasons for his decision to refuse a blood transfusion that legally ought to be respected, whereas an individual with the schizophrenic delusion that the blood for transfusion is poisoned does not have valid reasons. I used the disjunction between these two examples to motivate the concern that there could be different but valid epistemic standards governing the decision-making process. Having now closed off this conceptual possibility, how is the difference between these cases to be understood? I suggest that the following conclusions can now be drawn. Seeking a criterion to distinguish capacity from incapacity in these cases is a futile exercise. If we examine the process stripped of all content and context, there is nothing to differentiate them. If we look to the epistemic status of the beliefs

influencing decision-making, neither truth, nor plausibility, nor level of communal agreement can support a distinction between the cases. We cannot generate a context-independent checklist for determining capacity, but this does not imply that the judgement is arbitrary: there is content to the idea that there is rational shape to the appreciation of a person's reasons. Although there is an inherent impediment to using examples as they cannot provide all the facts that might be relevant to an assessment of capacity, I suggest that we can gesture towards the kinds of questions a clinician conducting an assessment ought to bear in mind in the two cases mentioned. For example, one might query the consistency of the patient's decision-influencing beliefs with his broader worldview, the empirical sensibility such beliefs have and whether they are amenable to revision or argument, and the cultural acceptability and rationality of the values impinging upon decision-making, to name but a few possible considerations. It is only if we acknowledge the historical, social, physical and environmental context surrounding each decision and the relational nature of the intentional states entering into the decision-making process that we are in a position to identify where anomalies occur that might indicate impairment to decision-making capacity. Whilst we cannot provide a clear-cut prescriptive specification of how capacity judgements ought to work, there is nonetheless a rational structure to the determination of capacity.

If an asserted belief is sustained in spite of being patently false or failing to be at all coherent with other beliefs an individual expresses or those of his community, there are grounds for suspecting the belief to be an irrational one, possibly caused by a psychopathological impairment. It is the failure of the rational and epistemic relations that ought to obtain that picks the belief out as being irrational, not the attributed content of the belief itself. This rational background accommodates potential divergences in epistemic norms of reasoning and evidence-weighing, as the individual's own belief and value system forms part of the rational structure within which irrationality is identified: the context of understanding the individual's other beliefs and the background of his social and historical environment provides this

framework. This point does not advert to the internal coherence of the individual's thoughts and actions conceived of as meaningful entities in isolation from the world. On the account I have developed, the identities of his thoughts, utterances and actions are partially constituted by the relations they bear to the world beyond his own brain, and by their interpretability by third-person observers. Essentially then, the holism and context-ladenness of the intentional realm provide a richer framework for identifying instances of irrationality that may undermine capacity than any attempt to grasp the singular process of a particular instance of decision-making can achieve.

Philosophical Insights in Practice

To gain an understanding of how the insights into the holistic and context-laden norms of rationality and reasons I have sought to clarify can be brought to bear on practical issues in clinical judgement, I will briefly consider a recent tragic case brought to the public's attention amidst significant debate among clinicians and law-makers about the ethical implications of implementing the Mental Capacity Act in practice. Kerrie Wooltorton was a 26 year old woman who had been diagnosed with an emotionally unstable personality disorder and depression for a number of years[120]. She had attempted suicide on nine previous occasions by drinking anti-freeze, but had subsequently accepted life-saving dialysis each time. In September 2007, she made a further suicide attempt and called an ambulance. A few days previously she had drafted an advance statement dictating that in the event of a suicide attempt she did not want to receive life-saving treatment. The statement asserted that if she called an ambulance, this did not indicate she had changed her mind and wished to be treated, but rather that she did not wish to die alone and in pain. On admission to hospital she presented this statement and directed doctors to read it when they asked her if she wanted treatment. The consultant renal physician assessed that Ms. Wooltorton possessed the capacity to refuse treatment and as a result she died four days later,

[120] Many details from the coroner's report are not in the public domain, and so all information regarding the circumstances of her admission to hospital and subsequent death has been gleaned from subsequently cited journal articles and responses.

having accepted only palliative care. As she was deemed to possess capacity, to disregard her decision or treat her on the basis of a best interests' judgement would have been unlawful. In September 2009, the Norfolk coroner upheld the decision of the treating doctors (Brannan et al., 2010). Whilst this case is undoubtedly complex and prompts many ethical, legal and clinical questions regarding the applicability of advance directives (MacLean, 2009) and the relation between different parts of the legislature (for example, the interface between the Mental Health Act (2007), the Mental Capacity Act (2005) and the Suicide Act (1961)), I wish to focus here on the determination of Ms. Wooltorton's capacity to refuse treatment that would likely have prevented her death.

Given the clear precedent at common law (*Re* C) that the presence of a mental disorder is not a bar to the presence of capacity, Ms. Wooltorton's capacity would have been assessed according to the criteria laid out in the MCA: understanding, retaining, using or weighing information, and communicating a decision[121]. From the detail of the advance statement, demonstrating awareness of how her actions might be interpreted and the consequences that would follow, it is evident that Ms. Wooltorton was aware of the implications of her actions and cognizant of the fact that without medical intervention, her suicide attempt would likely succeed. In recording a narrative verdict, the coroner agreed that her treatment refusal was made in full knowledge of the likely consequences (Dyer, 2009). There is thus little question that strictly speaking the patient could be thought to fulfil all the criteria, entailing that legally, her decision must be respected.

Much debate arose in light of this verdict, particularly among clinicians who were concerned that Ms. Wooltorton's mental capacity to refuse treatment might have been impaired. On the face of it, she had a recognisable reason for her decision: if we

[121] Although the MCA did not come into full legal force until 1st October 2007, clinicians ought to have undergone training about its provisions and put them into practice by the time of Ms. Wooltorton's admission in September of that year.

consider that a factor in the decision-making process was a desire to die (this was thought to be in part because she was unable to bear children due to a medical condition), her action of drinking anti-freeze and refusing treatment reasonably follows, even if we would not condone that desire. Abstracted away from its context, there is nothing about this decision-making process that undermines capacity. But ought the desire to die be taken to be a simple premise in an inferential process? I suggest not. The point I wish to draw out here is that her treatment refusal ought to be understood in light of the complex context in which it was made, before a judgement as to whether or not she possessed the requisite capacity could be determined. It is only then that the potentially distorting effects of beliefs and values caused by a mental impairment would come to light.

I am not suggesting that suicide may never be a choice made by a person with capacity, but rather that in order to frame a judgement about whether a person has capacity with regard to this decision, an understanding of the contextual and historical background of that decision needs to be in place. In the case of Ms. Wooltorton, whilst she clearly understood and accepted the consequences of treatment refusal, her history of failed suicide attempts by the same method, her depressive state, refusal to engage with the treating clinicians to discuss her options or her reasons for her decision, the clinical relevance of her unstable personality disorder, and the fact that she was a well known and frequent user of mental health services all contribute to a broad picture that could provide an enriched understanding of her mental capacity at the time of admission and treatment refusal.

Treating judgements of capacity as a reductive exercise in ticking off criteria of mental functioning leads to an impoverished understanding of what it is to possess the capacity to make a particular decision and undermines the subtleties and complexities of clinical judgement. Whilst judgements of capacity are and will remain difficult, particularly in cases where the effects of psychiatric conditions are implicated in a

person's decision-making, I suggest that taking the contextual and holistic nature of reasons seriously does not diminish the objectivity and reliability of such judgements but in fact enhances it. If we seek to understand reasons holistically and not in empirical isolation from their relations and implications beyond the specific decision at hand, then we have the resources available to distinguish odd and perhaps idiosyncratic beliefs from those that may undermine capacity. Taking a narrow criterial approach to capacity assessment risks overlooking the very features of intentionality and decision-making that would enable these kinds of distinctions to be recognised.

Clinical Judgements and Expertise

The gesture towards a constructive account of the justifications for capacity judgements has thus far indicated that determining whether or not an individual has recognisable reasons for a decision is a matter of seeking to understand that decision in light of the normative structure of shared practice. I do not, however, wish to imply that capacity judgements are unskilled observations, the ability for which is simply a facet of one's natural ability to interpret intentional behaviour. The capacity for making such judgements is indeed intrinsic to participating in a shared form of life, but the ability to make good, reliable and accurate assessments of capacity is a skill that requires training, experience and the development of considerable expertise. This subtle but essential aspect of clinical training and practice is easily overlooked if one is focused solely on checking off the capacity criteria in assessing a person's decision-making as though they were indices of cognitive functioning (Silberfeld & Checkland, 1999).

Expertise is a matter of practical wisdom that ought not to be construed as inferior to rule-based judgements or arbitrary because its standards cannot be abstracted from the practice in which the pronouncements of expertise are made. This is not a new problem: Dreyfus cites the examples of Socrates being frustrated by the failure of Euthyphro, an expert on piety, to articulate the piety-recognition rule by which Socrates

presumed he must make his judgements. Instead Euthyphro could only provide perspicuous examples of applications of his expertise (Dreyfus, 1992, p.67). But rather than taking this feature of expert judgement to be a limitation, or inferior to the objective rigour of a checklist of criteria, the arguments I have presented here and in the previous chapter point to the conclusion that such judgement is in fact crucial to a good assessment, as it is flexible and broad in scope, able to take into account subtle, complex contextual factors, a rich and possibly inconsistent history, the mitigating effects of particular beliefs, values, compulsions or fixations and innumerable other factors that influence a person's decision-making process.

To give due consideration of how expertise ought to be defined or what the particular skills necessary to expert judgement would be would require significant further empirical and conceptual research, but the claim that capacity judgements do rely on what are fundamentally uncodified norms of judgement indicates that attempting to increase objectivity and reliability through a criteriological approach alone is a misguided strategy. This is not to suggest that codified criteria are dispensable in making assessments of capacity, but only that a simple reading of such criteria ought not to exhaust the normative standards implicated in what is intrinsically a complex judgement:

> "Providing they are interpreted intelligently, diagrammatic codifications can be helpful guides to practice. What they cannot do, however, is capture good clinical judgement independently of the tacit background. In themselves they remain merely partial and schematic codifications of practice" (Thornton, 2006).

Again there are close parallels with this view in debates within the philosophy of science: for Kuhn, the judgements of a trained expert scientific community constitute the best criterion of objectivity and rationality we have for scientific theories (Kuhn, 1977). Although experts may not always achieve consensus in judgements, if capacity assessments are to be more robust and capable of reliably distinguishing capacity from incapacity, developing the skills necessary for complex, context-laden judgements would equip clinicians with the best possible tools for the difficult task at hand.

Future Directions for Research

The purpose of capacity legislation is to ensure respect for individual autonomy as far as possible, whilst protecting the interests of those whose autonomy is impaired in relation to a particular decision. Autonomy is generally taken to refer to a person's ability to self-govern; to act and make choices freely, without interference or external influence (Christman, 2009). In bioethics the clearest use of the notion of autonomy is the requirement that informed consent is obtained prior to any medical intervention, which suggests that autonomy is the entitlement to self-determination with respect to the treatment one is subject to (Manson & O'Neill, 2007). Dworkin also predicates much of his influential ethical theory of autonomy on the ability to make decisions in a specific context (e.g., Dworkin, 1986) and other authors in bioethics suggest that *"competency be regarded solely as a function of the capacity for autonomous action"* (Silver, 2002, p.465). The concepts of capacity and autonomy are thus closely connected, although it is not clear to what extent they co-refer. Given the increasing prominence of capacity as a medico-legal concept and the broadly liberal political agenda in the UK, further research exploring the conceptual relationship between capacity and autonomy would be fruitful. It would be particularly interesting to consider the connection between capacity and autonomy in relation to the disparity between increasingly paternalistic mental health law that revokes the right to autonomy and the promotion of respect for choices that are made by individuals with capacity.

These questions are also significant for broader fields of study in which the concept of autonomy is variously employed. There is a wealth of literature in the field of ethics, particularly derived from attempts to naturalise a Kantian conception of autonomy, that focuses on the evaluative elements of autonomy, and the conceptual exploration of the epistemic standards of capacity conducted here could potentially intersect with these discussions. I have argued that a test of capacity is not a purely procedural but depends on substantive criteria, and so if the concepts of capacity and autonomy

overlap there may be implications for the extensive debates in ethics and political philosophy about whether autonomy is a value-laden or content-neutral concept.

In rejecting the reduction of capacity tests to sets of codified principles and checklist criteria I have emphasised the role of clinical expertise in assessing capacity, arguing that such expertise supplies the best kind of practical judgement possible for distinguishing capacity from incapacity. I have given a broad defence here of the importance of expertise in the face of a competing demand to secure the objectivity of assessments through the use of reductive criteria, but there is significant scope for research on the nature, bounds and justifications of expertise in clinical decision-making, particularly when such judgements are as ethically and politically loaded as they are in mental health care. One potential area of contact lies in discussions about the distinction between explicit and tacit knowledge, and the role that the latter, as a form of uncodifiable good judgement, plays in evidence based medicine (e.g., Thornton, 2006). Furthermore, the argument that clinical judgement does not derive its normative justifications from principles but rather from the uncodified rationality of practice may carry over into the field of medical ethics, in which judgements are typically based on the four cardinal principles set out in Beauchamp and Childress's influential textbook (2001, pt.II).

The introduction of capacity legislation has placed the individual's ability to make particular decisions centre stage in conceptual and practical issues about the legitimate reach of healthcare and medical paternalism. Hingeing upon the determinations of a capacity assessment is the potential to deny an individual a fundamental human right, freedom of choice, and the potential to permit life-threatening decisions to be carried out by individuals whose health and welfare it is incumbent upon society to protect, even from themselves. Questions about how we ought to treat patients lacking capacity are ethically, politically and emotively loaded, and it is likely that the ethical debate will intensify as capacity legislation is tested in the courts and the rights of the individual in

medical decision-making are further championed. It is therefore imperative that decisions made paternalistically on behalf of patients are only taken when necessary, that is, when they lack the capacity to make particular decisions for themselves. In spite of sophisticated attempts to codify the criteria of capacity judgements to ensure objectivity and reliability across clinicians, in this thesis I have demonstrated the limitations of a narrow cognitive approach to decision-making and argued that objectivity is not sacrificed if the rich personal and interpersonal context in which decisions are made is acknowledged.

References

Re T (Adult: Refusal of Treatment) [1992] 4 All E.R. 649.

B v Croydon Health Authority [1994a] 2 W.L.R. 294.

Re C (Adult: Refusal of Medical Treatment) [1994] 1 W.L.R. 290.

South West Hertfordshire Health Authority v KB [1994b] 2 F.C.R. 1051.

Norfolk & Norwich Healthcare (NHS) Trust v W [1996] 2 F.L.R. 613.

Re MB (Medical Treatment) [1997] 2 F.L.R. 426.

Mental Capacity Act 2005 c. 9. [online] http://www.opsi.gov.uk/acts/acts2005/pdf/ukpga_20050009_en.pdf.

Trust A and Trust B v H (An Adult Patient) [2006] 2 FLR 958.

Mental Health Act 2007 c. 12. [online] http://www.opsi.gov.uk/acts/acts2007/pdf/ukpga_20070012_en.pdf.

American Psychiatric Association (1980) *Diagnostic & Statistical Manual of Mental Disorders, 3rd Ed.*, Washington, DC: American Psychiatric Association.

American Psychiatric Association (2000) *Diagnostic & Statistical Manual of Mental Disorders, 4th Ed., Text Revision*, Washington, DC: American Psychiatric Association.

Appelbaum, P. S. & Grisso, T. (1995) 'The MacArthur Treatment Competence Study. I: Mental illness and competence to consent to treatment.' *Law & Human Behavior,* 19:2, 105-126.

Appelbaum, P. S. & Roth, L. H. (1982) 'Competency to consent to research. A psychiatric overview.' *Archives of General Psychiatry,* 39:8, 951-958.

Aronson, E. (1969) 'The theory of cognitive dissonance: a current perspective.' In *Advances in Experimental Social Psychology: Volume 4*, L. Berkowitz (Ed.), New York: Academic Press, pp. 1-34.

Ashton, G., Letts, P., Oates, L. & Terrell, M. (2006) *Mental Capacity: The New Law*, Jordan Publishing.

Audi, R. (2004) 'Theoretical rationality: its sources, structure and scope.' In *The Oxford Handbook of Rationality*, A. R. Mele and P. Rawlings (Eds.), Oxford: Oxford University Press, pp. 17-44.

Banner, N. F. & Thornton, T. (2007) 'The new philosophy of psychiatry: its (recent) past, present and future: a review of the Oxford University Press series International Perspectives in Philosophy and Psychiatry.' *Philosophy, Ethics & Humanities in Medicine,* 2:9.

Bayne, T. & Pacherie, E. (2004) 'Experience, belief and the interpretive fold.' *Philosophy, Psychiatry & Psychology,* 11:1, 81-86.

Bean, G., Nishisato, S., Rector, N. A. & Glancy, G. (1994) 'The psychometric properties of the Competency Interview Schedule.' *Canadian Journal of Psychiatry,* 39:8, 368-376.

Beauchamp, T. L. & Childress, J. F. (2001) *Principles of Biomedical Ethics*, Oxford: Oxford University Press.

Bentall, R. P., Corcoran, R., Howard, R., Blackwood, N. & Kinderman, P. (2001) 'Persecutory delusions: A review and theoretical integration.' *Clinical Psychology Review,* 21:8, 1143-1192.

Bermúdez, J. L. (2001) 'Normativity and Rationality in Delusional Psychiatric Disorders.' *Mind & Language,* 16:5, 457-493.

Bermúdez, J. L. (2005) *Philosophy of Psychology: A Contemporary Introduction*, Routledge.

Bermúdez, J. L. (2009) *Decision Theory and Rationality*, Oxford: Oxford University Press.

Berrios, G. (1991) 'Delusions as "wrong beliefs": a conceptual history.' *British Journal of Psychiatry,* 159: 6-13.

Bielby, P. (2005) 'The conflation of competence and capacity in English medical law: A philosophical critique.' *Medicine, Healthcare and Philosophy,* 8:3, 357-369.

Block, N. (1987) 'Functional role and truth conditions.' *Proceedings of the Aristotelian Society,* LXI: 157-181.

Bloor, D. (1973) 'Wittgenstein and Mannheim on the Sociology of Mathematics.' *Studies in History and Philosophy of Science,* 4:2, 173-191.

Bloor, D. (1983) *Wittgenstein: A Social Theory of Knowledge*, London.

Bolton, D. & Hill, J. (2004) *Mind, Meaning and Mental Disorder*, Oxford: Oxford University Press.

Bortolotti, L. (2003) 'Inconsistency and interpretation.' *Philosophical Explorations,* 6:2, 109-123.

Bortolotti, L. (2004a) 'Can we interpret irrational behaviour?' *Behavior & Philosophy,* 32: 359-375.

Bortolotti, L. (2004b) 'Intentionality without rationality.' *Proceedings of the Aristotelian Society,* CV:3, 385-392.

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

Bortolotti, L. & Broome, M. R. (2008) 'Delusional beliefs and reason giving.' *Philosophical Psychology,* 21:6, 821-841.

Braine, M. D. S., Reiser, B. J., & Rumain, B. (1984) 'Some empirical justification for a theory of natural propositional logic.' In *The Psychology of Learning and Motivation*, G. H. Bower (Ed.), New York: Academic Press.

Brannan, S., Davies, M., English, V., Mussell, R., Sheather, J., Chrispin, E. & Sommerville, A. (2010) 'Ethics briefings.' *Journal of Medical Ethics,* 36: 63-64.

Breden, T. M. & Vollmann, J. (2004) 'The Cognitive Based Approach of Capacity Assessment in Psychiatry: A Philosophical Critique of the MacCAT-T.' *Health Care Analysis,* 12:4, 273-283.

British Medical Association & The Law Society (2004) *Assessment of Mental Capacity: Guidance for Doctors and Lawyers*, British Medical Association.

Broome, J. (1997) 'Reason and motivation.' *Proceedings of the Aristotelian Society,* Suppl. Vol. 71: 131-146.

Buchanan, A. & Brock, D. W. (1986) 'Deciding for others.' *Millbank Quarterly,* 64:2, 17-94.

Cairns, R., Maddock, C., Buchanan, A., David, A. S., Hayward, P., Richardson, G., Szmukler, G. & Hotopf, M. (2005a) 'Prevalence and predictors of mental incapacity in psychiatric in-patients.' *British Journal of Psychiatry,* 187: 379-385.

Cairns, R., Maddock, C., Buchanan, A., David, A. S., Hayward, P., Richardson, G., Szmukler, G. & Hotopf, M. (2005b) 'Reliability of mental capacity assessments in psychiatric in-patients.' *British Journal of Psychiatry,* 187:4, 372-378.

Campbell, J. (2001) 'Rationality, meaning, and the analysis of delusion.' *Philosophy, Psychiatry, and Psychology,* 8:2, 89-100.

Campbell, J. (2007) 'An interventionist approach to causation in psychology.' In *Causal Learning: Psychology, Philosophy, & Computation*, A. Gopnik and L. Schulz (Eds.), New York: Oxford University Press, pp. 58-66.

Campbell, J. (2009) 'What does rationality have to do with psychological causation? Propositional attitudes as mechanisms and as control variables.' In *Psychiatry As Cognitive Neuroscience: Philosophical Perspectives*, M. R. Broome and L. Bortolotti (Eds.), Oxford: Oxford University Press, pp. 137-149.

Carpenter, W. T., Gold, J. M., Lahti, A. C., Queern, C. A., Conley, R. R., Barkto, J. J., Kovnick, J. & Appelbaum, P. S. (2000) 'Decisional capacity for informed consent in schizophrenia research.' *Archives of General Psychiatry,* 57: 533-538.

Carroll, L. (1895) 'What the Tortoise Said to Achilles.' *Mind,* IV:14, 278-280.

Charland, L. C. (2006) 'Anorexia and the MacCAT-T test for mental competence: validity, value, and emotion.' *Philosophy, Psychiatry & Psychology,* 13:4, 283-287.

Charland, L. C. "Decision-making capacity" *The Stanford Encyclopedia of Philosophy* (Spring 2008 Edition), Edward N. Zalta (Ed.) [online] http://plato.stanford.edu/archives/spr2008/entries/decision-capacity/.

Charland, L. C. (2001) 'Mental competence and value: The problem of normativity in the assessment of decision-making capacity.' *Psychiatry, Psychology and Law,* 8:2, 135-145.

Cherniak, C. (1981) 'Minimal Rationality.' *Mind,* XC: 161-183.

Child, W. (1993) 'Anomalism, uncodifiability and psychophysical relations.' *The Philosophical Review,* 102:2, 215-245.

Child, W. (1996a) 'Anomalism, rationality, and psychophysical relations.' In *Causality, Interpretation & the Mind*, Oxford: Oxford University Press, pp. 56-90.

Child, W. (1996b) *Causality, Interpretation, and the Mind*, Oxford: Oxford University Press.

Chomsky, N. (1986) *Knowledge of Language: Its Nature, Origin and Use*, NY: Praeger.

Christman, J. "Autonomy in moral and political philosophy" *The Stanford Encyclopedia of Philosophy* (Fall 2009 Edition), Edward N. Zalta (Ed.) [online] http://plato.stanford.edu/archives/fall2009/entries/autonomy-moral.

European Convention for the Protection of Human Rights and Fundamental Freedoms 1998 Protocol No. 11.

Crary, A. (2000) 'Introduction.' In *The New Wittgenstein*, A. Crary and R. Read (Eds.), London: Routledge, pp. 1-18.

Culver, C. M. & Gert, B. (2004) 'Competence.' In *The Philosophy of Psychiatry: A Companion*, Jennifer Radden (Ed.), Oxford: Oxford University Press, pp. 258-271.

Danielson, P. (2004) 'Rationality & Evolution.' In *The Oxford Handbook of Rationality*, A. R. Mele and P. Rawlings (Eds.), Oxford: Oxford University Press, pp. 417-439.

Davidson, D. (1967) 'Truth and Meaning.' In *Inquiries*, (2001b) pp. 17-42.

Davidson, D. (1968) 'On saying that.' In *Inquiries*, (2001b) pp. 93-108.

Davidson, D. (1969) 'True to the facts.' In *Inquiries*, (2001b) pp. 43-54.

Davidson, D. (1970a) 'Mental events.' In *Essays*, (2001a) pp. 207-224.

Davidson, D. (1970b) 'Semantics for natural languages.' In *Inquiries*, (2001b) pp. 55-64.

Davidson, D. (1973a) 'Psychology as Philosophy.' In *Essays*, (2001a) pp. 229-245.

Davidson, D. (1973b) 'Radical Interpretation.' In *Inquiries*, (2001b) pp. 125-140.

Davidson, D. (1974a) 'Belief and the basis of meaning.' In *Inquiries*, (2001b) pp. 141-154.

Davidson, D. (1974b) 'On the very idea of a conceptual scheme.' In *Inquiries*, (2001b) pp. 183-198.

Davidson, D. (1974c) 'Replies to David Lewis & W.V. Quine.' *Synthese,* 27: 345-349.

Davidson, D. (1975) 'Thought and talk.' In *Inquiries*, (2001b) pp. 155-170.

Davidson, D. (1977) 'Reality without reference.' In *Inquiries*, (2001b) pp. 215-226.

Davidson, D. (1978) 'Intending.' In *Essays*, (2001a) pp. 83-102.

Davidson, D. (1982) 'Paradoxes of irrationality.' In *Problems of Rationality*, (2004) pp. 169-188.

Davidson, D. (1983) 'A Coherence theory of truth and knowledge.' In *Subjective, Intersubjective, Objective*, (2001) pp. 137-153.

Davidson, D. (1985a) 'Incoherence and irrationality.' In *Problems of Rationality*, (2004) pp. 189-198.

Davidson, D. (1985b) 'Replies to Essays X-XII.' In *Essays on Davidson: Actions and Events*, B. Vermazen and M. Hintikka (Eds.), Oxford: Clarendon Press, pp. 242-252.

Davidson, D. (1986a) 'A nice derangement of epitaphs.' In *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, 1989 E. LePore (Ed.), Oxford: Basil Blackwell, pp. 433-446.

Davidson, D. (1986b) 'Deception and division.' In *Problems of Rationality*, (2004) pp. 199-212.

Davidson, D. (1990) 'The structure and content of truth.' *The Journal of Philosophy,* 87:6, 279-328.

Davidson, D. (1991) 'Three varieties of knowledge.' In *Subjective, Intersubjective, Objective*, (2001) pp. 205-220.

Davidson, D. (1994) 'Radical Interpretation interpreted.' *Philosophical Perspectives,* 8, Logic and Language: 121-128.

Davidson, D. (1995a) 'Could there be a science of rationality?' In *Problems of Rationality*, (2004) pp. 117-134.

Davidson, D. (1995b) 'The objectivity of values.' In *Problems of Rationality*, (2004) pp. 39-52.

Davidson, D. (1997) 'The emergence of thought.' In *Subjective, Intersubjective, Objective*, (2001) pp. 123-134.

Davidson, D. (2001a) *Essays on Actions and Events*, Oxford: Oxford University Press.

Davidson, D. (2001b) *Inquiries Into Truth & Interpretation*, Oxford: Oxford University Press.

Davidson, D. (2001) *Subjective, Intersubjective, Objective*, Oxford: Oxford University Press.

Davidson, D. (2003) 'Responses to Barry Stroud, John McDowell, and Tyler Burge.' *Philosophy & Phenomenological Research,* 67:3, 691-699.

Davidson, D. (2004) *Problems of Rationality*, Oxford: Oxford University Press.

Davies, M., Coltheart, M., Langdon, R. & Breen, N. (2001) 'Monothematic delusions: towards a two-factor account.' *Philosophy, Psychiatry & Psychology,* 8:2-3, 133-158.

Dawson, J. & Szmukler, G. (2006) 'Fusion of mental health and incapacity legislation.' *British Journal of Psychiatry,* 188: 504-509.

Dennett, D. C. (1987) *The Intentional Stance*, Cambridge, MA.: MIT Press.

Department of Constitutional Affairs (2007) *Mental Capacity Act 2005: Code of Practice*, London: TSO.

Dodd, J. (1999) 'There is no norm of truth: a minimalist reply to Wright.' *Analysis,* 59:264, 291-299.

Dreyfus, H. L. (1992) *What Computers Still Can't Do: A Critique of Artifical Reason*, MIT Press.

Dreyfus, H. L. & Dreyfus, S. E. (1986) *Mind Over Machine: the Power of Human Intuition & Expertise in the Era of the Computer*, Mew York: MacMillan.

Dummett, M. (1959) 'Wittgenstein's Philosophy of Mathematics.' *Philosophical Review,* 68: 324-348.

Dummett, M. (1993) 'What is a theory of meaning?' In *The Seas of Language*, Oxford: Oxford Universoty Press, pp. 1-33.

Dworkin, R. (1986) 'Autonomy and the demented self.' *The Millbank Quarterly,* 64:Suppl. 2, 4-16.

Dyer, C. (2009) 'Coroner rules that treating 26 year old woman who wanted to die would have been unlawful.' *British Medical Journal,* 339: b4070.

Elliot, A. J. & Devine, P. G. (1994) 'On the motivational nature of cognitive dissonance: dissonance as a psychological discomfort.' *Journal of Personality & Social Psychology,* 67: 382-394.

Elster, J. (1984) 'The nature and scope of rational-choice explanation.' In *Actions & Events: Perspectives on the Philosophy of Donald Davidson*, E. LePore and B. P. McLaughlin (Eds.), Oxford: Blackwell, pp. 60-72.

Evans, J. & Over, D. E. (1996) *Rationality & Reasoning*, Psychology Press.

Evnine, S. (1991) *Donald Davidson*, Stanford University Press.

Eysenck, M. W. & Keane, M. T. (2005) 'Reasoning and deduction.' In *Cognitive Psychology: A Student's Handbook, 5th Edition*, Psychology Press, pp. 506-532.

Fennell, P. (2007) *Mental Health: The New Law*, Bristol: Jordan Publishing.

Finkelstein, D. H. (2000) 'Wittgenstein on rules and platonism.' In *The New Wittgenstein*, A. Crary and R. Read (Eds.), London: Routledge, pp. 53-73.

Flanagan, O. (1984) *The Science of the Mind*, MIT Press.

Fodor, J. & Lepore, E. (1992) *Holism: A Shopper's Guide*, Oxford: Blackwell.

Foley, R. (1993) *Working Without a Net*, Oxford: Oxford University Press.

Føllesdal, D. (1984) 'Causation and explanation: A problem in Davidson's view on action and mind.' In *Actions & Events: Perspectives on the Philosophy of Donald Davidson*, E. LePore and B. P. McLaughlin (Eds.), Oxford: Blackwell, pp. 311-323.

Frankfurt, H. (1977) 'Freedom of the will and the concept of the person.' *Journal of Philosophy,* 68:1, 5-20.

Freedman, B. (1981) 'Competence, marginal and otherwise: concepts and ethics.' *Int.J.Law Psychiatry,* 4:1-2, 53-72.

Garety, P., Hemsley, D. & Wessely, S. (1991) 'Reasoning in deluded schizophrenic and paranoid patients: biases in performance on a probablistic inference task.' *Journal of Nervous & Mental Disorder,* 179: 194-201.

Gerrans, P. (2004) 'Cognitive architecture and the limits of Interpretationism.' *Philosophy, Psychiatry & Psychology,* 11:1, 43-48.

Gigerenzer, G. (2006) 'Bounded and rational.' In *Contemporary Debates in Cognitive Science*, R. J. Stainton (Ed.), Oxford: Blackwell, pp. 115-133.

Gigerenzer, G., Todd, P. M. & ABC Research Group (1999) *Simple Heuristics That Make Us Smart*, New York: Oxford University Press.

Gigerenzer, G. & Selten, R. (2001) 'Rethinking rationality.' In *Bounded Rationality: The Adaptive Toolbox*, Gerd Gigerenzer and Reinhard Selten (Eds.), MIT Press, pp. 1-12.

Glock, H.-J. (2003) *Quine & Davidson on Language, Thought & Reality*, Cambridge University Press.

Gold, I. & Howhy, J. (2000) 'Rationality and Schizophrenic Delusion.' In *Pathologies of Belief*, M. Coltheart and M. Davies (Eds.), Oxford: Blackwell, pp. 145-166.

Goldman, A. I. (1989) 'Interpretation psychologized.' *Mind & Language,* 4: 161-185.

Grandy, R. (1973) 'Reference, meaning and belief.' *Journal of Philosophy,* 70: 439-452.

Grice, P. (1957) 'Meaning.' *The Philosophical Review,* 66: 377-388.

Grisso, T. & Appelbaum, P. S. (1995a) 'Comparison of standards for assessing patients' capacities to make treatment decisions.' *American Journal of Psychiatry,* 152:7, 1033-1037.

Grisso, T. & Appelbaum, P. S. (1995b) 'The MacArthur treatment competence study. III: Abilities of patients to consent to psychiatric and medical treatments.' *Law & Human Behavior,* 19:2, 149-174.

Grisso, T., Appelbaum, P. S. & Hill-Fotouhi, C. (1997) 'The MacCAT-T: A clinical tool to assess patients' capacities to make treatment decisions.' *Psychiatric Services,* 48:11, 1415-1419.

Grisso, T., Appelbaum, P. S., Mulvey, E. P. & Fletcher, K. (1995) 'The MacArthur treatment competence study. II: Measures of abilities related to competence to consent to treatment.' *Law & Human Behavior,* 19:2, 127-148.

Grubb, A. (2004) 'Consent ot Treatment: Competent Patient.' In *Principles of Medical Law*, A. Grubb and J. Laing (Eds.), Oxford University Press.

Grubb, A. & Laing, J. (2004) *Principles of Medical Law*, Oxford: Oxford University Press.

Gunn, M. (1994) 'The meaning of incapacity.' *Medical Law Review,* 2: 8-29.

Gunn, M. J., Wong, J. G., Clare, I. C. H. & Holland, A. J. (1999) 'Decision-Making Capacity.' *Medical Law Review,* 7: 269-306.

Harman, G. (1999) 'Rationality.' In *Reasoning, Meaning & Mind*, G. Harman (Ed.), Oxford: Oxford University Press, pp. 10-45.

Harman, G. (2004) 'Practical aspects of theoretical reasoning.' In *The Oxford Handbook of Rationality*, A. R. Mele and P. Rawlings (Eds.), Oxford: Oxford University Press, pp. 45-56.

Hattiangadi, A. (2006) 'Is Meaning Normative?' *Mind & Language,* 21:2, 220-240.

Heal, J. (1998) 'Understanding Other Minds from the Inside.' In *Current Issues in Philosophy of Mind*, A. O'Hear (Ed.), Cambridge University Press, pp. 83-100.

Heal, J. (2008) ''Back To The Rough Ground!' Wittgensteinian Reflections on Rationality and Reason.' In *Wittgenstein and Reason*, J. Preston (Ed.), Oxford: Blackwell, pp. 47-64.

Heil, J. (1989) 'Minds divided.' *Mind,* XCVIII:392, 571-583.

Henderson, D. K. (1987) 'A solution to Davidson's paradox of irrationality.' *Erkenntnis,* 27:3, 359-369.

Henderson, D. K. (1991) 'Rationalizing explanation, normative principles, and descriptive generalizations.' *Behavior & Philosophy,* 19:1, 1-20.

Higgs, R. (2004) 'The contribution of narrative ethics to issues of capacity in psychiatry.' *Health Care Analysis,* 12:4, 307-316.

Hollis, M. (1994) 'Rationality and relativism.' In *The Philosophy of Social Science: An Introduction*, Cambridge: Cambridge University Press, pp. 224-247.

Holroyd, J. (2010) 'Clarifying capacity: values and reasons'. *Unpublished draft*.

Hopkins, J. (2004) 'Wittgenstein, Davidson, and the methodology of interpretation'. *Unpublished draft*.

Horwich, P. (2005) *Reflections on Meaning*, Oxford: Oxford University Press.

Hotopf, M. (2005) 'The assessment of mental capacity.' *Clinical Medicine,* 5:6, 580-584.

Hunt, G. (1990) 'Schizophrenia and indeterminacy: the problem of validity.' *Theoretical Medicine,* 11: 61-78.

Hurley, S. & Nudds, M. (2006) *Rational Animals?*, Oxford: Oxford University Press.

Jackman, H. (2003) 'Charity, self interpretation and belief.' *Journal of Philosophical Research,* 28: 145-170.

Janofsky, J. S., McCarthy, R. J. & Folstein, M. F. (1992) 'The Hopkins Competency Assessment Test: A brief method for evaluating patients' capacity to give informed consent.' *Hospital & Community Psychiatry,* 43:2, 132-136.

Johnson-Laird, P. N., Legrenzi, P. & Legrenzi, M. S. (1972) 'Reasoning and a sense of reality.' *The British Journal of Psychology,* 63: 395-400.

Jones, R. (2005a) *Mental Capacity Act Manual*, Andover: Sweet & Maxwell.

Jones, R. (2005b) 'Review of the Mental Capacity Act 2005.' *Psychiatric Bulletin,* 29: 423-427.

Joseph, M. (2004) *Donald Davidson*, Acumen.

Kacelnik, A. (2006) 'Meanings of Rationality.' In *Rational Animals?*, S. Hurley and M. Nudds (Eds.), Oxford: Oxford University Press.

Kahnemann, D. & Tversky, A. (1972) 'Subjective probability: a judgment of representativeness.' *Cognitive Psychology,* 3: 430-454.

Kemp, R., Chua, S., McKenna, P. & David, A. S. (1997) 'Reasoning and delusions.' *British Journal of Psychiatry,* 170: 398-405.

Kennedy, I. (1997) 'Commentary on Re MB (Medical Treatment).' *Medical Law Review,* 5: 317-353.

Kim, S. Y. H. (2006) 'When does decisional impairment become decisional incompetence? Ethical and methodological issues in capacity research in schizophrenia.' *Schizophrenia Bulletin,* 32:1, 92-97.

Kim, S. Y. H., Caine, E. D., Currier, G. W., Leibovici, A. & Ryan, J. M. (2001) 'Assessing the competence of persons with Alzheimer's disease in providing informed consent for participation in research.' *American Journal of Psychiatry,* 158:5, 712-717.

Kim, S. Y. H., Karlawish, J. H. T. & Caine, E. D. (2002) 'Current state of research on decision-making competence of cognitively impaired elderly persons.' *American Journal of Geriatric Psychiatry,* 10:2, 151-165.

Kitamura, F., Tomoda, A., Tsukada, K., Tanaka, M., Kawakami, I., Mishima, S. & Kitamura, T. (1998) 'Method for assessment of competency to consent in the mentally ill: Rationale, development, and comparison with the medically ill.' *International Journal of Law & Psychiatry,* 21:3, 223-244.

Korsgaard, C. M. (1996) *The Sources of Normativity*, Cambridge: Cambridge University Press.

Kripke, S. (1982) *Wittgenstein on Rules and Private Language*, Oxford: Basil Blackwell.

Kuhn, T. (1970a) *The Structure of Scientific Revolutions, 2nd Ed.*, Chicago: University of Chicago Press.

Kuhn, T. S. (1970b) 'Reflections on my critics.' In *Criticism & the Growth of Knowledge*, I. Lakatos and A. Musgrave (Eds.), Cambridge: Cambridge University Press, pp. 231-278.

Kuhn, T. S. (1977) 'Objectivity, value judgment and theory choice.' In *The Essential Tension*, Chicago: University of Chicago Press, pp. 320-339.

Kupfer, D. J., First, M. B. & Reggier, D. A. (2002) *A Research Agenda for DSM-V*, American Psychiatric Association.

Kusch, M. (2006) *A Sceptical Guide to Meaning and Rules: Defending Kripke's Wittgenstein*, Chesham, Acumen.

Laing, R. D. (1967) *Politics of Experience*, Harmondsworth: Penguin.

Langdon, R. & Coltheart, M. (2000) 'The cognitive neuropsychology of delusions.' In *Pathologies of Belief*, M. Davies and M. Coltheart (Eds.), Oxford: Blackwell, pp. 183-215.

Law Commission (1991) No. 119. *Mentally Incapacitated Adults and Decision-Making: An Overview*, London: HMSO.

Law Commission (1993) No. 128. *Mentally Incapacitated Adults and Decision-Making: A New Jurisdiction*, London: HMSO.

Law Commission (1993) No. 129. *Mentally Incapacitated Adults and Decision-Making: Medical Treatment and Research*, London: HMSO.

Law Commission (1995) Report No. 231. *Mental Incapacity*, London: HMSO.

Lear, J. (1982) 'Leaving the world alone.' *The Journal of Philosophy,* 79:7, 382-403.

Lear, J. (1986) 'Transcendental Anthropology.' In *Subject, Thought & Context*, J. McDowell and P. Pettit (Eds.), Oxford: Clarendon Press, pp. 267-298.

Lear, J. & Stroud, B. (1984) 'The Disappearing 'We'.' *Proceedings of the Aristotelian Society, Supplementary Volumes,* 58: 219-258.

Leeser, J. & O'Donohue, W. (1999) 'What is a delusion? Epistemological dimensions.' *Journal of Abnormal Psychology,* 108:4, 687-694.

Lepore, E. & Ludwig, K. (2005) *Donald Davidson: Meaning, Truth, Language & Reality*, Oxford: Oxford University Press.

LePore, E. & Ludwig, K. (2006) 'Introduction.' In *The Essential Davidson*, E. LePore and K. Ludwig (Eds.), Oxford: Oxford University Press, pp. 1-22.

Lévy-Bruhl, L. (1923) *Primitive Mentality (Trans. L.A. Clare of La Mentalité Primitive (Paris, 1922)*, London: Allen & Unwin.

Lloyd, G. E. R. (2007) *Cognitive Variations: Reflections on the Unity and Diversity of the Human Mind*, Oxford: Oxford University Press.

Ludwig, K. (2004) 'Rationality, Language and the Principle of Charity.' In *The Oxford Handbook of Rationality*, A. R. Mele and P. Rawlings (Eds.), Oxford: Oxford University Press, pp. 343-362.

Luntley, M. (2003) 'Rules and other people.' In *Wittgenstein: Meaning & Judgement*, Oxford: Blackwell, pp. 93-123.

Luntley, M. (2006) 'Keeping track, autobiography, and the conditions for self-erosion.' In *Dementia: Mind, Meaning, and the Person*, J. C. Hughes, S. J. Louw, and S. R. Sabat (Eds.), Oxford: Oxford University Press, pp. 105-122.

MacLean, S. (2009) 'Live and let die.' *British Medical Journal,* 339: b4112.

Maher, B. A. (1999) 'Anomalous experience in everyday life: its significance for psychopathology.' *The Monist,* 82: 547-570.

Malpas, J. E. (1992) *Donald Davidson & the Mirror of Meaning: Holism, Truth, Interpretation*, Cambridge: Cambridge University Press.

Manson, N. C. & O'Neill, O. (2007) *Rethinking Informed Consent in Bioethics*, Cambridge: Cambridge University Press.

Martin, A. M. (2007) 'Tales Publicly Allowed: Competence, Capacity and Religious Beliefs.' *Hastings Center Report,* 37:1, 33-40.

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

McDowell, J. (1984a) 'Functionalism and Anomalous Monism.' In *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, E. LePore and B. P. McLaughlin (Eds.), Oxford: Basil Blackwell, pp. 387-398.

McDowell, J. (1984b) 'Wittgenstein on Following a Rule.' *Synthese,* 58: 325-363.

McDowell, J. (1986) 'Singular Thought and the Extent of Inner Space.' In *Subject, Thought and Context*, Oxford: Clarendon Press, pp. 137-168.

McDowell, J. (1994) *Mind and World*, Harvard University Press.

McDowell, J. (1998) *Mind, Value, and Reality*, President and Fellows of Harvard College.

McGinn, M. (1997) *Wittgenstein & the Philosophical Investigations*, London: Routledge.

McLaughlin, B. P. (1985) 'Anomalous Monism and the Irreducibility of the Mental.' In *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, E. LePore and B. P. McLaughlin (Eds.), Oxford: Blackwell, pp. 331-368.

McLeod, P., Plunkett, K., & Rolls, E. T. (2006) 'The attraction of Parallel Distributed Processing for modelling cognition.' In *Philosophy of Psychology: Contemporary Readings*, J. L. Bermúdez (Ed.), Routledge, pp. 182-202.

Mele, A. R. & Rawling, P. (2004) *The Oxford Handbook of Rationality*, Alfred R. Mele and Piers Rawling (Eds.), Oxford: Oxford University Press.

Merleau-Ponty, M. (1962) *The Phenomenology of Perception, Trans. Colin Smith*, London: Routledge & Kegan Paul.

Millar, A. (2004) *Understanding People: Normativity and Rationalizing Explanation*, Oxford: Oxford University Press.

Miller, A. (1998) *Philosophy of Language*, Routledge.

Mišcevic N. (2000) *Rationality & Cognition: Against Relativism-Pragmatism*, University of Toronto Press.

Moran, R. (2001) *Authority & Estrangement: An Essay on Self-Knowledge*, Princeton University Press.

Mukherjee, S. & Shah, A. (2001) 'The prevalence and correlates of capacity to consent to a geriatric psychiatry admission.' *Aging and Mental Health,* 5:4, 335-339.

Mulhall, S. (1987) 'Davidson on interpretation and understanding.' *The Philosophical Quarterly,* 37:148, 319-322.

Nozick, R. (1993) *The Nature of Rationality*, New Jersey: Princeton University Press.

Nygaard, H. A., Naik, M. & Ruths, S. (2000) 'Mental impairment in nursing home residents.' *Tidsskr.Nor Laegeforen.,* 120:26, 3113-3116.

Nys, H., Welie, S., Garanis-Papadatos, T. & Ploumpidis, D. (2004) 'Patient capacity in mental health care: Legal overview.' *Health Care Analysis,* 12:4, 329-337.

Oaksford, M. & Chater, N. (2001) 'The probabilistic approach to human reasoning.' *Trends in Cognitive Sciences,* 5:8, 349-357.

Okai, D., Owen, G., McGuire, H., Singh, S., Churchill, R. & Hotopf, M. (2007) 'Mental capacity in psychiatric patients: systematic review.' *British Journal of Psychiatry,* 191: 291-297.

Owen, G. S., Cutting, J. & David, A. S. (2007) 'Are people with schizophrenia more logical than healthy volunteers?' *British Journal of Psychiatry,* 191: 453-454.

Owen, G. S., David, A. S., Richardson, G., Szmukler, G., Hayward, P. & Hotopf, M. (2009a) 'Mental capacity, diagnosis and insight in psychiatric in-patients: a cross-sectional study.' *Psychological Medicine,* 39: 1389-1398.

Owen, G. S., Freyenhagen, F., Richardson, G. & Hotopf, M. (2009b) 'Mental capacity and decisional autonomy: an interdisciplinary challenge.' *Inquiry,* 52:1, 79-107.

Pettit, P. (1990) 'The Reality of Rule-Following.' *Mind,* 99: 1-21.

Pettit, P. & McDowell, J. (1986) 'Introduction.' In *Subject, Thought and Context*, P. Pettit and J. McDowell (Eds.), Oxford: Clarendon Press.

Pettit, P. & Smith, M. (1990) 'Backgrounding Desire.' *The Philosophical Review,* 99:4, 565-592.

Quine, W. V. O. (1960) *Word and Object*, Cambridge, MA.: MIT Press.

Radden, J. (1985) *Madness and Reason*, London: Allen and Unwin.

Ramberg, B. T. (1989) *Donald Davidson's Philosophy of Language: An Introduction*, Cambridge, MA: Blackwell.

Raymont, V. (2002) ''Not in perfect mind' - the complexity of clinical capacity assessment.' *Psychiatric Bulletin,* 26: 201-204.

Raymont, V., Bingley, W., Buchanan, A., David, A. S., Hayward, P., Wessely, S. & Hotopf, M. (2004) 'Prevalence of mental incapacity in medical inpatients and associated risk factors: cross-sectional study.' *Lancet,* 364:9443, 1421-1427.

Raz, J. (1999) 'Explaining normativity: on rationality and the justification of reason.' *Ratio (new series),* 12: 354-379.

Rhodes, J. & Gipps, R. G. T. (2008) 'Delusions, certainty, and the Background.' *Philosophy, Psychiatry & Psychology,* 15:4, 295-310.

Richards, S. & Mughal, A. F. (2006) *Working With the Mental Capacity Act 2005*, Hampshire: Matrix Training Associates.

Rosenberg, J. "Wilfrid Sellars" *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (Ed.) [online] http://plato.stanford.edu/archives/fall2008/entries/sellars/.

Rudnick, A. (2002) 'Depression and competence to refuse psychiatric treatment.' *Journal of Medical Ethics,* 28: 151-155.

Sackett, D. L., Haynes, R. B., Guyatt, G. H. & Tugwell, P. (1991) *Clinical Epidemiology: A Basic Science for Clinical Medicine*, Boston, MA: Little, Brown & Co.

Sass, L. A. (1994) *The Paradoxes of Delusion: Wittgenstein, Schreber and the Schizophrenic Mind*, Cornell University Press.

Schulte, J. (2008) 'Rules and Reason.' In *Wittgenstein and Reason*, J. Preston (Ed.), Oxford: Blackwell, pp. 107-122.

Searle, J. R. (2001) 'The Classical Model of Rationality and its weaknesses.' In *Rationality In Action*, MIT Press, pp. 1-33.

Sesardic, N. (1986) 'Psychology without Principle of Charity.' *Dialectica,* 40:3, 229-240.

Silberfeld, M. (1994) 'Evaluating decisions in mental capacity assessments.' *International Journal of Geriatric Psychiatry,* 9:5, 365-371.

Silberfeld, M. & Checkland, D. (1999) 'Faulty judgment, expert opinion, and decision-making capacity.' *Theoretical Medicine & Bioethics,* 20:4, 377-393.

Silver, M. (2002) 'Reflections on determining competency.' *Bioethics,* 16:5, 455-468.

Smith, M. (2004) 'Humean Rationality.' In *The Oxford Handbook of Rationality*, A. R. Mele and P. Rawlings (Eds.), Oxford: Oxford University Press, pp. 75-92.

Spitzer, M. (1990) 'On defining delusions.' *Comprehensive Psychiatry,* 31:5, 377-397.

Stauch, M., Wheat, K. & Tingle, J. (2006) *Text, Cases and Materials on Medical Law*, New York: Routledge Cavendish.

Stein, E. (1996) *Without Good Reason: The Rationality Debate in Philosophy & Cognitive Science*, Oxford: Clarendon Press.

Stich, S. (1983) *From Folk Psychology to Cognitive Science: The Case Against Belief*, Cambridge, Massachusetts: Bradford Books, MIT Press.

Stich, S. P. (1999) 'Dennett on Intentional Systems.' In *Mind & Cognition: An Anthology*, W. G. Lycan (Ed.), Blackwell, pp. 87-100.

Stone, T. & Young, A. W. (1997) 'Delusions and brain injury: The philosophy and psychology of belief.' *Mind & Language,* 12:3-4, 327-364.

Stroud, B. (1979) 'Inference, Belief, and Understanding.' *Mind,* LXXXVIII:1, 179-196.

Tan, J. & Hope, T. (2008) 'Treatment refusal in anorexia: a challenge to current concepts of capacity.' In *Empirical Ethics in Psychiatry*, Widdershoven G., T. Hope, and J. McMillan (Eds.), Oxford: Oxford University Press, pp. 187-211.

Tan, J., Hope, T. & Stewart, A. (2003) 'Competence to refuse treatment in anorexia nervosa.' *International Journal of Law & Psychiatry,* 26:6, 697-707.

Tan, J. O. A., Stewart, A., Fitzpatrick, R. & Hope, T. (2006) 'Competence to make treatment decisions in anorexia nervosa: thinking processes and values.' *Philosophy, Psychiatry & Psychology,* 13:4, 267-282.

Taylor, C. (2002) 'Foundationalism and the inner-outer distinction.' In *Reading McDowell: On Mind & World*, N. H. Smith (Ed.), New York: Routledge, pp. 106-119.

Thornton, T. (1997) 'Reasons and Causes in Philosophy and Psychopathology.' *Philosophy, Psychiatry & Psychology,* 4:4, 307-317.

Thornton, T. (2005) *John McDowell*, Acumen.

Thornton, T. (2006) 'Tacit knowledge as the unifying factor in evidence based medicine and clinical judgement.' *Philosophy, Ethics & Humanities in Medicine,* 1:2.

Tomoda, A., Yasumiya, R., Sumiyama, T., Tsukada, K., Hayakawa, T., Matsubara, K., Kitamura, F. & Kitamura, T. (1997) 'Validity and reliability of structured interview for competency incompetency assessment testing and ranking inventory.' *Journal of Clinical Psychology,* 53:5, 443-450.

Tversky, A. & Kahneman, D. (1983) 'Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment.' *Psychological Review,* 91: 293-315.

Tversky, A. & Kahnemann, D. (1974) 'Judgement under uncertainty: heuristics and biases.' *Science,* 185: 1124-1131.

Van Staden, C. W. & Kruger, C. (2003) 'Incapacity to give informed consent owing to mental disorder.' *Journal of Medical Ethics,* 29:1, 41-43.

Vellinga, A., Smit, J. H., van Leeuwen, E., van Tilburg, W. & Jonker, C. (2004) 'Instruments to assess decision-making capacity: An overview.' *International Psychogeriatrics,* 16:4, 397-419.

Viglione, V., Muratori, F., Maestro, S., Brunori, E. & Picchi, L. (2006) 'Denial of symptoms and psychopathology in adolescent anorexia nervosa.' *Psychopathology,* 39:5, 255-260.

Vollman, J., Bauer, A., Danker-Hopfe, H. & Helmchen, H. (2003) 'Competence of mentally ill patients: a comparative empirical study.' *Psychological Medicine,* 33: 1463-1471.

Waldfogel, S. & Meadows, S. (1996) 'Religious issues in the capacity evaluation.' *General Hospital Psychiatry,* 18:3, 173-182.

Wallace, R. J. "Practical Reason" *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (Ed.) [online] http://plato.stanford.edu/archives/sum2009/entries/practical-reason/.

Wason, P. C. (1983) 'Realism and rationality in the selection task.' In *Thinking and Reasoning: Psychological Approaches*, J. St Evans (Ed.), Routledge & Kegan Paul, pp. 44-75.

Wedgwood, R. (2007) *The Nature of Normativity*, Oxford: Oxford University Press.

Wiggins, D. (1975) 'Deliberation and Practical Reason.' In *Needs, Values, Truth: Essays in the Philosophy of Value*, Oxford: Basil Blackwell, pp. 215-237.

Williams, M. (1991) 'Blind Obedience: Rules, Community and the Individual.' In *Meaning Scepticism*, K. Puhl (Ed.), Berlin: Walter de Gruyter & Co., pp. 93-125.

Wilson, N. L. (1959) 'Substances without substrata.' *Review of Metaphysics,* 12: 521-539.

Wittgenstein, L. (1953) *Philosophical Investigations*, G. E. M. Anscombe (Ed.), Oxford: Basil Blackwell.

Wittgenstein, L. (1956) *Remarks on the Foundations of Mathematics*, Oxford: Basil Blackwell.

Wittgenstein, L. (1969) *On Certainty*, G. E. M. Anscombe and G. H. von Wright (Eds.), Oxford: Blackwell.

Wong, J. G., Clare, I. C. H., Holland, A. J., Watson, P. C. & Gunn, M. (2000) 'The capacity of people with a 'mental disability' to make a health care decision.' *Psychological Medicine,* 30: 295-306.

Wright, C. (1980) *Wittgenstein on the Foundations of Mathematics*, London: Duckworth.

Wright, C. (1984) 'Kripke's account of the argument against private language.' *The Journal of Philosophy,* 81:12, 759-778.

Wright, C. (2002) 'Human Nature?' In *Reading McDowell: On Mind & World*, N. H. Smith (Ed.), New York: Routledge, pp. 140-159.

Zahavi, D. (2005) *Subjectivity & Selfhood: Investigating the First-Person Perspective*, Cambridge, MA: MIT Press.