SENTIMENTALISM:

A Humean Analysis of Moral Belief

(PhD thesis)

Fritz Martin Kretschmer
Department of Philosophy
University College London
Gower Street
London WC1E 6BT

ABSTRACT

In this thesis, I investigate the nature of moral belief and judgment. Moral beliefs come out as dispositional sentiments of sanction; that is moral judgments express dispositions to experience something like compunction, guilt, remorse or shame (in the first-person) and blame, resentment, indignation or outrage (in the second- and third-person). This analysis constitutes one particular dispositional theory of value typically captured by a biconditional of the following form:

$$\text{x is P} \Leftrightarrow \text{x is such as to produce a P response in subjects S}$$

Many moral philosophers, cognitivists and non-cognitivists, have adapted the general form of the biconditional for their purposes. The content of moral belief and judgment may be said to be dependent on psychological capacities of a various kind: interests, preferences (Hare), desires, second-order desires (Lewis), motivations, responses (Wiggins), attitudes (Blackburn) or plain dispositions. Common to most of these proposals is that the specific value-making states of mind remain underdefined. Attitudes, for example, are often merely characterized as what they are *not* supposed to be: standardly conceived beliefs with genuine truth-conditions (modelled, say, on a correspondence theory of truth).

It is one aim of this dissertation to supply an independent account of the relevant states of mind. My proposal therefore starts from the *psychological reality* of sentiments, giving a full analysis of what sentiments are, before developing the relation of specific sentiments to moral belief in form of another biconditional:

$$\begin{array}{l}\text{x is of moral value V} \Leftrightarrow \\ \text{x is such as to produce sentiments of sanction associated with V in subjects S}\end{array}$$

In filling in this formula I argue (1) for a version of Internalism about moral belief, and I reassess (2) the epistemic status of moral beliefs as dispositional sentiments concentrating on the notion of a sentimental cause. In the case of the non-moral emotion of fear, for example, the sentimental cause is what I am afraid *of* – a dangerous circumstance. But what are sentiments of sanctions directed at, what is their content? My answer seeks to make room for the notion of an *appropriate* sentiment while remaining epistemologically non-cognitive: sentiments of sanction are subject to conditions of critical reflection, but these conditions may only imply reasonable convergence of sentiments under substantial psychological assumptions.

In an appendix it is suggested that what has been previously argued on independent grounds may count as our best reading of Hume's moral philosophy.

PREFACE

It is the ambition of this dissertation to present the framework to a complete moral theory. What are moral beliefs? Can they be contrasted with other mental entities? Are they cognitive or non-cognitive (and in which sense)? How do they respond to justificatory demands?

Any promising answer to these pivotal questions, I contend, would ultimately have to be grounded on assumptions of human psychology. In particular, moral beliefs are inextricably linked to a psychological capacity to experience and exert sanctions. Here we may distinguish two key groups: first-personal sentiments of sanction such as guilt, shame, compunction and remorse; second- and third-personal sentiments of sanction such as blame, outrage, resentment and indignation. Moral beliefs are dispositional sentiments of sanction, moral judgments therefore express dispositions to feel guilty, ashamed etc. or morally angry in some way.

Human social behaviour may be largely characterized by a web of lines limiting the acceptable. If you address somebody politely, you invite an answer, and your facial play will mimic the stimuli received or reflect a pointed difference in social role and standing. If your expectations are disappointed – your smile is ignored, your hand rejected – you will either show some kind of anger (sanction in the second-person) inducing your counterpart to come up with an explanation or apology ("I was distracted", say), or you may wonder whether you yourself had overstepped some line inadvertently. If you discover you caused offence you may feel guilty (sanction in the first-person) and try to make amends. Alternatively, the communication may be aborted altogether or escalate into open conflict. It is already here in these minutiae of human interaction that the corrective role of sentiments of sanction appears.

Moral sentiments are the regulators of mutuality: they are intrinsically practical, and they hover and mediate between non-cognitive responses (marking perhaps the social intercourse of many animals) and highly reflected and flexible capacities of cognitive reason (typical of normative argument and long-term thinking). In Chapter

1, I give with the help of recent emotion-theory a general analysis of sentiments fitting this bill. While in the latter parts of the dissertation I seek to show that it is just such an analysis that captures, in dispositional form, the crucial features of moral belief.

During the course of this systematic strategy I am led into confrontations with major alternative theoretical accounts. The most obvious competitors are, on the one side, various forms of Moral Realism or Factualism, claiming that moral beliefs are only contingently connected to human psychological dispositions and answer, as factual beliefs, directly to reality. One sophisticated version of this realism I take to be Peter Railton's. His theory is discussed in Chapter 3.4. On the other side, there are theories more closely related to mine that acknowledge the keyrole psychological capacities must play in an adequate account of moral belief and judgment. Here we can distinguish between doctrines we may label Response-Dependent Realism (which includes Secondary-Quality Realism or Sensibility Realism – e.g. McDowell, Wiggins; Chapter 4.1.1 - 4.1.3), Projectivism (or Quasi-Realism – Blackburn; Chapter 4.1.4) and accounts that fall under the heading of Practical Reasoning Theories either of a Kantian (e.g. Hare; Chapters 3.5, 4.2) or of a Contractualistic (Gauthier, Mackie; Chapter 4.3) complexion. Unsurprisingly, I reject all these options in favour of my own version of Sentimentalism.

In contrast to some of its closer relatives, Sentimentalism, as it is developed in this dissertation, exhibits two central features.

(1) Sentimentalism characterizes the psychological reality of moral responses as sentiments of sanction. Thus it gives some content to the abstract, and too often interchangeably used notions of attitudes, desires, dispositions, preferences, responses etc.

(2) Sentimentalism remains epistemologically non-cognitive. It is conceded that not everybody is normatively compelled to develop uniform sentimental

> responses to any given act, character or situation. At the same time, Sentimentalism claims to be able to accommodate significant corrective and justificatory resources.

Finally, Sentimentalism sketches a non-foundational decision-procedure: Among competing sentimental responses towards any given act, character or situation the response counts as reasonable that on the weakest motivational grounds allows for the widest acceptance of a normative system (Ch. 4.3.5). The motivation licensing the keysentiments of sanction is identified as a version of sympathy. *If*, contingently, we are motivated by some form of sympathy (that is, we have preferences beyond strictly self-seeking concerns, including others in some way), we are then normatively compelled to agree on a system of mutual norms that is roughly just.

I am pleased to acknowledge the encouragement and criticism of many people, first and foremost of my supervisor at UCL, Ted Honderich, who insisted that one should not defend a doctrine to be labelled Sentimentalism without giving first an account of what sentiments are. I am also greatly indebted to other members of University College's philosophy department, in particular Sarah Richmond and Malcolm Budd who read the whole of the penultimate draft. Anthony Price, now of Birkbeck College London, suggested valuable improvements to the last printout. In the intercollegiate Thesis Seminar of the University of London, I had on many occasions the opportunity to benefit from stimulating discussions with, among others, Mark Sainsbury and David Ruben.

The philosophical path I have travelled was first suggested to me by Ernst Tugendhat, then professor at the Freie Universität Berlin, where I spent some of my undergraduate years. He was convinced that thoughts circulating among the "moral sense" or "sentimentalist" school of 18th-century British Moralists like Shaftesbury, Hutcheson, Hume and Adam Smith could contribute to contemporary moral debate. I found this to be true, though I owe by far the most to only one of these philosophers,

David Hume. In an Appendix to this dissertation, I hope to make the allusions to Hume more explicit and bring them into a cohesive picture.

I also would like to acknowledge the participants of three successive reading-groups instigated during the year 1993-4: mainly James Cornwell, Hallvard Lillehammer, Dominic Murphy and Sarah Richmond. There I could try out some of my ideas on Gibbard's book *Wise Choices, Apt Feelings*, on "Trends in Recent Moral Philosophy" and on "Response-Dependence and Projectivism". Last, I have to thank two of my friends, Peter Schaber of Zürich university and Max Kölbel, PhD student at King's College London, for many incisive comments, and my wife Sharon Sanbrook-Davies who reads my work with a sharp mind.[1]

[1] Editorial note: Throughout this dissertation, single quotation marks are used in the logician's sense to name words or expressions. Double quotation marks quote and perform other, looser tasks (e.g. using a word and drawing attention to it, naming a use (as in irony) etc.). In this I follow Gibbard (1990, 6, n. 4). 'ise'-spellings have been retained for words in common use, such as 'apologise' and 'recognise'; 'ize'-spellings are used for terms of art like 'universalize'.

1. SENTIMENTALISM

Moral beliefs are sentiments. This I call the thesis of Sentimentalism. Before considering how it can be argued for (Ch. 2) and then arguing for it (Chs. 3 & 4), I shall be concerned with stating the thesis more fully and clearly. This will occupy me for the following pages.

1.1 Two Pictures of The Mind

In calling moral beliefs sentiments I claim moral beliefs to be mental entities of a certain kind. How are sentiments to be characterized, and how can sentiments be contrasted with other mental entities? It may be useful to mention a few traditional distinctions I do *not* wish to make.

Received opinion has it that there is an obvious distinction between reason and sentiment. Here, 'sentiment' seems to be a term for a mental faculty governing feelingful states (states we may call "emotion" or "passion") which are to be contrasted with states under the control of "reason". In fact, reason and sentiment are often conceived of as two competing and antagonistic forces pulling on human character and action. Dr. Johnson in his moral tale *The History of Rasselas, Prince of Abissinia* lends his mighty rhetoric to a professor expressing this traditional view:

> He [the professor] shewed ... that human nature is degraded and debased, when the lower faculties predominate over the higher; that when fancy, the parent of passion, usurps the dominion of the mind, nothing ensues but the natural effect of unlawful government, perturbation and confusion; that she betrays the fortresses of the intellect to rebels, and excites her children to sedition against reason their lawful sovereign. He compared reason to the sun, of which light is constant, uniform, and lasting; and fancy to a meteor, of bright but transitory lustre, irregular in its motion, and delusive in its direction.
>
> He then communicated the various precepts given from time to time for the conquest of passion, and displayed the happiness of those who had obtained the important victory, after which man is no longer the slave of fear, nor the fool of hope; is no more emaciated by envy, inflamed by anger, emasculated by tenderness, or depressed by grief; but walks on calmly through

> the tumults or the privacies of life, as the sun pursues alike his course through the calm or the stormy sky.[1]

This picture of the mind with its sharp divide between cognitive and non-cognitive mental space originates in ancient Greek thought. Among its prominent philosophical ancestors are Plato, to some extent Aristotle, and later Aquinas. They all posit a mental faculty of reason which is to distinguish the wise man from the fool, but they lack, like Dr. Johnson's professor, an account of the somehow mysterious psychic force in virtue of which cognitive reason is said to combat and sometimes overcome non-cognitive sentiment, passion or emotion.[2]

There is a more recent and altogether different tradition that also contrasts reason and sentiment. Here, reason is restricted to one aspect of thought, typically belief, and is understood as wholly passive. The *locus classicus* of this view is Hume's section "Of the influencing motives of the will" from Book II of the *Treatise.*[3] There, Hume argues "*first*, that reason alone can never be a motive to any action of the will; and *secondly*, that it can never oppose passion in the direction of the will" (T. 413). A wise man's and a fool's passions, so Hume is often read as saying, are of the same

[1] Johnson 1968 (1759), 46-7. Inspired by this oratory, Rasselas desires to visit the professor. After buying his way into the professor's house, Rasselas finds him in a room "half darkened, with his eyes misty, and his face pale", grieving over the loss of his daughter. The chapter carries the rather cruel title "The prince finds a wise and happy man".

[2] In the *Republic* (Book 4), Plato partitions the soul into three "homunculi": reason (*to logistikon*), spirit (*to tymoeides*) and appetite (*to epithymêticon*). Reason controls the gate to theoretical and practical truths, appetites are bodily states of the type of hunger, thirst and lust, spirit again is something common to children and animals that can take sides with the appetites against reason (e.g. as impulsive anger), or with reason against the appetites (e.g. as controlled anger). For Plato, desire and emotion operate in all three psychic parts. Aristotle defines the cognitive rational against the non-cognitive vegetative mental sphere more sharply by contrasting reason (in its theoretical and practical versions) with desire. While some desires (here *alogon*) may take account of reason (*logos*) as one does of "one's father" (*Nichomachean Ethics,* 1103a3), sentiments or emotions are typically characterized as non-cognitive impulses or desires (e.g. *De Anima,* 403a29). (Aristotle's generic term for desire is *orexis*, *epithymia* denotes one particular bodily species of desire). In the *Rhetoric* (II, 2-11) Aristotle appears to suggest a more subtle account of particular emotions (amongst them anger, hatred, fear, pity, envy and *schadenfreude*) as involving cognitive elements (cf. John Cooper, 1993). Aquinas takes up the traditional interpretation of Aristotle: "... passion and therefore emotion, is seated in the orectic rather than the cognitive part of the soul" (Aquinas, 1967 (1267-73), la 2ae, Q22, 2, p.11).

[3] Hume, 1978 (1739-40). Throughout this dissertation, quotes from the Treatise will be revealed by the page number preceded by 'T', quotes from the Enquiries (Hume, 1975 (1748, 1751) by the page number preceded by 'E'.

non-rational complexion. "'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger" (T. 416), as long it is destruction I ultimately prefer. Beliefs may only (1) inform the passion of "false suppositions" (such as my being part of the world and, in consequence, the suicidal nature of my preference), or (2) discover "insufficient means" (e.g. I may employ to my destructive purpose). Thus a passion can only be unreasonable relative to its own purpose. In this picture, passions (emotions, sentiments) turn out to be rather monolithic. They spring up without apparent cause, or at least, seem to be identifiable without regard for their causes. "When I am angry, I am actually possest with the passion, and in that emotion have no more a reference to any other object, than when I am thirsty, or sick, or more than five foot high" (T. 415). In the Appendix, I shall dispute that this famous section "Of the influencing motives of the will" constitutes the whole truth about Hume's view of the passions. In the present context, Hume's alleged view must count as the predecessor of later psychologies, in particular it seems to anticipate the so-called James-Lange theory of the emotions.[1] In *The Principles of Psychology*, William James writes:

> One natural way of thinking about coarser emotions is that the mental perception of some fact excites the mental affection called the emotion, and that this latter state of mind gives rise to the bodily expression. My theory, on the contrary, is that *the bodily changes follow directly the perception of the exciting fact, and that our feeling of the same changes as they occur IS the emotion*. Common-sense says, we lose our fortune, are sorry and weep; we meet a bear, are frightened and run; we are insulted by a rival, are angry and strike. The hypothesis here to be defended says that this order of sequence is incorrect, that the one mental state is not immediately induced by the other, that the bodily manifestations must first be interposed between, and that the more rational statement is that we feel sorry because we cry, angry because we strike, afraid because we tremble.[2]

According to James, my psychic state can be identified as, say, fear whether I *believe* myself in fearful circumstances or not. Sentiments, thus, can be characterized without

[1] The James-Lange theory derives its name from James' collaboration with the Danish physician Carl Lange.

[2] James, 1890, Vol. II, ch. 25. The "subtler" emotions (e.g. love, indignation, pride) are modelled on the paradigmatic "coarser" emotions (e.g. rage, grief, fear), i.e. subtler emotions are themselves to be understood either as sensations or as etiolated sensations.

reference to their situational causes. They may be undirected non-cognitive experiences we simply undergo.[1]

So much for two philosophical psychologies assuming a clear divide between cognitive and non-cognitive mental space. Both conceptions of the mind seem to locate sentiments on the non-cognitive side of the divide, in the first case to be combated by a powerful but mysterious rational faculty, in the second case to be assisted by a weak and slavish servant. These are the traditional disguises of reason.

1.2 An Analysis of Sentiments

The view of the sentiments I am going to sketch now does not presuppose the mental divide of reason and sentiment. I think we can define sentiments as distinct mental entities in the following way:[2]

> The concept entails: each sentiment (a) has a typical cause, (b) is marked by a typical symptom, and (c) issues a typical action. Together, cause, symptom and action are sufficient to identify a mental state as being of the type of a particular sentiment.

[1] This view has been exposed over the years to a fair amount of criticism, mainly on the grounds (1) that hardly distinguishable symptoms seem to occur in such different emotions as fear and anger, envy and jealousy, and (2) that in Jamesian terms there is no room for the notion of an *appropriate* emotional response. For recent discussions of Jamesian theories of the emotions see Pitcher, 1965; Lyons, 1980, 12-16; Gordon, 1987, ch. 5; Oakley, 1992, 16-22. Again, there is exegetical disagreement whether an analysis of the emotions as unstructured sensations represents really James' most considered opinion on this matter (cf. Gibbard, 1990, 134 n.7).

Some later doctrines of the Logical Positivist movement may be seen as growing directly out of a Jamesian conception of the emotions. Thus A.J. Ayer's branch of ethical "emotivism" allegedly commits not fully descriptive statements to the class of unstructured "boo" or "hurray" utterances (Ayer, 1946 (1936), ch. 6).

[2] The use of 'sentiment' here may need some terminological clarification. 'Sentiment' may denote a mental faculty (to be contrasted with "reason") as well as the states it governs – states we call "emotion" or "passion". In the following I use 'sentiment', 'emotion' and 'passion' interchangeably for the mental state whose concept and status is in question.

This definition starts from a proposal by Kenny in his *Action, Emotion and Will.* There, Kenny elucidates: "The concept, for example, of *fear* stands on three struts: (a) fearful circumstances (b) symptoms of fear (c) action taken to avoid what is feared."[1] Kenny's transcription of fear does not constitute a full analysis; rather it comes as a fleeting remark on how concepts of emotions are linked to non-emotional concepts, or, more specifically, how emotions can be attributed in particular instances. Still, Kenny takes his proposal seriously enough to react to an obvious charge of circularity.

> "Fearful circumstances" ... could be replaced by "dangerous circumstances"; the concept of *danger* does not involve reference to the emotions, since we can speak of danger to plants or artefacts, which cannot have emotions. The symptoms of fear could be given a purely physical description; many of them, such as fluctuations of breathing-rate, are shared with other emotions and might be produced by purely physical causes. Avoiding action can be explained purely in terms of intention, without reference to emotion. (70)

Kenny elaborates less on other aspects of his proposal. In particular his characterization of emotional symptoms is unclear, and, in consequence, the distinction between symptom and action remains problematic. Either we understand emotional symptoms as expressive behaviour (in this case symptoms of fear, say, should include things we would normally label as actions, such as nail-biting or the uttering of sentences expressing fear), or we restrict symptoms to unintentional manifestations of the visceral system such as trembling, acceleration of the heart-beat or irregular breathing (in the latter case, we have a clean divide between symptom and action but at a cost – the cost of being left with two distinct classes of actions: actions expressive of fear and actions taken to avoid what is feared). For the moment, I suggest we adopt the broader reading of "symptom" as expressive behaviour.

How adequate, then, is an analysis of sentiments in terms of cause, symptom and action? Immediately, there come to mind various cases the analysis does not seem

[1] Kenny, 1963, 67. Kenny's three struts are briefly discussed in Lyons (1980, 47-8). Lyons rightly goes on to give a cognitive interpretation to Kenny's overall treatment of the emotions. Gibbard (1990, 133) uses Kenny's proposal as the basis of his non-cognitive, evolutionary account of the moral emotions. For a more detailed look at Gibbard's proposal, see my Chapter 2.1.

to fit. I shall briefly discuss three apparent counter-examples, the second and third leading to revisions of the initial proposal.

First, in many instances of fear some elements of the threefold analysis seem to be missing. People are afraid in less than fearful circumstances (feature (a) missing), sometimes they suppress signs of fear (feature (b) missing), sometimes they tremble but withstand danger (feature (c) missing). Perhaps one may even experience fear when exhibiting only one of the features. I might be afraid of subways or bridges but successfully avoid being put in these fearful situations (no fearful circumstances, no symptoms of fear); I might be a fearful but composed fire-fighter (no symptoms of fear, no action taken to avoid what is feared); I might sit trembling in a safe but dark room (no fearful circumstances, no particular action taken). *In extremis*, I might be safe, exhibit no sign of fear, take no particular action but profess to being afraid. No other term might spring to my mind suitable to label what I feel. In all these cases, we would not easily discount avowals of fear but the further we deviate from paradigmatic fear the more we feel the need for an explanation. "Why do you feel afraid?", we want to ask, and "Why do you do nothing about it?", "Why do you appear so calm?". In making sense of avowals of fear we refer to all three elements of the standard analysis, and some answers simply would not satisfy us. As Kenny says (68), I am afraid "because it is five to three" is without further qualifications hardly acceptable as an account of the causes of my fear. Similar claims could be made with regard to symptom and action. It would cast doubt on my ability to understand what I say if I professed to being afraid appearing unmoved, wearing an easy smile. To be sure, there could be an explanation why someone does smile in the face of danger. John Wayne, we understand, will not let his enemy enjoy an undue advantage. The point is that we would *want* such an explanation. *If any of the features of paradigmatic fear is missing, there is a need for an explanation why it is missing. If no explanation is available whatever somebody feels cannot count as fear. Ceteris paribus*, each sentiment will exhibit a typical cause, symptom and action. This deals with the first objection.

A second objection to the threefold analysis arises from the consideration of a group of sentiments, emotions or passions which do not imply obvious action-tendencies. Under this heading, we may include reactive emotions like grief, awe, wonder, disappointment, emotional conditions like happiness, melancholia or nostalgia as well as the emotions of self-assessment such as pride, self-respect, humility, and perhaps in some passive sense guilt or shame. In the case of grief, for example, one might say that the emotion is exhausted by its causes and symptoms. When I grieve, I grieve. I might be asked after the cause of my grief but I will *not* be asked why I do not perform a particular action for no particular action seems paradigmatic of grief.

In order to salvage the analysis, we might move in two directions. Either we could say that (1) for some emotions there is no clear distinction between elements (b) and (c), between symptom and action (weeping and other expressions of sorrow *are* the actions appropriate to grief[1]), or we could say (2) that grief counts somehow indirectly as motive for actions other than the expressions of grief. The first move concedes readily that some sentiments may not explain any particular action apart from actions expressive of that sentiment, the second move seeks to preserve a conception of sentiments as motives which seems to be essential to the full scope of the analysis. Unfortunately, this second possibility is obscure. Though we sometimes say "He made amends *because* he felt guilty" (citing "guilt" as an explanation or motive for a particular action other than the expression of guilt) it is odd to say "She slammed the door because she was mourning", or "He went to the fun-fair to forget his grief." Such ways of talking may contain a grain of truth but it is difficult to see precisely where. How are we to explicate the way grief figures here as a motive? If it were conceptually linked to particular actions relieving grief we could ask any mourner why he doesn't go to a fun-fair. This, however, seems grossly inappropriate.

[1] Does weeping constitute an action? Though typically unintentional, it can sometimes be induced or controlled. On the proposed broad reading of "symptom", I do not think I have to commit myself on the action character of weeping. I only contend that expressive behaviour *may* include items commonly labelled as actions. (I owe this point to David Ruben.)

When grief does not give rise to the seeking of relief, we do not need an explanation as we do when fear does not lead to the evasion of danger. We seem to understand and believe professions of grief by two elements alone, their causes and symptoms.

We have to conclude that for some sentiments the threefold analysis appears strained and may fail. Some sentiments lack the distinct action-element (c). In consequence, those sentiments may not be intrinsically motivational; if they explain particular actions (other than those counting as immediate expressive symptoms) they do so only contingently.

I turn now to a third and last objection to the proposed analysis of sentiments in terms of cause, symptom and action. In some instances, the same apparent causes, symptoms and actions may be common to two different emotions. How, for example, do we know whether somebody is jealous or envious? Both jealousy and envy may be seen as forms of anger; persons in envious and jealous states often exhibit angry symptoms and behaviour. Where jealousy and envy differ, it might be suggested, is with regard to their causes. Now an eruption of jealousy and envy may be occasioned by the same event, say, a friend's being courted by a person of fame. If the friend's behaviour in your opinion is impeccable you may feel envious at what you believe is his gain in social standing; if, however, you suspect undue flattery on your friend's part, or negligence of your appearance on the scene, you may feel jealous i.e. your state of mind may arise from an apprehension of rivalry. An observer may be able to decide which outer event caused your early and apparently upset departure from the gathering, but he cannot know whether you blame your friend for what has happened, whether you experienced envy or jealousy. It seems therefore not sufficient to analyse an emotion, passion or sentiment in terms of cause, symptom and action; we need to make room for how the bearer of an emotion *views* the cause of his emotional experience. This seems to necessitate the introduction of a cognitive element into the analysis. The cause of my being afraid is not that something is dangerous but that I *believe* something to be dangerous; the cause of your being jealous is not that somebody is a rival but that you *believe* somebody to be a rival.

We may revise the initial analysis in the following way. In order for something to be identified as a particular sentimental state, the following three conditions have to obtain (now taking the perspective not of an observer but of the bearer of that state): (a) a belief about a typical situational cause has to be entertained, (b) a typical symptom has to be expressed, and (c) a typical action has to be taken.

Though this cognitive account is cogent for envy and jealousy it is admittedly less plausible for more primitive, evolutionarily older emotions such as fear, rage, lust, hunger and thirst. It seems that fear as a biological mechanism (with its typical causes, symptoms and actions) can be identified without attributing beliefs. A mouse perceives the shadow of a bird of prey and flees agitatedly. That's it. A non-cognitive analysis of the emotions thus may characterize both jealousy and envy as instances where the biological mechanism of anger is operating; still, the non-cognitivist will not be able to distinguish these instances of anger as instances of two different emotions.

In response to some difficulties, I have now arrived at four versions of an analysis which, between them, should cover most things we want to call sentiments. Still in the game are the following.

> The concept of a particular type of sentiment entails
>
> (1) that (a) a belief about a typical cause is entertained, that (b) a typical symptom is expressed, and that (c) a typical action is taken (*cognitive analysis* as in the discussion of envy and jealousy);
>
> or (2) that (a) there is a typical cause, that (b) a typical symptom is expressed, and that (c) a typical action is taken (*non-cognitive analysis*, possibly adequate to fear);
>
> or (3) that (a) a typical cause is believed, and that (b) a typical symptom is expressed (*reduced cognitive analysis*, applies to reactive sentiments like grief);

or (4) that (a) there is a typical cause, and that (b) a typical symptom is expressed (*reduced non-cognitive analysis*, may apply to some moods, perhaps euphoria and depression).

1.3 Sentimental and Epistemological Cognitivism

Up to this point, I have used the crucial terms 'cognitive' and 'non-cognitive' in the sense in which these terms normally are employed in emotion theory. These uses are not always consistent and do not, in any case, match with the uses in meta-ethics. A cognitive theory of sentiments (emotions or passions) does not imply a cognitivist meta-ethics, nor does a non-cognitive reading of sentiments imply meta-ethical non-cognitivism. Since my programme combines work in both areas, emotion theory and meta-ethics, I should clarify my terminology before proceeding to the central thesis of this chapter.

Cognitivism in meta-ethics is an *epistemological* doctrine. It claims that we can know the truth of propositions like "slavery is wrong", "burning cats is cruel" and so on.[1] Cognitive theories of the emotions do minimally stipulate that emotions cannot be identified without referring to some *intentional content, thought or belief.* Some theories do not require that these thoughts are epistemologically cognitive (i.e. can be a matter of knowledge);[2] some theories even deny that an emotion's intentional

[1] Exceptions here might be Blackburn and Hare: Blackburn calls his projectivism a form of non-cognitivism while claiming at the same time that moral judgments are apt for truth-evaluation. He evidently uses 'cognitivism' in a sense closer to emotion-theory. In Blackburn's terms, moral judgments express non-cognitive attitudes or desires, not standardly truth-apt factual beliefs (e.g. Blackburn, 1984, esp. chapter 6). Richard Hare is another philosopher resisting the distinction suggested above: "If to think that [prescriptive questions] can be determined rationally is to have an epistemology or theory of knowledge, then one who thinks this, as I do, should perhaps be labelled a cognitivist. But I do not recommend the label, because those who are unable to envisage any other kind of reasoning than factual will think that if I am a cognitivist I must be a descriptivist, which I am not." (Hare, 1985, 52)

[2] Lyons, 1980, 59.

content can be forced into propositional form.[1] Before I complicate things further, I shall give a few examples.

Standard cognitive theories of the emotions often use the "belief-desire" terminology. "Belief", here, corresponds to what I called "cause" in a cognitive analysis (i.e. how the bearer of an emotion views the cause of his state – analyses (1) and (3)), "desire" captures perhaps the remaining elements. In terms of belief and desire, the sentiment of compassion could be reconstructed as the belief that another person is suffering plus the desire to help her. The sentiment of *schadenfreude* could be said to be constituted by the very same belief that another person is suffering plus the desire to see her suffering. "Envy" is perhaps the belief that another person enjoys superior advantages plus the desire to enjoy them yourself; "jealousy" may be the belief that another person is a rival plus the desire to preserve one's own rights. These examples, I am aware, are less than subtle. For the present purposes, they do not need to be subtle nor would I want to commit myself to beliefs and desires as the basic mental entities. I seek to distinguish varieties of cognitivism, and here the reconstructions of sentiments as clusters of beliefs and desires suggest a *type* of analysis. The underlying theory of the sentiments I shall call weak sentimental cognitivism, for on this standard cognitive account, sentiments are not exclusively cognitive. They contain a non-cognitive component, that is, a desire.[2]

Strong cognitive theories claim that sentiments *are* beliefs, they are a way of looking at the world. Perhaps the sternest cognitive theory of the emotions has been suggested by Davidson. In his piece "Hume's Cognitive Theory of Pride", Davidson

[1] A. Baier, 1976; 1977; 1978.

[2] Exponents of belief-desire theoretical accounts of the emotions are e.g. O. H. Green (1992), *The Emotions: A Philosophical Theory*, and I. Thalberg (1977), *Perception, Emotion and Action*. Most authors argue that emotions, though containing beliefs and desires, cannot be reduced to these two kinds of intentional states. Gordon (1987, 8-9) gives an example of two farmers who, knowing that their crops will not survive another week of drought and being told that there is a fifty percent chance of rain within the next week, differ neither in any relevant desires nor beliefs. Yet one is afraid it will not rain, the other is hopeful it will; one prepares for irrigation, the other does not. Oakley (1992, ch. 1) argues for an account of emotions as complexes of beliefs, desires and affects.

analyses a particular case of pride as entailing the beliefs that (i) I own a beautiful house, that (ii) all who own beautiful houses are praiseworthy (in so far as they own beautiful houses), and therefore that (iii) I am praiseworthy (in so far as I own a beautiful house). The last belief expresses the emotion.[1]

Both strong and weak sentimental cognitivism are compatible with epistemological non-cognitivism. One or more of the beliefs contained in the strong cognitive analysis of sentiments may be, so to speak, a matter of opinion not knowledge. Thus, the epistemological status of "All who own beautiful houses are praiseworthy" is not settled by its characterization as belief.[2] Similarly, weak sentimental cognitivism might be epistemologically non-cognitive in two ways: (1) A belief about another person's superior advantages (as in the analysis of envy, say) may be a matter of opinion not knowledge; or (2) even if it were a matter of knowledge, the belief may not imply any particular desire. Another person's superior advantages may lead to admiration as well as envy.

[1] Davidson, 1980 (1976), 277. Davidson, though, does not claim that all emotions are cognitive in his strong sense. For a critique of Davidson's piece, see A. Baier (1978), and G. Taylor (1985, ch. 1). Robert Solomon (1976) also defends a strong version of sentimental cognitivism. The Stoics were the first to analyse affections (*pathê*) as beliefs, though as false ones, to be corrected by a kind of cognitive therapy. For a good overview of the Stoics' account of mental life, see A. Price, 1995, ch. 4.

[2] I argue here from a wider notion of belief. Strong sentimental cognitivism claims that sentiments, emotions or passions are a way of looking at the world; they may be better characterized as evaluations than as feelings. If strong sentimental cognitivism relied on a narrow notion of belief as propositional attitude, i.e. as expressing a proposition that possesses truth-value, strong sentimental cognitivism might be incompatible with epistemological non-cognitivism. But again, this may depend on whether a minimal or substantial conception of truth is presupposed. (Cf. Chapter 4.1 below).

1.4 Sentimental Analysis and Moral Belief

After this short excursion into the maze of recent theories of the emotions, it should have become clear that the thesis of Sentimentalism (if it holds true) does not preempt the traditional *epistemological* questions about the status of moral beliefs. Even under a cognitive interpretation of sentiments, meta-ethical non-cognitivism remains a possibility. Where, then, does the thesis bite? At the heart of the proposed analyses of sentiments lies, I believe, the interpretation of sentiments as motives.[1] In attributing sentiments, we can explain and make sense of behaviour. On the full, threefold analysis (versions (1) and (2)), the mental entities of sentiments explain two kinds of behaviour – first, behaviour immediately expressive of a particular sentiment (i.e. fear explains the symptoms of fear), and secondly other behaviour somehow directed at the sentimental cause (i.e. fear explains actions taken to avoid what is being feared). On the reduced, twofold analysis (versions (3) and (4)), sentiments only explain their immediate expressions. The task is then, first to choose one version of sentimental analysis, secondly to apply it to moral beliefs, and thirdly to show that the analysis is suitable to distinguish moral beliefs as sentiments from other mental entities.

Moral beliefs, I declare, are members of the class of sentiments individuated by versions (1) or (2) of the analysis i.e. they entail, and therefore explain, actions other than the immediate expressions of moral belief. While a moral belief is typically expressed by an utterance, the ascription of a moral belief explains more than that utterance. "I apologised because I believed I had offended her" is as fully an explanatory way of speaking as we could wish, citing a moral belief "I offended her (and this was wrong)" as reason and motive for a particular action (the apology). The act of apologising is distinct from the action expressing the moral belief.

[1] Again Kenny has seen that very clearly. He writes (1963, 38): "it is not just an unfortunate accident of idiom that we use the same words, such as "love", "anger", and "fear", in the description of feelings as we do in the attribution of motives. The two uses of an emotion-word are two exercises of a single concept; for it is through their connection with motivated behaviour that feelings are identified as feelings of a particular emotion."

I shall not now defend the contention that all moral beliefs are of this explanatory, practical nature. The whole of Chapter 3 will be devoted to arguing this claim. To be sure, whether the thesis of Sentimentalism holds true, depends on the success of this later defence. At the moment, however, I am engaged in stating the scope of the thesis of Sentimentalism. Here I simply stipulate that moral beliefs are sentiments of a kind with fear, anger, envy and jealousy, and not of a kind with grief or other reactive or assessing emotions. Versions (3) or (4) of the analysis do not apply to moral belief.

This claim, I suspect, is still less than clear. To be justified in comparing the mental state of fear, say, to the mental state of entertaining a moral belief, it does not suffice to state structural similarities; we will have to overcome structural dissimilarities. Most strikingly, one may want to resist the analogy because of what may be called the propositional surface of moral beliefs. Moral beliefs issue in propositional judgments (e.g. "I believe *that* slavery is wrong"), but do we not merely *feel* (and then perhaps *voice*) sentiments? The alleged asymmetry between the propositional properties of beliefs and sentiments is treacherous; it covers two distinct charges. First, the uttering of "I am afraid" is often a symptom of fear, i.e. it becomes part of emotional behaviour itself. Sentimental statements are used to express emotions at the scene of emotional happenings. The statement "slavery is wrong", on the other hand, may be uttered at a scene of slavery but more often it is not. The first charge of asymmetry then goes: while sentimental statements are typically expressions of occurrent sentiments, statements of moral belief are not. This charge, however, is easily defused. There are no conceptual restrictions on dispositional uses of sentimental statements. I may say "I am afraid of nuclear power stations" without expressing occurrent fear. I might then express that I would feel afraid (dispositionally) if I were to think of the potential dangers of nuclear power stations. Similarly, the statement "slavery is wrong" may express not an occurrent but a dispositional sentiment. Conceding this point, the question remains of what kind the underlying occurrent sentimental state is to which a typical moral statement expresses a disposition. Here, the second charge of asymmetry butts in. If moral beliefs are parasitic on occurrent moral sentiments just as dispositional fear is parasitic on occurrent fear, do not the propositional properties of belief disappear in this sentimental translation? Moral belief issue in judgments; judgments can be negated ("not: slavery is wrong"), conditionalized ("if slavery is wrong, then preaching slavery is wrong, too"), they can be conjoined and disjoined with other moral judgments. Do statements of occurrent or dispositional sentiments allow for these semantic operations? Does the propositional surface of moral belief resist a translation into "that" clauses which parallel the features of the proposed analysis of sentiments? On the proposed analysis, may we call moral beliefs sentiments without subscribing to an error theory, i.e. the view that the ordinary propositional features of moral belief are inexplicable?

Just as we can distinguish in the case of fear an occurrent state, and, dependent on it, a dispositional state, we may interpret moral beliefs as dispositional sentiments parasitic on occurrent sentiments such as guilt, compunction or shame (in the first-person) and anger, resentment or indignation (in the second/third-person). In order to preserve the more narrowly semantic properties of moral judgments, two options may be pursued. For a non-cognitive analysis of sentiments, moral statements express dispositions we may or may not have; in a cognitive reading, moral statements are about the *believed* causes of moral sentiments. Summarily, the constitutive elements of sentimental analysis may be set against features of moral belief in the following way.

Cause. Sentiments imply situational causes, and so do moral beliefs. In the case of fear, the cause is what I am afraid *of*;[1] either non-cognitively (version (2) of the analysis) as a dangerous circumstance, or cognitively (version (1) of the analysis) as entailing a belief about a dangerous circumstance. Analogously, a moral belief

[1] Malcolm Budd suggested to me this may not be true in some cases: In fearing a third World War, the belief that there will be, or the thought that there might be, is not what you are afraid *of*. This objection can be circumvented by specifying the cause. You are afraid of Word War III in respect of the dangerous circumstances and suffering it will generate.

encloses a cause, something I approve or disapprove *of*; either non-cognitively (2) as instances of slavery (say), or cognitively (1) as entailing a belief: "slavery is humiliating", or "I would not want to be enslaved", or "everybody does feel compassion with slaves", or "slaves will harbour resentment, thus breeding general unrest" etc. These possible beliefs, though sentimentally cognitive as thoughts, may not be epistemologically cognitive.

Symptom. Sentiments exhibit symptoms. Sentimental statements such as "I am afraid of nuclear power stations", however, may express no more than that under certain conditions certain symptoms *would* occur. The same can be said of moral beliefs. Though moral beliefs are often symptomless (apart from their occasional utterance), my belief that slavery is wrong, for example, entails that I would experience symptoms, say, akin to symptoms of anger when faced with instances of slavery, and perhaps akin to shame when found guilty of practising slavery.

Action. Sentiments considered relevantly similar to moral beliefs exhibit actions other than the actions expressing those sentiments. Thus it is part of my fear of nuclear power stations that I avoid them. We can say that I avoid nuclear power stations *because* I am afraid of them.[1] With moral beliefs, we should say that I do not own slaves, or that I take or support measures to censure slavery *because* slavery meets with resentment and anger, or guilt and shame.

Reading the analysis backwards, we arrive at an explanatory chain. I avoid nuclear power stations because I am afraid. I am afraid because nuclear power stations are fear-inducing or believed to be dangerous. The thesis of Sentimentalism claims in effect that such an explanatory chain is valid for moral beliefs.[2]

[1] Remember, it is no objection to this claim that you may demonstrate at a nuclear power station because you are afraid of it. We need only the explanation that you see the demonstration as a means to get rid of the power station. If no such explanation is available, whatever your mental state is, it cannot count as fear.

[2] We have seen that the propositional features of moral belief can be, *prima facie*, relocated in a sentimental analysis, interpreting moral judgments as judgments expressing dispositions or as judgments about sentimental causes. This does not preclude that much work needs to be done to find a logic of sentimental judgments parallel to the ordinary semantics of declarative sentences. One main hurdle here is the so-called Frege-Geach objection of embedded contexts (Geach, 1965). To conclude,

1.5 Sentiments and Other Mental Entities

One analysis of moral beliefs as dispositional sentiments has now been stated. But is not the proposed reading of sentiments so broad that it is in danger of losing its edge? Are not, at least on a well-known functionalist account, all mental states identified by their typical causes and behavioural effects?[1] How, then, does an account of moral beliefs as sentiments succeed in distinguishing moral beliefs from other non-sentimental states of mind?

One crucial distinction, we may want to draw, is between moral beliefs as sentiments and non-normative beliefs. According to the proposed sentimental analysis, moral beliefs can figure as motives; non-normative beliefs, I contend, cannot. What is the difference between "I apologised because I believed I had offended her" and "I crossed at London Bridge because I believed that Tower Bridge was closed"?

David Hume, as well as being one of the more notorious moral sentimentalists, gives a sentimental reading to beliefs about "matters of fact". Thus he blurs the distinction between moral and non-normative beliefs I wish to make. What does a sentimental analysis of non-normative belief look like? Beliefs about the world, or at least beliefs about the presently unobserved, are distinguished, so Hume thinks, by their influence on the mind, by their "force and vivacity" (T. 199). In the *Appendix* to his *Treatise*, Hume adds:

> We may ... conclude, that belief consists merely in a certain feeling or sentiment; in something, that depends not on the will, but must arise from certain determinate causes and principles, of which we are not masters. When we are convinc'd of any matter of fact, we do nothing but conceive it, along with a certain feeling, different from what attends the mere *reveries* of the imagination. And when we express our incredulity concerning any fact, we mean, that the arguments for the fact produce not that feeling. Did not the

as we would, from (1) "slavery is wrong", and (2) "If slavery is wrong, then preaching slavery is wrong, too" that (3) "preaching slavery is wrong", "slavery is wrong" in (1) and (2) must have the same meaning. In (1) "slavery is wrong" expresses a disposition (on at least one of my analyses), but does "slavery is wrong" express a disposition in (2)? I address the problem in Chapter 2.2.

[1] For a classic statement of functionalism, see Lewis (1966, 1972).

> belief consist in a sentiment different from our mere conception, whatever objects were presented by the wildest imagination, wou'd be on an equal footing with the most establish'd truths founded on history and experience. There is nothing but the feeling, or sentiment, to distinguish the one from the other. (T. 624)

In Humean terms, to believe that Tower Bridge is closed is to have a certain feeling lively enough to render that belief "the governing principle of our actions" (E. 50). "I believe that Tower Bridge is closed" expresses that I would use London Bridge rather than Tower Bridge if I were to cross the river; even stronger: only if the latter action is taken does it become clear that I really believed that Tower Bridge was closed. According to Hume, moral beliefs and beliefs about matters of fact are both sentimental states in that they govern actions. As sentiments, moral beliefs and beliefs about matters of fact are distinguished from "the mere *reveries* of the imagination".

Hume's account of belief is problematic on many grounds.[1] In the present context, I am only concerned with the conception of belief as "the governing principle of our actions". Are moral beliefs and beliefs about matters of fact alike in that they both motivate actions? Along with most people, I happen to think they are not. One might believe that Tower Bridge is closed without ever having the intention to cross the river. Thus beliefs about matters of fact can be separated from action-tendencies. *Given my belief that Tower Bridge is closed, we do not need an explanation why I do not cross the river at all.* If I cross at London Bridge because I believe that Tower Bridge is closed, the latter belief motivates my crossing at London Bridge, but it does so only in conjunction with my intention to cross the river. For moral belief, on the other hand, it is part of its concept (as suggested by the three-part analyses) that it

[1] As it is obvious from his treatment of "reveries", Hume's sentimentalism cannot amount to a functionalist account of all mental states. Reveries are defined as having no lasting influence on action; they are mental states without effects. How do we know that we have reveries? Hume here appears to rely on residues of the Cartesian introspective picture of the mind where "nothing is ever really present with the mind but its perceptions or impressions and ideas" (T. 67). On the charge of psychologism, cf. my Appendix below p. 162-3.

counts as a motive. If I believe "I offended her" but do not apologise we need an explanation why I do not apologise.[1]

In the last few paragraphs I argued that an application of sentimental analysis to belief is apt to distinguish moral from non-normative beliefs. Does this not conflate moral beliefs in a narrow sense with other normative beliefs which certainly include beliefs of prudence, etiquette, aesthetics, epistemology and perhaps rationality? To this reservation, my answer is essentially terminological. Call normative beliefs moral beliefs in a broad sense. Moral beliefs in this broad sense concern what ought to be done (as including considerations of prudence, etiquette, etc.). Moral beliefs in a narrow and traditional sense, as when we call something the right moral choice, are an especially colourful sub-class. They may compete with other normative considerations. As it stands, the thesis of Sentimentalism fits both moral beliefs in the narrow and in the broad sense. In developing Sentimentalism, I accepted some support from narrowly moral phenomena such as the reality of states of mind like indignation, resentment, guilt and shame. Since we talk with less psychological certainty of broadly normative sentiments (i.e. occurrent states of mind that underlie choices of action), my programme is for the time being orientated on moral beliefs in the narrow sense. Yet, it should turn out that the distinction between broad and narrow morality is not as critical as it sometimes has been made out to be.[2]

[1] The reader may discover little new in this. The distinction I wish to make has often been cast in terms of a "belief" and "desire" model of intentional action. Following this model, for the belief that Tower Bridge is closed to make up a sufficient cause of action, we have to add a relevant desire. Without setting out a specific form of the belief-desire theory, this terminology advances little. To explain what specific things these alleged mental states of belief and desire are might be most of the work of explaining what moral belief is. I only claim that any account of action would have to make room for the distinction I drew between mental states for whose motivational inefficacy we would need an explanation and mental states for whose we wouldn't.

[2] My use of 'moral' in the broad sense coincides with Bernard Williams' term 'ethical' (Williams, 1985b, ch. 1). Other distinctions between morality in a narrow and a broad sense can be found in Mackie (1977, 106) and Brand (1979, chs. 9-10).

1.6 Résumé and Outlook

In this chapter, I sketched an analysis of sentiments which I then applied to moral beliefs, claiming, finally, that it identifies moral beliefs as distinct mental entities. To call moral beliefs sentiments is to place moral beliefs in an explanatory chain. For a standard emotion, the scope of sentimental analysis can be stated thus: My fear, say, of nuclear power stations entails that I avoid nuclear power stations (a) *because* I am afraid, and that I am afraid (b) *because* nuclear power stations are dangerous. The explanation of a sentiment's action-tendency (as symptom and other behaviour) ultimately points to some cause, in the case of fear it is a dangerous circumstance. This cause we may give a cognitive or non-cognitive reading, first sentimentally: "Is the cause a belief about danger or simply a dangerous circumstance?", and secondly epistemologically: "Is the (sentimentally cognitive) belief about danger a matter of knowledge or not?". The remainder of this dissertation may be seen as a detailed application of this standard case of sentimental analysis to moral belief.

Chapter 2 on methods rules out (among other things) a non-cognitive sentimental reading of moral belief. If moral beliefs were sentimentally non-cognitive the thesis of Sentimentalism could be verified empirically. I argue this cannot be done. In Chapter 3, I vindicate a reading of moral beliefs as motives, that is, I defend the first of the "because" clauses (a) I identified above; "He apologised because he had given offence" is, I claim, as fully explanatory as "She demonstrates against nuclear power stations because she is afraid of them". In Chapter 4, I consider the vexed question of the epistemological status of beliefs behind the second "because" clause (b). A belief that nuclear power stations are potentially dangerous may give the reason for someone's fear. For sentiments like indignation (about an instance of slavery) or guilt (about my offensive behaviour) does anything correspond to that belief? In my sentimental transcription of the moral belief that slavery is wrong [p. 22-3 above], I tentatively suggested four possible beliefs to fill in the second "because" clause (b): I may feel indignation about instances of slavery because "slavery is humiliating", because "I would not want to be enslaved", because "everybody does feel compassion with slaves", or because "slaves will harbour resentment, thus breeding general unrest".[1] In Chapter 4, I assess various ways of determining the epistemological status of beliefs filling this second "because" clause.

[1] The familiar meta-ethical doctrines lingering behind these formulations are perhaps: Moral Realism, Kantianism, sympathy ethics and Contractualism.

2. METHODS

In Chapter 1, I stated the thesis of Sentimentalism. Moral beliefs, as mental entities, are sentiments. Moral judgments express dispositions to experience sentiments of a certain kind. How are we to assess such claims of moral psychology? Three methods come to mind: According to the first, in calling moral beliefs sentiments we make an *empirical* claim. Secondly, it may be thought, the dispute is really *linguistic*. Sentimentalism, here, is an answer to a question of meaning. In a third reading, Sentimentalism is suggested as an *explanatory* device. Given certain features of moral belief, we infer to what has to be the case for mental entities to possess those features.

2.1 Sentimentalism As An Empirical Thesis

In some ways it is most natural to understand Sentimentalism as an empirical hypothesis. Morality after all is an empirical phenomenon, it consists in more than a set of propositions. People in normative communities behave in certain ways. They may not only say "thou shalt not kill" but in fact not kill, feel outraged by killing and punish those who kill. We can investigate empirically how people in normative communities behave and suspect these data to be somehow connected to the people's normative experiences, the way they believe and feel.

How then is an empirical programme of moral psychology to proceed? One promising approach starts from the psychological reality of narrowly moral sentiments. We respond in certain ways to what we take to be moral failure and achievement in ourselves and others. We possess a good intuitive grasp of concepts like respect, indignation, outrage, blame, resentment, guilt and shame. *If we find we can give an account of narrow morality in terms of moral sentiments without giving an account of moral belief we may take moral belief to be explained in terms of moral sentiments.* There may not be more to narrow morality than certain psychological responses captured by sentimental terms.

This claim has recently gained prominence with Gibbard's book *Wise Choices, Apt Feelings: A Theory of Normative Judgment* (1990). Part of Gibbard's larger project is an attempt to revive what he takes to be a Millean tradition in defining narrow morality in terms of sanctions. Gibbard gives the following quote from Mill's *Utilitarianism*:[1] "Morality pertains to what is wrong or not wrong, and to say that an act is wrong is to say that there ought to be a sanction against it, a sanction of law, of public opinion, or of conscience." Gibbard then rejects Mill's emphasis on the law. A penal code, he argues, does not always concern moral matters. "... we do not think overparking morally wrong; we merely think that a price should be charged." Instead Gibbard proposes the sentiments of guilt and anger as the key sanctions of conscience and public opinion.[2] Experimental psychology, anthropology and evolutionary biology can be expected to deliver data on this matter. It may turn out, such at least is the hope of the programme, that there is no role for moral beliefs independently of moral sentiments. We understand moral beliefs only because we have the relevant moral sentiments.

First difficulties here concern how to ask the right questions. Where does the moral scientist look? How does empirical inquiry get started? Hume remarked that the pleasant sentiment of moral approval must be distinguished from other pleasures like

[1] Mill, 1863, ch. 5; Gibbard, 1990, 41.

[2] Along with others (e.g. Sturgeon, 1985, 23, note 2), I do not find Gibbard's Millian credentials very convincing. First, cases where but a price is charged may not fall within the scope of Mill's enterprise. Overparking may be like an exchange of goods, implying a trade agreement. Millian sanctions will come into play only if the terms of the trade agreement are violated, for example, when a charged price is not paid. This, of course, would turn overparking into a moral issue. Only then (this is my second point) Mill's central question arises on what grounds a sanction against the violation of trade agreements, say, can be justified. In other words, Mill is interested in the status and force of the 'ought' in the quoted locution "to say that an act is wrong is to say that there ought to be a sanction against it." Mill mentions the sanctions of public opinion and conscience because sanctions of law may, in particular cases, constitute "inconveniences" (ch. 5, par. 13), i.e. they may be impractical and difficult to enact. However, Gibbard's concern at this point is not, whether and on what grounds we *ought* to sanction (justificatory question) but whether we can *understand* moral belief apart from sentimental sanctions (conceptual question).

Another recent proposal (Tugendhat, 1989) defines narrow morality in terms of sentimental sanctions in first, second and third-person perspective. Named are (with reference to Strawson, 1974) the sentiments of shame, resentment and indignation. Tugendhat's proposal, however, has lesser empirical ambitions then Gibbard's project.

those arising from "A good composition of music and a bottle of good wine" (T. 472). This may not seem too challenging a demand in a Humean world of music, wine and morals. In more profane circumstances it is less obvious how to individuate moral pleasure (if pleasure it is to be). Similarly, in Gibbard's world it is not enough to look at any guilt or any anger, we need to identify moral guilt and anger. The last point opens the door to a more general problem. Shall we direct our attention to moral pleasures or to the darker sentiments of moral censorship, to moral approval or moral sanction? Gibbard invites us to take our chances, make a choice that strikes us as plausible and see how far it might carry us. Let us follow Gibbard onto the side of sanctioning sentiments.

Narrowly moral beliefs in the first-person here concern the sentiments of guilt, remorse or shame. I am prone to experience these internal sanctions if I violate moral norms I accept. The second- and third-person counterparts to guilt, remorse or shame are the sentiments of resentment, indignation, outrage or blame. If someone breaches norms we hold as moral standards we will resent him, blame him, feel outrage and indignation. This seems a plausible first approximation to morality in the narrow sense. In this vocabulary we may expect an anthropologist to be able to tell us a story about a tribe's morality, an experimental psychologist to report his experiments, a socio-biologist to speculate on evolutionary history. The anthropologist might identify the prevalence of a strong concept of guilt in the Catholic Tribe; the psychologist might test the workings of guilt, say, in people who use contraceptives; the socio-biologist might explain the evolutionary rationale behind censoring limited reproduction in particular and behind the workings of guilt in general.

This approach, however, may appear to get things the wrong way around. Should we not explain sentimental sanctions in terms of moral beliefs, not moral beliefs in terms of sentimental sanctions? Though we may possess ample empirical evidence for the psychological reality of moral sentiments we cannot conclude that moral beliefs depend on sentiments. The catholic might feel guilty *because* he believes contraception to be a sin. The evil of contraception, he might hold, can be understood independently of feelings of guilt.

Gibbard is alert to this charge. To avoid circularity, he writes, we need to characterize the moral sanction through a sentiment "that can be felt even by a person who thinks it makes no sense to feel that way" (126). "Indignation" or "outrage" suggest full moral judgment ("contraception is a sin"); "anger" and "guilt", Gibbard claims, are innocent in this respect. They can be empirically investigated outside the sphere of narrow morality. As *adaptive syndromes,* they may even be found in the animal world. Gibbard presents us with a non-cognitive analysis of a dog's anger – a dog whose territory is approached by another dog.

> The dog stands up, it takes a special kind of stance or it runs back and forth, it barks, and it is primed to attack if the other dog keeps approaching. This story combines features of various kinds: a cause (territorial intrusion), expressive behavior (barking and taking threatening stances), and other behavioral tendencies (the dog is primed to attack) ... The combination of these constitute a syndrome, and the emotion is whatever state of the organism is behind the syndrome. (132)

As human adaptive syndrome, anger is coordinated with guilt. Guilt may motivate to placate anger through "apology, restitution, and open contrition" (139). Guilt and anger thus "mesh" (140). They contribute to the development of advantageous cooperative schemes, so Gibbard's evolutionary speculation goes.

To succeed as an empirical account of moral belief and avoid circularity, Gibbard has to show that moral sanctions are indeed forms of guilt and anger and operate as such across cultural borders. Can we on purely empirical grounds reject the competing hypothesis that we have to understand moral belief already if we are to characterize moral sentiments, that the special standpoint of morality is less than continuous with the standpoint of non-moral guilt and anger? If guilt and anger were *universal* adaptive syndromes of the human condition, this would lend overwhelming plausibility to the empirical enterprise of understanding moral belief in terms of moral sentiments. Everything hinges here on the strength of the empirical evidence.

Threatening to Gibbard's theory must be, first, that there seem to exist cultures with a entirely different set of emotional concepts (thus, so called "shame-cultures" are said to possess no concept of guilt in their repertoire[1]), and secondly, that even concepts of evolutionarily old emotions are susceptible to change through time (e.g. the concept of anger of classic antiquity differs in important ways from modern anger[2]). If guilt and anger are universal adaptive syndromes of the human condition, Gibbard admits, how could some cultures have failed to have noticed them (145), and, we should add, how can they disappear through time? Gibbard introduces a second, *attributional* (we might say, sentimentally cognitive) theory in order to cope with the cultural diversity of concepts for emotions. On the attributional picture of emotions, seeing oneself as guilty plays a part in producing the syndrome we call guilt. In this picture, biological adaptations operate only in the background as "tendencies to label and guide emotional agitation with items drawn from one's cultural repertory" (147).

The possible truth of the attributional theory, however, re-enforces the charge of circularity. If the workings of, say, guilt is guided by a view of oneself as guilty, that view may include a belief of one's situation as calling for guilt. This is the way we normally think. Our moral outrage seems to include a belief that a wrong has been done. If it turns out that the act was morally excusable our outrage should cease immediately. We do not blame the blind for not seeing us. Thus it seems that moral anger is specific to the moral sphere. It cannot occur outside narrow morality. It is a cognitive emotion.

Gibbard thinks we are deceived in thinking so. Crucially, he cites the fact that we sometimes feel an emotion when we don't believe we have grounds to feel it. If the

[1] e.g. the Maoris of New Zealand (Smith, 1981). On the standard account, shame is related to public exposure while guilt is internalized: "shame requires an audience", as G. Taylor writes (1985, 57). Williams (1993) rejects this account for the culture of Classical Greece. According to Williams, ancient Greek shame involves a perceived loss of power with respect to an *internalized* observer, while modern Christian guilt is a subject's reaction of fear at the anger of an internalized victim.

[2] Miles Burnyeat in a paper *Anger and Revenge* (1993) argues that the ancient emotion of anger, as epitomized in Achilles' wrath and discussed by Aristotle and the Stoics, implies revenge. "Anger is a craving or desire to punish one who is thought to have done you an undeserved injury." (Diogenes Laertius, VII, 113-4) Only the pleasure of successful redress is apt to extirpate this emotion. Modern, post-Christian anger, however, may be placated by the mere recognition that a wrong has been done. For this, Burnyeat suggests, as little as an offender's listening to you voicing anger may suffice.

occurrence of a moral sentiment of sanction necessarily included the belief that a wrong had been done we could not make sense of such an experience. Think again of a person with a rigid catholic upbringing, is it not the case that such a person may in later life (after renouncing all catholic doctrines) experience feelings of guilt she takes to be unwarranted? If she can feel guilty, though she does not think herself at fault, must it then not be possible to conceive of guilt independently of morality?

This is an interesting argument. The phenomenon of experiencing feelings believed to be unwarranted could be called *weakness of the heart*, i.e. a malformation in the formative process of a feeling. A feeling deviates from what are believed to be its grounds (just as in *weakness of the will* an action deviates from its supposed reasons).[1] For the argument to succeed, I think, Gibbard would have to show that weakness of the heart is not a malformation. From the phenomenon of weakness of the will most people would not conclude that there is no room for intentional explanations of action. Though I sometimes do not sit at this table and write when I intended to, often I do sit at this table and write (as you may testify). I then do so *because* I chose to do so. Weakness of the will does not undermine the force of this explanation. Similarly, one might hold, weakness of the heart does not undermine cases where we feel guilty *because* we believe we are at fault.

There may be ways to unsettle the suggested analogy between weakness of the will and weakness of the heart. *If* we had to admit that emotional inertia is more than a puzzling deformation and that sentimental non-cognitivism is in a better position to cope with the occurrence of emotions believed to be unwarranted, a further argument would have to be called upon. Not all cases of unwarranted emotions, one might argue, speak in favour of sentimental non-cognitivism. For fictional emotions (or better, emotions towards fictions), the cognitivist is in a better position. How, on an empirical non-cognitive account of moral emotions, are we to deal with emotions

[1] In her essay "Actions, Passions, and Reason", Annette Baier uses 'weakness of the heart' in this sense (A. Baier, 1985, 109). Emotional inertia is also discussed in A.O. Rorty, 1980; de Sousa, 1980; Greenspan, 1980; all in A.O. Rorty (ed.), *Explaining Emotions*. See also Gosling's chapter "passionate akrasia" (1990, ch. 10).

towards plays, novels, films, pictures and music? In one central sense, these emotions are unwarranted. In the theatre, I may feel jealous without having a rival, angry without having been offended, frightened without being in danger, perhaps even guilty without having done a wrong. Yet, what I experience may be perfectly appropriate. A cognitive analysis can make some sense of this intuition. As a member of the audience, one might say, I view a situation as if I were a participant. It is the *belief* about a situational cause which makes a fictional emotion appropriate. The non-cognitivist can say no more than that, in the theatre, the biological mechanisms called emotions are out of control; they are somehow mistakenly triggered off.

With fictions, we are always on slippery territory; and I do not claim that a cognitivist account of fictional emotions would run through smoothly.[1] Still, an empirical account of the nature of moral emotions will finally be tested on the strength of its empirical evidence and not by philosophical arguments claiming that such an empirical account must be possible. Gibbard's evolutionary speculations, as far as they go, I take to be inconclusive. However, they are not necessarily inconclusive. Richer empirical investigation may shed new light on Gibbard's enterprise.

[1] In Mozart's *Cosi fan tutte*, for example, it may seem appropriate to feel Dorabella's and Fiordiligi's emptiness even boredom behind their official grief with the strange, circulating D major motifs at the beginning of the first *finale* ("Ah che tutta in un momento"). Were I totally unfamiliar with the musical vocabulary of late eighteenth century my experiences would differ significantly. So it seems plain that a cognitive process must have taken place, appropriating what I feel.

On the other hand, the emotions the average Hollywood film of the Eighties and Nineties operates with, we may find evolutionarily primitive, that is non-cognitive. We sit in a thriller, we know it is fiction, and still we can't help feeling afraid; we see Tom Cruise return, we hear the clapping of the masses, and we cannot help ancient hero worship creeping up our backs.

2.2 Sentimentalism As A Linguistic Thesis

I turn now to a second methodological approach to Sentimentalism. Here, Sentimentalism is to be a theory about the meaning of moral language. This has sometimes been thought to constitute Hume's view:

> when you pronounce any action or character to be vicious, you mean nothing, but that from the constitution of your nature you have a feeling or sentiment of blame from the contemplation of it. (T. 469)[1]

In more recent years, most prominently A.J. Ayer and R.M. Hare have shown some allegiance to this quotation. For Ayer's Logical Positivism only propositions about empirically verifiable data or deductive relations could be meaningfully asserted. Moral statements, according to Ayer, contain neither; they are cognitively empty. They add to the underlying factual statement a certain tone.

> Thus if I say to someone, "You acted wrongly in stealing that money," I am not stating anything more than if I had simply said, "You stole that money." In adding that this action is wrong I am not making any further statement about it. I am simply evincing my moral disapproval of it. (Ayer, 1946 (1936), 107)

Hare, on the other hand, holds that by "investigating the meaning of the moral words" (Hare, 1981, 20) moral judgments turn out to express *universal* approvals or disapprovals in an imperatival form:

> [Universalizability] comes to this, that if we make different moral judgements about situations we admit to be identical in their universal descriptive properties, we contradict ourselves... The prescriptivity of moral judgements can be explained formally as the property of entailing at least one imperative.[2]

Do these statements give an account of the actual usage of moral language? Since G.E. Moore's time[3] there has been a standard argument against sentimentalist accounts of meaning. If moral judgments are in some way about a speaker's psychological state, Moore said, it would preempt meaningful disagreement towards

[1] For now, I am not concerned with the reasons we might or might not have for attributing Sentimentalism in meaning to this quotation and Hume. (In the Appendix, I shall say more on this matter).

[2] Hare, 1981, 21. For a fuller discussion of Hare, see Chapters 3.5 and 4.2 below.

[3] Moore, 1912, 91. Sidgwick (1907 (1874), 26) offers a similar argument.

any given act, character or situation since apparently conflicting judgments would be mutually consistent. However, it is part of the ordinary use of moral language to meaningfully disagree; therefore moral judgment cannot express subjective states.

Moore's scapegoat was a caricature subjectivist position of which we do not know whether anybody actually held it.[1] It is hardly plausible that in judging morally we actually *say* that we are in a certain state. To assert that I am in a certain state of mind, however, is distinct from expressing that state – as it is in the case of non-normative judgment: In judging that *p* I normally do not assert that I judge that *p*; I do not talk about myself, I talk about *p*. The question must be about the content of *p*. Perhaps then Moore's claim can be transformed into a semantic argument: The contents expressed by moral judgments are such that they allows certain inferences subjective contents would not allow. This we find by looking at semantic features of our actual moral language.

Moral predicates seem to behave like ordinary predicates. We assert of something that it is good (or bad) just as we assert that it is green or round. We negate moral claims, we even make valid inferences from moral premises. Now we have a simple semantic theory that tells us how that is possible. You have claimed, say, that (1) "Slavery is wrong", conditionalized (2) "If slavery is wrong, then preaching slavery is wrong, too" and concluded from (1) and (2) that (3) "Preaching slavery is wrong". A Fregean explaining the meaning of sentences as truth-functions of the meaning of their component expressions might give the following account of the situation. If your argument is valid, as we accept it is, the locution "Slavery is wrong" must have the same meaning in (1) and (2). (Otherwise the argument would equivocate). If "Slavery is wrong" expressed a truth-apt belief, sameness of meaning would be guaranteed. But if, on a sentimentalist account of meaning, (1) "Slavery is wrong" expressed your attitude, disposition or approval, the locution "Slavery is wrong" in (2) would carry a different meaning since in (2) you do not *express* your

[1] It is sometimes wrongly attributed to Hume (e.g. by Harrison, 1976, and in his article on "Ethical Subjectivism" in the *Encyclopaedia of Philosophy*).

approval or disapproval of slavery at all, you conditionalize over it. So the anti-sentimentalist point can be neatly summed up: If the locution "Slavery is wrong" has the same meaning in (1) and (2), the argument from (1) to (3) is valid. Sentimentalism cannot assign the same meaning to the locution "Slavery is wrong" in (1) and (2). But the argument from (1) to (3) is valid. Therefore Sentimentalism cannot be correct as a thesis of meaning.[1]

I do not want to discuss this objection in detail. Others have done that more or less convincingly; and I am confident that a semantic theory can be found that does preserve sameness of meaning in unasserted contexts.[2] What I want to do here is to offer a general argument against the use of certain surface grammatical features of our actual language in order to block accounts of the nature of belief and judgment, in our case, moral belief and judgment.

How closely can a theory about the nature of moral belief depend on the *actual* workings of moral language? Consider an example of embedded contexts from religious language: A believer claims (1) "Praying is pious." Then she concludes from (1) and (2) "If praying is pious, encouraging others to pray is pious too" that (3) "Encouraging others to pray is pious". It may be right that 'pious' is used by true believers descriptively and that its predication is treated as true or false – as Geach claims our best semantic theory must do. But does that mean that 'pious' is a *substantially* truth-apt predicate, that it secures objectivity or reality of a kind at issue in debates about the status of moral belief?

Now it is perfectly possible to conceive of a society in which actual religious or moral disagreement is limited to the *application* of moral norms to particular cases,

[1] This has become known as the Frege-Geach objection. The argument was first sketched in Geach, 1958, 54, note; Geach, 1965, attributes the point to Frege.

[2] Gibbard, 1990, suggests a possible world semantics: "The content of a normative statement is the set of factual-normative worlds for which the statement holds." (97) Here the meaning of "Slavery is wrong" is determined by what it rules out; and that is the same in (1) and (2). Blackburn makes two distinct proposals in 1984 (189-196) and 1988 from the notion of consistently realizable attitudes (cf. Chapter 4.2.1 below). Max Kölbel (1994, 1995) sets out the problem very clearly and rejects Blackburn's solutions. Blackburn (1992b) discusses Gibbard.

where the moral standards themselves remain unquestioned. This society may possess a foundation myth: a book of revelations, a Sermon on the Mount, a Magna Charta etc., containing a manageable list of prescriptions, rights, or perhaps an ultimate monistic maxim. In this society, moral statements are purely descriptive, the ascription of moral predicates truth-conditional. Moral language will be as safe in its use as the Highway Code in the hands of a traffic warden. Still, nobody should hold that reliability in application does tell us anything about the metaphysical status of the Highway Code. The nature of the Highway Code is not revealed by the linguistics of traffic-language. Likewise it may be a mistake to think that moral or religious language can reveal sound modes of normative reflection or distinctively metaphysical beliefs.[1]

This fiction appears to be as damaging to the truth-conditional descriptivist as to a sentimentalist in meaning. One may indeed be tempted into an *error theory* – the view that our moral language is systematically mistaken.[2] Moral or religious arguments, like the ones we considered, only appear to be valid, in fact they are semantically inexplicable.

Sentimentalists in meaning might hope to reply that truth-conditional use of moral terms within a closed society cannot be what moral language really means. Moral beliefs are not like hard-line religious or traffic beliefs. Even hyper-traditional societies move to some degree; they are not static. Moral language must provide for this possibility. In every society there can be someone who asks, without misuse of language, whether a given norm is right. Moral language is often used to question existing norms. Any account of meaning excluding moral reflection in this fundamental sense must be incomplete.

[1] Brandt (1979, 2-10) criticizes appeals to linguistic intuitions as vague and bad guidance for normative reflection. Williams, too, (1985b, ch. 7) opposes what he calls the *linguistic turn* in ethics.

[2] John Mackie has introduced the term *error theory* in *Ethics: Inventing Right and Wrong* (1977, esp. ch. 1, sc. 7 and ch. 1, conclusion).

The question then is not what moral language actually means (at least for the majority of hard-liners) but what it might mean, or perhaps ought to mean, given other empirical and metaphysical beliefs. The answer to this is not given by actual linguistic analysis but by assessment of what we want to be able to say. Sentimentalism in this picture comes out as the meaning we *assign* to moral language in order to account for certain features of moral thinking. Sentimentalism in meaning, thus conceived, is not a linguistic thesis but an explanatory device. This is the understanding of Sentimentalism I support.

2.3 Sentimentalism As An Explanatory Device

Taking sides with this third methodological approach, how is sentimental explanation to proceed? In principle, I think, no different from other philosophical explanation. We ask for philosophical explanation when we find a tension between a corpus of beliefs we accept and a phenomenon we find paradoxical in terms of that corpus, or when we want to understand how a phenomenon we accept can occur.

In the latter mood, the belief-desire theory, for example, may tell us how intentional action is possible (given there is such a thing). It explains an agent's intentional action by setting out his reasons for doing what he did. The explanation stipulates entities such as desires and beliefs by inference. Beliefs and desires cannot be observed directly, they have a controversial phenomenology. When I cross a road, I am not aware of a desire to cross nor, normally, of a representation of the road. I simply cross the road. Yet, if beliefs and desires existed they would elucidate the difference between cases where I cross the road intentionally and cases where I didn't.

Suppose we have successfully explained how my intentionally crossing the road was possible by pointing to my reasons for doing so. Now comes along *weakness of the will* where my reasons for an action apparently do not explain the action I performed. Explanation in a further (i.e. my former) sense is then required. The theoretical corpus of the belief-desire theory we may have come to accept seems to rule out that there is such a thing as weakness of the will. Philosophical explanation then would have to show either how we can be deceived into thinking that a phenomenon like weakness of the will could have occurred, or it would have to revise the underlying assumptions that appeared to rule out weakness of the will as incoherent.

The strategy for a systematic explanation of moral belief is twofold. On the one hand, we need some agreement on the *explanandum*: Which are the central features of moral thinking and argument that call for explanation? Here we must take the lead from prephilosophical conceptions, linguistic intuitions and perhaps empirical research but we are likely to find that the phenomena to be explained are contradictory. This is why appeal to actual usages and empirical data are at best a guidance, at worst misleading (as I have argued throughout sections 2.1 and 2.2): Empirical research in ancient Greece might have revealed that anger was always and only placated by successful redress; yet as history moved on anger began to respond to other considerations. Linguistic intuitions would support inferences from conditionalization upon "pious" as valid; yet "pious" might not be a substantially truth-apt predicate.

Once we have identified features of moral thinking and argument that strike us as central in that they won't give way easily, both historically and conceptually, we then proceed to the second stage of the explanatory project. We suggest a theory about moral belief, i.e. the states of mind that would make these central features possible. If there remains a tension between the corpus of the theory and particular features already identified as central, explanation often takes the form of explaining away.

In this dissertation I have reversed the natural order of the explanatory strategy. In Chapter 1, I boldly outlined one particular theory of moral belief. The task is now to collect the *explananda* and see how Sentimentalism would deal with them. Among the persistent phenomena which I count as central are the following:

(a) The issue of Internalism vs. Externalism: we think both (1) that we abstain from certain actions *because* we believe them to be wrong, and (2) that we perform some actions *though* we believe them to be wrong. (see Ch. 3 below)

(b) The issue of Moral Psychology: many factors are formative for the moral dispositions we acquire (they might fill anevolutionary role, they might concur with our interests, etc.), yet we think what is morally right can be understood independently from these dispositions and their functions. (Chapters 2.1, 4.1)

(c) The issue of Moral Disagreement: Certain considerations uniquely compel us to modify our judgments and attitudes (e. g. if we find some members of the set of our judgments and attitudes to be directly contradictory). Still, moral agreement remains partial and elusive. (Chapters 4.2, 4.3)

In my opinion there can be no complete explanation of these phenomena. Tensions are likely to persist. The Internalist will explain more easily how we come to be moved by normative considerations than how we can find something of value without being moved by it. The opposite holds for the Externalist. No conceptual analysis will state eternal necessary and sufficient conditions for what it means to call something right or wrong. Often, reform is needed – a revision of received usages and opinion. Reform, however, should not go too far if a new explanation is to be the recognisable successor of received concepts.

3. INTERNALISM

The analysis of sentiments stated in Chapter 1 aims at providing an unmysterious account of what it means for a mental state to be practical. If a given sentimental state (e.g. fear) does not lead to paradigmatic behaviour (the symptoms of fear and the avoidance of what is feared, say) we need an explanation why it does not. *A mental state is practical if it requires an explanation of this type.* No stronger tie between mind and behaviour is needed to defend a substantial thesis of practicality. I propose to understand moral beliefs in exactly this way. This seems to require that there is a need for an understanding of moral beliefs as practical. In recent years, something like this has become known as the thesis of Internalism. In the following, I defend a version of Internalism about moral belief.

3.1 Preliminary Remarks

What does the thesis of morality's practicality say? In some weak sense, it is a trivial claim. Throughout history, philosophers have agreed that morality is *somehow* practical. People have pondered about the right and the good as part of an inquiry into what ought to be done. Human action is the focus of moral (or more general, normative) reflection. The thesis, however, becomes controversial if it aspires to do more than characterising the *subject-matter* of moral reflection. The stronger thesis claims that moral beliefs *are* practical. On this reading, moral beliefs are action-guiding in themselves. To entertain a moral belief is to endorse (recommend, prescribe) a certain course of action. For moral beliefs to be practical in this sense becomes a condition of sincerity (namely to be motivated to do what you believe to be right and good), while on the weaker interpretation of the thesis, morality's practicality merely constrains what moral beliefs are *about*. Here sincerity is a separate and substantial claim. You might see something as right and good and not do it. In this sense, it is a *demand* of morality that you do what you believe right and good.

On the strong reading of the thesis of practicality, the issue turns out to be whether it is possible to sincerely believe something to be right or good and, at the same time, not be motivated in some relevant sense. Both the denial and the affirmation of this claim are supported by widely shared intuitions. On one hand, we are susceptible to treating moral considerations as motivational. As Hume puts it:

> ... men are often govern'd by their duties, and deter'd from some actions by the opinion of injustice, and impell'd to others by that of obligation. (T. 457)

and further;

> If morality had naturally no influence on human passions and actions, 'twere in vain to take such pains to inculcate it; (ibid.)

Yet, we also (and again Hume is typically faithful to common sense) will often not do what morality demands of us.

> 'Tis one thing to know virtue, and another to conform the will to it. (T. 465)

This is certainly true of the A- or Immoralist. A vandal, one might say, breaks trees or throws benches on the railway track knowing this to be wrong. It is perhaps even the point of his destructive activity. In a less dramatic sense, most of us are capable of neglecting moral demands we have come to accept. Often, this occurs in areas of our lifes which we can expect to remain shielded from public scrutiny. When an agent becomes aware of this, we call it hypocrisy – "the tribute vice pays to virtue", as La Rochefoucault defined it succinctly.[1] Both the vandal and the hypocrite believe certain things to be right and good but apparently fail to be motivated in the relevant sense.

In recent years, these puzzles have been discussed mostly in the terminology of Internalism vs. Externalism about moral motivation. This way of speaking is not entirely fortunate for two reasons. First, internal/external distinctions are among the most favoured terminological pairs in analytical philosophy, and in some cases they

[1] Francois de la Rochefoucauld, *Reflexions, ou sentences et maximes morales,* 1976 (1665). I borrow the remark from Railton, 1986, 203.

are easily confused.[1] Secondly, it has proved difficult to give a clear statement of what it is for moral motivation to be internal or external. "Internal or external to what?" one is immediately tempted to ask. In his 1958 paper "Obligation and Motivation in Recent Moral Philosophy", William Frankena introduces the distinction in the following way:

> ... the question is not whether or not moral philosophers may or must introduce the topic of motivation. Externalists have generally been concerned about motivation as well as about obligation; they differ from their opponents only about the reason for this concern. Internalists hold that motivation must be provided for because it is involved in the analysis of moral judgments and so is essential for an action's being or being shown to be obligatory. Externalists insist that motivation is not part of the analysis of moral judgments or of the justification of moral claims.[2]

As always in philosophy, it is difficult to change or modify a terminology, once two or three papers have adopted it.[3] In recent years, the blame falls most probably on Thomas Nagel and his book *The Possibility of Altruism*. There he revives Frankena's distinction:

> Internalism is the view that the presence of a motivation for acting morally is guaranteed by the truth of the ethical propositions themselves ... Externalism holds, on the other hand, that the necessary motivation is not supplied by ethical principles or judgments themselves, and that an additional psychological sanction is required to motivate our compliance.[4]

The philosophical problem, of course, is much older than Falk, Frankena or Nagel. It makes its earliest appearance in Plato's *Protagoras*. There Socrates vexes Protagoras with the following questions:

> What is your attitude to knowledge? Do you share the common view about that also? The opinion generally held of knowledge is that it is nothing strong, no guiding or governing thing... Is this your view too, or would you rather say

[1] A particularly dangerous, since related distinction is suggested in Bernard Williams' paper "Internal and External Reasons" (Williams, 1980). There Williams asks whether there can be reasons which are not relative to an agent's motivational set.

[2] Frankena (1958, 40). Frankena credits W.D. Falk (1948) with the invention of the internalist/externalist labels.

[3] Hare's analysis of moral language as *prescriptive* (1952) and Gibbard's account of normative judgment as expressing the *acceptance of norms* (1990) are attempts at alternative, perhaps more successful, terminologies.

[4] Nagel, 1970, 7.

> that knowledge is a fine thing quite capable of ruling a man, and that if he can distinguish good from evil, nothing will force him to act otherwise than as his knowledge dictates?[1]

Plato, like Nagel, casts the problem in terms of moral knowledge not moral belief. Can we *know* some course of action to be right and still fail to be motivated in any relevant sense? This way of putting the question leaves things rather muddled. A moral cognitivist typically holds that moral beliefs can be a matter of knowledge. They are, the cognitivist often says, apt for truth-evaluation. Yet, do moral beliefs motivate only if they are secured to be true? Even a moral cognitivist can change his mind. Was then his previous mental state (which turned out to fall short of knowledge) less motivating than his subsequent state (now thought to constitute knowledge)? It seems not. Plato's and Nagel's considered thesis should be that moral beliefs can be true or false but that a moral belief's motivating force does not depend on its truth or falsehood.

Both Nagel's and Plato's positions seem to indicate a link of Internalism to some form of moral cognitivism, but they differ in one important way. How are moral beliefs thought to be motivating? Plato speaks of the *dictate* of moral knowledge while for Nagel to be in a state of moral knowledge is to be motivated merely *prima facie*. Nagel writes:

> Internalism's appeal derives from the conviction that one cannot accept or assert sincerely any ethical proposition without accepting at least a prima facie motivation for action in accordance with it.[2]

[1] *Protagoras* 352b. This is the perhaps genuinely Socratic view of the early Plato. It does not appear to allow for the notion of conflict between parts of the soul the mature Plato maps out in the *Republic*. The claim that no one is voluntarily bad is now generally taken to constitute two distinct paradoxes: (1) the prudential paradox that "no one desires evil things and that all who pursue evil things do so involuntarily"; and (2) the ethical paradox that "virtue is knowledge and that all who do injustice or wrong do so involuntarily" (Santas, 1964, 147).

[2] Nagel (1970), ibid. In order to clarify the difference between "dictating" and "prima facie" motivation it may be thought helpful to couch the distinction in terms of reasons. To be motivated "prima facie" is to possess *a* reason for action, to be motivationally "dictated" is to have an overriding reason to act. For the moment, I am reluctant to adopt this way of speaking for it imports Williams' problem of external and internal reasons (1980). Can I have a reason for something which is not part of my motivational set? That is, can I have a reason not to smoke even though I am not motivated to stop smoking? Dancy (1993, 253) claims: "A Nagelian internalist is not committed either to accepting or to denying the existence of external reasons in Williams' sense; an internalist may allow that there are external reasons

The distinction between *prima facie* and overriding motivation is crucial though, I believe, one should resist tying Internalism to moral cognitivism. Armed with these initial distinctions it is perhaps time to step back and survey the field. As I see it, we have at least the following options. Internalism (1) may or may not require moral cognitivism. Again, on the cognitive (1)(i) as well as on the non-cognitive (1)(ii) version we may opt for an account of moral motivation as *prima facie* (1)(i)(a)/(1)(ii)(a) or overriding (1)(i)(b)/(1)(ii)(b). On the other side, Externalism (2) may or may not go together with moral cognitivism. We can conceive of cognitive (2)(i) as well as non-cognitive (2)(ii) Externalism. Yet, a distinction between *prima facie* and overriding *external* motivation can be avoided. Let us map out these six options.

3.2 Options

(1) Internalism (roughly the claim that there is an intrinsic or internal connection between moral belief and action).

(1)(i) Cognitive Internalism: Beliefs about the right and good are a matter of knowledge. These beliefs intrinsically motivate us to perform certain actions and abstain from others.

(1)(i)(a) Cognitive Overriding Internalism: To hold a moral belief (which is a suitable matter of knowledge) is to be overridingly motivated. The spirit of this view is perhaps best captured by the Chinese Saying "To know and not to act is not to know".[1]

(in Williams' sense), so long as when those reasons come to motivate they require no additional psychological sanction to motivate our compliance. Equally, a Nagelian externalist can happily hold that there are no external reasons. An externalist only holds that where a moral truth or judgement is a reason, it still requires some additional psychological sanction to motivate our compliance."

[1] Nadine Gordimer in *Burger's Daughter* (1980 (1979), 213) attributes the saying to Wang Yang-ming. Cognitive Overriding Internalism is perhaps currently the most prevalent version of moral realism.

(1)(i)(b) Cognitive *prima facie* Internalism: To hold a moral belief (which is a suitable matter of knowledge) is to be motivated *prima facie*. Moral considerations, however, depend for their causal efficacy on certain background conditions and may be overridden by other considerations. This is probably the position of common sense.[1]

(1)(ii) Non-Cognitive Internalism: Moral beliefs are not a suitable matter of knowledge. What we believe to be right and good, however, is intrinsically motivating.

(1)(ii)(a) Non-Cognitive Overriding Internalism: To hold a moral belief (which is not a suitable matter of knowledge) is to be overridingly motivated. The most elaborate version of this kind of Internalism is Richard Hare's Universal Prescriptivism.[2]

McDowell defends it in a well-known series of papers (1978; 1979; 1981; 1983; 1985) where he argues that it is part of the concept of a moral consideration that it is motivating. For the virtuous person there is no possibility that something other than the right action is done. McDowell is adamant to contrast overriding and silencing reasons (1978, 26). In the present context, I believe, not much hinges on the distinction. Related forms of Internalism are held by Raimond Gaita (1991) and Mark Platts – e.g. in *Moral Realities* (1991) and his earlier paper "Moral Reality and The End of Desire" (1980). There Platts suggests that anyone who claims to recognise an act as honest, say, but fails to see it as desirable has not in fact seen it as honest. Historically, Cognitive Overriding Internalists include Plato, some of the so-called "eighteenth century British Moralists" (e.g. Butler, Samuel Clarke and Richard Price) as well as Post-Wittgensteinian of the late 1950s (Foot, Anscombe). In her early writings (1958a, 1958b – before turning Externalist with "Goodness and Choice", 1961; cf. 1972a and 1972b), Philippa Foot suggests that moral motivation might be tied to naturalistic content understood as a suitable object of knowledge.

[1] Even Plato's Socrates admits that most people believe that moral considerations can be defeated: "They [the people] maintain that there are many who recognise the best but are unwilling to act on it. It may be open to them but they do otherwise." (*Protagoras*, 352d). Jonathan Dancy holds that what he calls "intrinsically motivating states" are for their causal efficacy dependent on certain background conditions. "... a state which is here sufficient for action might elsewhere not be." (1993, 25) Dancy then pursues an analogy with the theory of causation. "We might allow that the causes of my attending a conference in the US would not have been sufficient if the US had recently declared war on England, without accepting the fact that this had not happened was among the causes of my presence there". (24) Wiggins (1990, 82) writes: "... we need not disturb the claim of necessary connection between our thoughts of value and our having defeasible reasons of some kind." In a reply to Peter Railton (1993, 307), Wiggins concedes that the motivational force of a moral requirement "need not necessarily be a reason that can under all circumstances outweigh all others".

[2] Hare believes that moral preferences (expressed in moral judgments as prescriptions) are distinguished from other kinds of desirability by being universalizable and overriding. "... 'ought' aspires to the status of 'must', and, as we shall see, in rigorous, critical, moral reasoning has to be used like it. I shall therefore ... continue to use the word 'ought', with the proviso that it is to be used in our reasoning *as if* it were always fully prescriptive, and *as if* its prescriptions were not to be overridden ..." (1981, 24, cf. 60-1). Hare is perhaps not an *epistemological* non-cognitivist: "If to think that

(1)(ii)(b) Non-Cognitive *prima facie* Internalism: To hold a moral belief (which is not a suitable matter of knowledge) is to be motivated *prima facie*. This version seems to be implicit in many kinds of Internalism but I do not know of a detailed account.[1]

(2) Externalism (roughly the claim that there is a contingent or external connection between moral belief and action).

(2)(i) Cognitive Externalism: Beliefs about the right and good are a matter of knowledge. These beliefs motivate only in conjunction with contingent moral dispositions. Of this position, there exists at least a Neo-Aristotelian and a Scientistic version.[2]

[prescriptive questions] can be determined rationally is to have an epistemology or theory of knowledge, then one who thinks this, as I do, should perhaps be labelled a cognitivist. But I do not recommend the label, because those who are unable to envisage any other kind of reasoning than factual will think that if I am a cognitivist I must be a descriptivist, which I am not." (Hare, 1989 (1985), 97) Perhaps we should formulate non-cognitive Internalism as "Moral beliefs, though intrinsically motivating, are not a matter of *descriptive* knowledge". On "intrinsic action-guidingness", see also Mackie (1977, 23; 1980, 54).

[1] Stevenson (1937, 13) says: "'Goodness' must have, so to speak, a magnetism. A person who recognizes X to be 'good' must ipso facto acquire a stronger tendency to act in its favor than he otherwise would have had." Gibbard (1990, 56) supports the common-sensical notion "that the acceptance of a norm is motivating, at least to a degree: believing I ought to stop tends to make me stop." Otherwise weakness of will were the rule and coordinated activity impossible. "Humans plan together; they make agreements; they exhort. Their language facilitates both complex coordination among individuals and complex individual plans – and if words lacked all power to move us, none of these things would be possible. Words, then, must motivate." (57) Gibbard, however, believes that different motivational systems can be in conflict.

[2] Neo-Aristotelians like Peter Geach (1977), Philippa Foot (1978) and Martha Nussbaum hold that the "goal of human choice ... is *eudaimonia* or "human florishing"". Virtues, then, are "modes of characteristic human function" (Nussbaum, 1992, 10, 11). Here it is an external facts that the exercise of virtue makes human beings flourish. McDowell has recently challenged the externalist interpretation of Aristotle according to which there could be "standards of worth-whileness that any human being ... could accept, independently of any acquired values and the motivational dispositions that are associated with it". (McDowell, 1994, 2). For the Utilitarians, Sidgwick has defended some kind of Cognitive Externalism (1907, 498-503). More recently a American school of philosophers (sometimes called the "Cornell-Realists") has become influential likening moral facts to facts in science (Sturgeon, Boyd, Brink, Railton).

(2)(ii) Non-Cognitive Externalism: Beliefs about the right and good are not a matter of knowledge but perhaps of opinion or preference. Again, it is contingent whether these opinions or preferences become motivational. This might be thought to be the position of a Moral Cynic.

On the externalist side, of course, there is no corresponding distinction between *prima facie* and overriding motivation. Why? The fact that on the externalist picture the motivational pull is but contingent excludes that motivations can be necessarily overriding. Contingent in these circumstances means precisely that other considerations may play a part.

It is now claimed that some form of Internalism must be true. This I seek to establish by an argument from the remainder. Having mapped out the options, it turns out that all versions of Externalism are flawed in some irreparable way.

3.3 Non-Cognitive Externalism – The Moral Cynic

First, I shall briefly consider position (2)(ii), the position I dubbed Moral Cynicism. Moral beliefs, the Moral Cynic announces, are up to him. They are mere opinions or preferences, and preferences again are to be followed randomly. Sometimes the Moral Cynic does what he prefers, sometimes he doesn't. There is no recognisable pattern to his motivation. His beliefs and actions, he says, "float". Is such a person a coherent possibility? Are preferences that do not matter preferences?

Take again the case of vandalism. An Honest Citizen might declare "The vandal wilfully does what he believes to be wrong. This makes vandalism particularly evil." Is a vandal as seen by the Honest Citizen a Moral Cynic as in (2)(ii)? I think he is not. The Honest Citizen treats the evil of throwing benches on the railway track as a matter of fact. He then imputes this view to the vandal. On the non-cognitivist reading, however, the Moral Cynic thinks of the wrong of vandalism not as a matter of fact but of preference – a preference he professes not to care about. This sounds incoherent. If the Moral Cynic says he prefers an action A to an action B, yet, when faced with a choice consistently chooses B over A he cannot be said to have preferred A in the first place.[1] The coherent view for the Moral Cynic, therefore, must be to deny that moral beliefs are preferences, or even that there are such things as preferences. In the first case he is not a Cynic in the sense of (2)(ii) but (2)(i); he takes moral facts to be independent of his preferences but is not motivated by them. In the second case, he cannot hold a view on the relation of moral belief and action at all. A Moral Cynic as in (2)(ii) is deceived about what his beliefs are. He says things he does not really think. Moral Cynics in this sense may exist but cannot be made intelligible.

[1] Hare (1981, 21) makes a similar case: "... if we say [of a hotel] that it is a better hotel than the one on the other side of the road, there is a sense of 'better than' (the prescriptive sense) in which a person who assented orally to our judgement, yet, when faced with a choice between the two hotels (other things such as price being equal), chose the other hotel, must have been saying something he did not really think."

3.4 Cognitive Externalism

I turn now to (2)(i), the position introduced as Cognitive Externalism. For brevity's sake, this view is sometimes called Factualism. This label seems apt to distinguish externalist kinds of moral realism from their internalist cousins. For the present expository purposes, Neo-Aristotelianism and Scientism as versions of Factualism do not differ sufficiently to warrant separate treatment. Though I shall concentrate on Factualism of the scientific kind, the central points should apply equally to all kinds of Externalism.

3.4.1 Internalism vs. Externalism: A Terminological Dispute?

Is there an interesting account of moral facts conceived of as external? It may seem that any account of moral fact as part of "the fabric of the world"[1] is vulnerable to a well rehearsed line of argument. We do not perceive values directly as we perceive that a table is round, nor do we need to postulate such things in order to fully explain the natural world (including human minds). The objection then has to do with a general principle of economy: We ought not to populate the world with things we do not need. "If there were objective values, then they would be entities or qualities or relations of a very strange sort, utterly different from anything else in the universe." This is the substance of Mackie's "argument from queerness".[2]

Part of the persuasiveness of the argument derives from the epistemological worry how values as parts of the universe could stand in relation to us. Even if there exists the GOOD, one likes to challenge proponents of objective values, how does it reach out to us? Why should we care about the "eternal fitness of things", and why

[1] Mackie's discussion of objectivity in *Ethics: Inventing Right and Wrong* (Mackie, 1977, ch. 1) centres around this formulation. See also Williams' paper "Ethics and the Fabric of the World" (1985a).

[2] Mackie, 1977, ch. 1 sect. 9. Harman, too, stresses the need for an explanatory role of entities we postulate (1977, ch. 1, esp. 6-7).

should the "eternal fitness of things" care about us? Historically the main force of the argument from queerness thus was directed at Internalists, i.e. theorists who believed *both* that moral facts existed independently from human concerns and that moral facts intrinsically motivated those who perceive them.[1] The Factualist denies precisely this conjunction. He claims that we can know moral facts but are not necessarily motivated by them. Has the argument from queerness any force against him?

As things stand, Factualists accept the principle of explanatory economy but they claim that, as externalists, they are less vulnerable to the argument from queerness. They do not have to show how some features of the world are intrinsically motivating, the onus is merely to establish an explanatory need for moral facts of an external kind. This remains an ambitious task.

One of the first problems faced by the Factualist is how to select moral facts from a list of facts of a natural kind. We need a preconception of the sphere in which we are to find moral facts. The challenge is here that there are not too few but too many facts in question. We have moral intuitions, for example, that morality has somehow to do with (a) individual human well-being. This may lead us to analyse judgments as similar to, say, "mountain air is beneficial to tuberculosis" which by Neo-Aristotelians may be thought to be a crucial judgment of fact. Alternatively we may think, for example, that it is a fact about (b) the institution of morality that it promotes social cooperation and stability. Or we might characterize (c) the moral point of view formally as e.g. impartial in that it seems to exclude the use of indexicals. This too might constitute a moral fact. From this list of substantial moral intuitions, facts about the social institution of morality and formal constraints (or shall

[1] The phrase "the eternal fitness of things" stems from the 18th century moralist Samuel Clarke. He writes: "And by this understanding or knowledge of the natural and necessary relations, fitnesses, and proportions of things, the *wills* likewise of all intelligent beings are constantly directed and must needs be determined to act accordingly." (Clarke, 1969 (1706), 198-90). In response to such "fitness"-claims, Hume invented what is perhaps the first argument from queerness: "Take any action allow'd to be vicious: Wilful murder, for instance. Examine it in all lights, and see if you can find that matter of fact, or real existence, which you call *vice*. In which-ever way you take it, you find only certain passions, motives, volitions and thoughts. There is no matter of fact in the case. The vice entirely escapes you, as long as you consider the object. You never can find it, till you turn your reflection into your own breast, and find a sentiment of disapprobation, which arises in you, towards this action." (T. 468-9)

we say: facts of meaning) we select not easily on empirical grounds, hence it may be argued that by directing moral inquiry into one area of facts important decisions have already been made. As we shall see, the Factualist does indeed seek to combine conceptual claims with facts of individual well-being and social engineering:

> ... the notion of social justice might pick out an array of conditions that enhance the possibility of psychologically self-respecting and attractive individual lives while at the same time promoting social cooperation, stability, and prosperity.[1]

While my first reservation urges but caution about the scientific aspirations of some Factualists, a second objection I take to be more serious. Suppose we have arrived at some kind of convergence about the facts in question. Would then the nature of a factualistic moral inquiry be sufficiently characterized as the confirmation by scientific means of the instantiation of properties one has already identified as moral?[2]

There is a good tradition of theorists denying the possibility of a coherent conception of *moral* facts who do not deny that there are facts *about* morality. Often they engage themselves in speculations about what might be the "object of morality" (as Mackie titles a chapter in his *Ethics*). Morality, he suggests, is "a device for counteracting limited sympathies", a device "which is beneficial because of certain contingent features of the human condition".[3]

> If men had been overwhelmingly benevolent, if each had aimed only at the happiness of all, if everyone had loved his neighbour as himself, there would have been no need for the rules that constitute justice. Nor would there have been any need for them if nature had supplied abundantly, and without any effort on our part, all we could want, if food and warmth had been as inexhaustibly available as, until recently, air and water seemed to be. The making and keeping of promises and bargains is a device that makes possible mutually beneficial cooperation between people whose motives are mainly selfish, where the contributions of the different parties need to be made at different times.

[1] Thus Darwall, Gibbard and Railton (1992, 170) characterize one moral cluster-property of a natural kind as envisaged by Boyd (1988).

[2] This is David Copp's definition of "confirmalism" in his "Explanation and Justification in Ethics" (1990).

[3] Mackie (1977) 107, 110. Mackie attributes these, or similar views to Protagoras (in Plato's dialogue), Hobbes (chs. 13-17 of the *Leviathan*), Hume (*Treatise* III ii) and Warnock (1971).

These, presumably, are statements of fact – though admittedly on speculative evolutionary grounds. The qualifications on justice proposed here sound suspiciously like the Factualist's "array of conditions" above. Still, Mackie and other Non-Cognitivists did not feel tempted to treat facts of this kind as *moral* facts. Why? One must assume, because these facts seem to be in no recognisable sense *normative*. We may, for example, not see how these facts could play a part in the explanations of individual behaviour. We say "I apologised because I had offended her", but is it similarly explanatory to say "I apologised because morality is a device for counteracting limited sympathies"? This gives rise to the charge that the dispute between Internalism and Externalism is really a terminological dispute. They answer to two different questions. Externalism, one might say, seeks to identify the purpose of morality, it might even succeed in specifying a moral point of view. Internalism, on the other hand, deals with normative problems, with the adoption of a moral point of view, with what one ought to do.[1] If this is right, we could be Externalists and Internalists at the same time. It would then be difficult to derive any meta-ethical conclusions from the truth of either Internalism or Externalism. This would be an unwelcome end to the dispute for both parties. I shall devote now some time to the exposition of one externalist position which has taken great care to avoid this cul-de-sac.

[1] Peter Singer has suggested a related, terminological explanation of the long-running debate over 'Is' and 'Ought'. For the Non-Cognitive Internalist (or Prescriptivist), there is a gap between factual beliefs and moral judgments, for the Cognitive Externalist (Factualist or Descriptivist) the gap remains between moral judgments and dispositions for action. (Singer, 1973). Wiggins on some counts is an Externalist about value as well as an Internalist about obligation. "... we then conceive of a distinction between *is* and *must* as corresponding to the distinction between appreciation and decision ..." (1987 (1976), 96).

3.4.2 Railton's Externalism

Peter Railton in his paper "Moral Realism"[1] states plainly that the description of moral inquiry as the instantiation of moral properties by empirical means falls short of what is required of an externalist project. The aim must be to establish a "linkage of the normative to the empirical" (163). It is not sufficient to find out what is right and good and suggest empirical procedures for actual cases, we further need to know how such identifications impinge on individual and collective choices of action, moral learning and other forms of normative orientation. Both Internalism and Externalism seek to give an analysis of the same "central evaluative functions", the factualistic ambition being that such an analysis "could be carried out within existing (or prospective) empirical theories" (164). In a more recent paper, the stronger claim is made that "without a suitable account of the normativity of ... purported moral properties, one could not identify them as *moral* properties"[2]. On this view, the search for merely empirical ethical procedures is not only insufficient but mistaken.

Railton attempts to bridge the gap between the empirical and the normative with a method he calls "critical explanation relative to objective interests" (cf. 187-8): "facts exist about what individuals have reason to do, facts that may be substantially independent of, and more normatively compelling than, an agent's occurrent conception of his reasons." (189) This works in two ways. On the one hand, it delivers *reasons* which are typical for long-term thinking. Though occurrently I may have reason to smoke (because it promises an immediate and unique satisfaction) I have a better reason – an "objective interest" – not to smoke (because it seems likely to prolong the years of other immediate and unique satisfactions). And similarly, I may have an occurrent reason (the prospect of imminent pain) not to go to the dentist, yet I have a better reason (the prevention of future pain) to go now. On the other hand,

[1] Railton, 1986. Quotations from "Moral Realism" in this chapter will be revealed by page numbers only.

[2] Darwall/Gibbard/Railton 1992, 128, n.30.

these inoccurrent objective reasons have a role to play in the *explanation* of individual behaviour. My objective interest in seeing the dentist preventatively may explain the evolution of non-deliberative habits which are in accordance with my long-term objective interests.

> ... as children we may have been virtually incapable of making rational assessments when a distant gain involved a proximate loss. Yet somehow over time we managed in largely nondeliberative ways to acquire various interesting habits, such as putting certain vivid thoughts about the immediate future at the periphery of our attention ... (187)

This account of individual rationality seems to be both empirical and normative in the required sense. What constitutes my objective interest may be a matter of empirical investigation while at the same time being normatively compelling and explanatory of individual behaviour.

Moral inquiry now extends this type of critical explanation to collective interests. For Railton, moral facts are constituted by what is "instrumentally rational from a social point of view" (200). There is a need to explain why people act as they do, and these explanations partly operate on a collective level and take again account of "inoccurrent" or non-subjective interests.

> When we seek to explain why people act as they do, why they have certain values or desires, and why sometimes they are led into conflict and other times into cooperation, it comes naturally to common sense and social science alike to talk in terms of people's interests. Such explanations will be incomplete and superficial if we remain wholly at the level of subjective interests, since these, too, must be accounted for. (184)

Railton briefly pursues an analogy with the explanation of "the world's consumption of refined sugar" (184, n.24). Again it would be insufficient to cite as explanatory the fact that people simply liked the taste of sugar:

> Facts about the way we are constituted, about the rather singular ways sugar therefore affects us, and about the ways forms of production and patterns of consumption co-evolved to generate both a growing demand and an expanding supply, must supplement a theory that stops at the level of subjective preferences. (ibid.)

What are the facts involved in the explanation of social rationality? Railton gives the following rather familiar characterizations. Social rationality requires avoidance of certain sorts of dissatisfaction. Thus persistent discounting of the interests of a particular group has the potential for social unrest.

> ... certain social and historical circumstances favor the realization of this potential for unrest, for example, by providing members of this group with experiences that make them more likely to develop interest-congruent wants, by weakening the existing repressive apparatus, by giving them new access to resources or new opportunities for mobilization, or merely by dispelling the illusion that change is impossible. (191-2)

This is, however, compatible with the possibility that under less favourable social and historical circumstances the discrimination of one group may not breed unrest and thus in no way endanger social peace. Under such circumstances what is rational from a social point of view may not be "just", as we normally understand this term. As a matter of historical fact, unjust societies do often flourish. Railton's position is thus vulnerable to a line of objections normally addressed at contractualistic moral theories. Children, the handicapped, the elderly, future generations (to name but a few) all constitute social groups that may never be in a position to press effectively for better social arrangements on their behalf. Such groups may settle for a peace we should call less than just. Railton concedes in a footnote that his account must rely on the possibility of individuals including "other individuals within their own interests" (194, n.35). Thus parents may fight for their children, present generations for future generations, and so forth. Still this does not guarantee, as Railton admits, that there will be "a univocal trend toward greater social rationality" (194). Railton believes he needs for his realism about facts of social rationality no more than that such facts explain a shift towards, say, more equal distribution of resources *when* it occurs. This is said to be supported by recent work in "social history and historical sociology" (192-3).

In the present context, i.e. the issue of Internalism and Externalism, the problems touched in the last paragraph are not central.[1] Let us grant Railton the existence of somehow satisfactory facts of social rationality. There is little doubt that explanations in terms of social rationality should then fall roughly within the realm of empirical investigation (though evolutionary, sociological and perhaps psychological facts may be rather softer than Railton's analogy to our constitutional predilection for

[1] In Chapter 4.3 I discuss contractualism in more detail.

sugar suggests). But how do facts of social rationality satisfy the requirement of normativity? How can the empirical be linked to the normative? In Railton's view, facts about what is rational from a social point of view are sufficiently normative (i.e. qualify as genuinely moral facts, as characteristic for our "central evaluative functions") if "moral rightness could participate in explanations of behavior or in processes of moral learning that parallel explanatory uses of the notion of degrees of individual rationality" (191).

> Morality surely can remain prescriptive within an instrumental framework, and can recommend itself to us in much the same way that, say, epistemology does: various significant and enduring – though perhaps not universal – human ends can be advanced if we apply certain evaluative criteria to our actions. (170)

One of these contingent human ends is for example an interest in impartial justification. "... in public discourse and private reflection we are often concerned with whether our thinking is warranted in a sense that is more intimately connected with its truth-conduciveness than with its instrumentality to our peculiar personal goals ... (202)."

Though most of us have such contingent inclinations, our beliefs in certain moral truths do not necessarily influence our conduct. This is the burden of the externalistic thesis. Our central evaluative function, Railton thinks, is seeing that something is of value (which under favourable circumstances carries a contingent normative commendation). Thus Railton defends a crucial distinction: *Observing* that one thing is more valuable than another should not be conflated with *valuing* one thing over another. The latter is intrinsically connected to desire and action, the former is not (cf. 168). Does a coherent conception of moral belief allow for this distinction?

3.4.3 Norms and States of Mind – The Highway Code

At the beginning of this chapter, I tried to do some justice to the intuitive appeal of divorcing moral belief and moral motivation by introducing the cases of vandalism and hypocrisy. Railton uses Hume's somewhat related example of the Sensible Knave from the end of the *Second Enquiry* (E. 282). According to Hume, justice is conventional. Our obligation to justice arises from the mutual advantage each of us enjoys from reliable cooperation and the stability of property. Justice is thus based on self-interest, i.e. on an instrumental conception of rationality. Hume's problem is that on this conception, a Sensible Knave has no reason to be just. He is *sensible* in that, outwardly, he performs all his public duties, keeps his promises, respects other people's property etc. – thus receiving all the benefits arising from a stable society; he is a *knave* in exploiting any opportunity to steal, cheat and break his word as long as he can get away with it. As a Sensible Knave he does better than as an Honest Citizen.

Railton then wonders whether the Sensible Knave might not accept that justice, as a matter of fact, is "directed at the general welfare" (168) and still admit that he is unjust. Under this reading, the Sensible Knave entertains a belief but is not motivated by it. What kind of belief is "justice is directed at the general welfare"? Railton writes:

> This is in a recognizable sense an evaluative or normative notion – "a value" in the loose sense in which this term is used in such debates... (168)

Compare Railton's "value" belief with the belief I ascribed earlier to the vandal: "Throwing benches on railway tracks is wrong". On the face of it, this also seems to be a straightforward moral belief without motivating character. Yet here a solution to the puzzle suggests itself easily. The vandal, one might say, uses 'wrong' in the sense of the moral community he lives in. He does not believe that throwing benches on railway tracks is wrong, he believes that people *call* throwing benches on railway tracks wrong. The vandal himself subscribes to different values.[1]

[1] Richard Hare introduced the related notion of *inverted commas* judgments. In inverted commas use, "'I ought to do X' becomes roughly equivalent to 'X is required in order to conform to a standard which people in general accept'." (Hare, 1952, 11.2, 167; also sections 7.5, 9.3; cf. Hare, 1963, 10.2 and Hare

Hume's example of the Sensible Knave (as well as Railton's exposition of social rationality) seems to resist this move. Railton and the Sensible Knave both seem to *accept* an account of justice (that is, the morally best) they do not feel committed by. Are value beliefs in the sense we need of a kind with "justice is directed at the general welfare" or "morally best is what is instrumentally rational from a social point of view"? In Chapter 2.1, although resisting an empirical analysis of the moral sentiments, I found plenty of scope for empirical inquiry into morality. How does Railton's moral realism measure up to this claim? To put it differently, could it be that the norms people *de facto* follow in a particular society have only a contingent connection to the norms they accept?

This is a complex issue since people, in a modern society at least, cannot be said to follow one and the same set of norms. Though a core of norms is enforced by the Law, these norms can be questioned and do not map with the moral domain. The point I wish to bring out can be best located if we concentrate on a finite, limited set of norms where we know where to look in empirical inquiry. Once again the Highway Code provides an enlightening example. Morality, I contend, shares some relevant features with traffic rules.

Imagine a social scientist, a specialist in traffic research, who being ignorant of the customs of this country seeks to extract the Highway Code by doing field-work. She stands a good chance to end up with at least an approximation to the Highway Code as it is enshrined by Law. She will find people driving on the left side; occasionally, there will be drivers who jump red lights though most times they don't; and some rules she might consider as sensible (such as the keeping of braking-distances) do not appear to be in place at all – here, she may get it wrong since some rules which are part of the Highway Code are simply not enforced. With all the data collected by the social scientist, what can we say about the beliefs of the road user? Are *traffic beliefs*, as we might call them, practical?

1981, 3.7, 58). For a closer discussion of the problem of amoralism, moral weakness and Hare's response to it, see the following section.

It seems to me clear that traffic beliefs must be practical. My belief that it is part of the Highway Code in Great Britain to drive on the left side is non-contingently connected to my driving on the left side. Questioned, why I am driving on the left side, I shall respond that I believe it to be part of the Highway Code in Great Britain to drive on the left side. I drive on the left side *because* I believe this. Of course, a Factualist may object that I drive on the left side out of an instinct of self-preservation or because I fear sanctions for not complying with the Highway Code. Traffic beliefs, the Factualist might say, are really cases of valuing something not of observing that something is of value. Here, however, he misses the point. I did not suggest that traffic beliefs *are* moral beliefs. I said that morality as a system of norms is open to empirical inquiry just as are traffic rules. Morality shares a feature with the Highway Code and this feature may be lost in Factualist terms. This suspicion we may support by extending the analogy between traffic beliefs and moral beliefs. Suppose now traffic beliefs *are* indeed moral beliefs. If this is a coherent idea we will have to say something about the respects in which morality and the Highway Code are similar.

For a hyper-traditional society (as sketched, for example, in Chapter 2.2), we declare the norms of moral conduct to be *constituted* by the Highway Code. People believe it to be evil (i.e. not a matter of prudence, etiquette or beauty) to drive on the right side, ignore red lights etc. Such a society seems not only to be a logical but arguably an empirical possibility. Traditional human communities, I argued, often treat moral judgment as a matter of application of given norms, they have little conceptual room for a more fundamental inquiry. It seems coherent to think of a closed and uniform society of the radical type of a Highway Code society, devoid of fundamental moral reflection (not to speak of meta-moral theories), yet undeniably incorporating (a) norms (as patterns of behaviour) and (b) moral beliefs (from the point of view of the society's individual members). What conclusions can we draw from this thought experiment?

Some may want to conclude that the factualistic account of morality is somehow defective, for without fundamental moral reflection there cannot be beliefs about morality as promoting the social good. Though there may exist many facts about a traditional society's social rationality, these facts cannot be the objects of moral belief from within that society. Moral beliefs from within the Highway Code society therefore must be characterized other than as beliefs about "critical explanations of social goods" (to use Railton's phrase). To this, the Factualist will reply that moral beliefs within a traditional society are outside the scope of moral beliefs as he likes to treat them. Moral facts as objects of contingently motivating moral beliefs appear only from a *standpoint of reflection.* Moral beliefs within a traditional society may motivate intrinsically, from a standpoint of reflection however that intrinsic motivation is to be exposed as a myth.

> Weak minds and moralists have ... surrounded justice with certain myths – that justice is its own reward, that once one sees what is just one will automatically have a reason to do it, and so on. But then ... weak minds and moralists have likewise surrounded wealth and power with myths – that the wealthy are not truly happy, that the powerful inevitably ride for a fall... (169)

On this reading, traditional moral beliefs can only *improperly* be called moral beliefs. This imposes a strain on the factualistic position. Factualism remains only intact as an alteration of received conceptions of moral belief. The Factualist, however, might not be too shaken by this result. Philosophy has often been revisionary. The Factualist may seek to provide (wholly in accordance with my explanatory credo of Chapter 2.3) what has been called a "reforming definition" of morality.[1] The main challenge, I suggest, must concentrate on how great a loss is constituted by the Factualist's revision. In which respects should the Factualist uphold but cannot that moral beliefs and traffic beliefs are similar?

I have suggested that people within the Highway Code society are norm-governed in a way intrinsically connected to their conception of the right and good. We may surely ask (1) whether people adopting a reflective attitude towards the institution of morality are themselves norm-governed, and if so, (2) to which state of mind norms that reflective people are governed by relate to. And finally we may ask,

[1] See Brandt, 1979, 10.

(3) how reflective norm-governed behaviour may relate to the Factualist's conception of morality.

In advance, let me briefly extract and clarify the notion of norm-governed behaviour which enabled my social scientist to speak confidently of rules of traffic in an otherwise inaccessible society. The notion of a norm presupposes no more than that it is possible to group patterns of behaviour under the heading of generalized claims. "Dogs bark whenever their territory is invaded" or "Unfamiliar and unagressive human beings do not sustain eye-contact for more than x seconds" may thus count as instances of norms. Among human animals, norms can be linguistically encoded and taught as precepts. The usual linguistic form is here a generalized imperative: "Do not stare at strangers!" or "Look right!" (as is written for the tourists' benefit on numerous pedestrian crossings in Great Britain). People within the Highway Code society are governed by fairly precisely formulizable imperatives they seem to accept, and so may be a society of reflective, epistemologically scrupulous Factualists. Where they differ is in the way these societies are prepared to describe norms they are governed by as moral norms. In both cases, however, the norms they are governed by, and the linguistic elements which typically express such norms, seem intrinsically motivational. Now Railton allows that there may be linguistic elements which directly express intrinsically motivating states of minds. These, he holds, must be instances of valuing *tout court*. Norm-governance in this sense is accounted for by the distinction between valuing and observing that something is of value. As a linguistic expression of intrinsically motivating states of mind Railton suggests "the thing to do".[1] A person X is non-contingently motivated to perform an action A if X believes A to be "the thing to do". This seems to me to fall short of what is needed to provide for normative governance even in a society of reflective Factualists. We need linguistic elements which express *patterns* of "things to do". Accepting the norm "look right" does imply more than looking right once. It requires,

[1] Railton, 1992, 966.

we might say, patterns of motivational states. At the next British crossing, other things being equal, "look right" cannot be substituted by "look left" – as the formulation "the thing to do" would suggest.[1] Thus it may only be open to Railton to either deny that there is normative governance at all or allow for norms as patterns of intrinsically motivating states. Railton's Factualism seems to rely on an account of norms, thus clearly opting for the latter alternative: "individuals can significantly influence the likelihood of norm-following behavior on the part of others by themselves following norms." (198) Most prominently this is revealed with, Railton says, "prohibitions of aggression and theft, and of the violation of promises" (ibid.). How precisely operate norms in a factualistic account of social rationality? Consider the case of promises. Mackie, in a Humean vain, gives us a vivid illustration how, without the institution of promising, we may fail to attain the social good.

> 'Your corn is ripe today; mine will be so tomorrow. It is profitable for us both that I should labour with you today, and that you should aid me tomorrow. I have no kindness for you, and know you have as little for me. I will not, therefore, take any pains upon your account; and should I labour with you upon my own account, in expectation of a return, I know I should be disappointed, and that I should in vain depend upon your gratitude. Here, then, I leave you to labour alone: you treat me in the same manner. The seasons chance; and both of us lose our harvests for want of mutual confidence and security.'[2]

With the device of promising this impasse may be overcome. How? To a linguistic expression like "If you help me now I promise I will help you later" a motivational state becomes attached. Sincerely uttering "I promise" then commits me to a norm "whenever one promises one is motivated to deliver (all other things being equal)", or in the corresponding imperatival form "whenever you promise, deliver!". It is not that I see it this one time as "the thing to do" to help you. Rather I accept a norm telling me to be motivated in like circumstances. With the sincere utterance of "I promise" I

[1] To be sure, there are cases where I accept a norm "look right" and do not look right at a relevant occasion. In extension of what has been said in Chapter 1 on sentiments, all that is needed to uphold the intrinsically motivating character of my state of mind is an explanation why I violated a norm I accepted. I might have been dreaming, in a rush etc.

[2] Mackie, 1977, 110-11.

express a state of mind of commitment to patterns of "things to do". Only with this conceptually motivating power is the institution of promising liable to induce cooperation which may then tend to be rational from a social point of view too.

We can now see clearer that Railton's conception of moral facts, while rejecting morality as a system of norms, relies (at least in part) on a notion of norms to create the possibility of social arrangements that moral facts are said to be constituted by. As I presented it, a system of intrinsically motivating states is presupposed by the existence of social rationality. Why not be Internalist, call these states moral beliefs and a system of such states a morality? Since a system of intrinsically motivating states seems to be conceptually prior to Railton's Factualism we should at least explore the possibility of locating morality here.

3.5 Internalism

There is one central problem for Internalism: Doing what one thinks one ought not (or: Not doing what one thinks one ought). We all are familiar with such cases, a few have already been mentioned – the vandal, the hypocrite, perhaps Hume's Sensible Knave; these are persons we expect to acknowledge that they are not doing the right thing (in a moral sense). It seems to be one of the virtues of Factualism that it can easily account for such cases. Moreover, if there are external facts moral judgments answer to, as the Factualist claims there are, we should expect to find persons not doing what they believe they ought to be doing. The notion of moral facts implies that there be a gap between what is morally the case and what people do. For the Externalist, the facts of moral desirability are thus divorced from an agent's motivational features. Here, the Internalist is in trouble. If moral judgments depend in some special way on motivational features of the moral believer – her desires, preferences, inclinations and apprehensions – how (1) do we explain cases of amoralism, hypocrisy and moral weakness, i.e. cases where a person appears to do what she thinks she ought not to be doing, and (2) how do we preserve a distinctive moral desirability, that is, how do a person's *moral* tendencies differ from her other motivational features, generating non-moral value?[1] Richard Hare has developed a characteristically forthright response to the second question which leaves him, I believe, dangerously little choice with the first question. The shortfalls of Hare's conception of "assenting to a value-judgement" should enable us to see clearer what is required of a coherent Internalism.

3.5.1 Hare's Internalism

Before I go into some textual matters, a few comments on Hare's terminology may be called for, since his way of casting moral theory as inquiry into the "meaning of moral terms" or the "logic of moral concepts" seems at variance with some of my earlier claims. In Chapter 2, I argued that we need a broader philosophical explanation of the *mental states* of moral belief and judgment, and that analyses of moral language revealing actual usages must at best fall short, at worst be seriously misleading. (A hyper-traditional society's moral language, for example, may be perfectly descriptive.) Now, it is easy to think that Hare's moral language arguments fall foul of these methodological ideals. Hare sometimes talks with strange certainty of "misuse" and "abuse" of language (e.g. F&R, 37).[2] In fact, Hare's investigations are much subtler than this terminology suggests. At one point, he concedes explicitly that his inquiry will be "at one and the same time about language and about what happens" (F&R, 75). His more radical theses claim no more than that there is "a central use" of 'ought' which is prescriptive, universalizable and overriding though there may be other uses and not all people employ moral concepts in the way Hare thinks they should be

[1] Railton mentions similar objections in a footnote (170, n.8): "it is necessary to have a contentful way of characterizing criteria of moral assessment so that moral approval does not reduce to "is valued by the agent."" To conceive of distinctive moral value as that which the agent prizes above all else would have the peculiar effect of making amoralism a "virtual conceptual impossibility."

[2] In this chapter quotations from Hare will be revealed in the following way: Hare, 1952, *The Language of Morals* = LoM; Hare, 1963, *Freedom and Reason* = F&R; Hare, 1981, *Moral Thinking* = MT.

employed. I prefer to read Hare's defence of the centrality of his analyses as a broader explanation in my sense.[1] Hare writes typically:

> I am merely suggesting a terminology which, if applied to the study of moral language, will, I am satisfied, prove illuminating. (LoM, 169)

or again:

> The substantive part of the prescriptivist thesis is *that there are* prescriptive uses of these words, and that these uses are important and central to the words' meaning. That they are important and central is shown by the fact that the problems which notoriously arise concerning moral language would not arise unless there were these uses. (F&R, 84; see also MT, e.g. 27, 80ff.)

Thus Hare's discussion of the prescriptivity of moral language should translate without strains into the terminology of Internalism and Externalism about moral motivation.

Hare's conception of value-judgments as entailing imperatives makes its first appearance in his 1952 classic *The Language of Morals*. Since Hare's basic convictions have changed little over the years it will be useful to recapitulate briefly. In part I.2.2 he gives an analysis of singular imperatives which in part III.11.2 is applied to moral 'ought' judgments. The following conditions underwrite the acceptance of a singular imperative addressed in the second-person to ourselves (1):

> ... we are said to be sincere in our assent if and only if we do or resolve to do what the speaker has told us to do; if we do not do it but only resolve to do it later, then if, when the occasion arises for doing it, we do not do it, we are said to have changed our mind; (LoM, 20)

From this, the application to value-judgment reads (2):

> ... the test, whether someone is using the judgement 'I ought to do X' as a value-judgement or not is, 'Does he or does he not recognize that if he assents to the judgement, he must also assent to the command "Let me do X"?' (LoM, 168-9)

The transition from an analysis of singular imperatives in the second-person to value-judgment is less straightforward than it looks. As Hare presents it, both (1) and (2)

[1] Under my reading, Blackburn's criticism of Hare seems slightly unfair (Blackburn, 1993a, 202). "Meaning is properly talked of only where we have *convention.* But there is no convention governing the selection of standards in ethics: someone who approves of the wrong things is not unconventional in the way of someone who uses a word wrong ... This is why, in my view, Professor Hare's battle to make universalizability, and hence perhaps utilitarianism, emerge from the meaning of ethical terms is quixotic."

seems to suggest one *overriding* motivation "Let me do X" – a resolution to act in only one way, or to accept X as "the thing to do", as Railton might put it. For (1), however, a closer examination reveals the motivation immediately as being merely *prima facie*. Often we sincerely accept an imperative and still fail in our expressed resolve, sometimes we plunge into a conflict of motivational states, having accepted two or more incompatible imperatives. I may, for example, have accepted that I should take the train at 10.02 but leave it simply too late, or leave it too late because I also sincerely accept that I should complete this paragraph first. If this is true for the sincerity of professed intentions, (that is, in Hare's words, the assent to imperatives in the second-person) it might by way of Hare's analogy also be true for value-judgment. In *The Language of Morals*, Hare is unhappy with this implication but postpones the problem.

> ... our criteria, in ordinary speech, for saying 'He thinks he ought' are exceedingly elastic. If a person does not do something, but the omission is accompanied by feelings of guilt, &c., we normally say that he has not done what he thinks he ought. It is therefore necessary to qualify the criterion given above for 'sincerely assenting to a command', and to admit that there are degrees of sincere assent, not all of which involve actually obeying the command. But the detailed analysis requires much more space than I can give it here, and must wait for another occasion. (LoM, 169-70)

The occasion came with *Freedom and Reason.* There Hare defends the more hard-line view that sincere assent to moral principles cannot be overridden.

> ... suppose that I have in my room in College a scarlet sofa, and that my wife gives me for my birthday a magenta cushion to go on it; and suppose that I am, so far as aesthetics go, vehemently of the opinion that one ought not to juxtapose scarlet and magenta. I may nevertheless think that I ought to keep the cushion on the sofa; because I may think, so far as morals go, that one ought not to hurt the feelings of, or lie to, one's wife. (F&R, 168)

This introduces a distinction between sincerity with regard to singular second-person imperatives and imperatives as they appear in moral judgment. The former are *prima facie*, the latter *overriding*. On account of the overridingness of morals, it becomes impossible for a person to sincerely accept a moral judgment and not be motivated in a relevant sense, magenta and scarlet be as they may. The disanalogy, so it seems at first, may be a virtue for it provides a criterion for the distinctive moral desirability the Externalist missed in internalist accounts. But how does Hare's conception of moral judgment as entailing overriding imperatives cope with cases of doing what one thinks one ought not and not doing what one thinks one ought to be doing? Hare rightly notes (F&R, 67-8) that we understand the problematic cases of amoralism, moral weakness etc. as *deviations*. If a person does not do what she says she ought to be doing she has something to answer for. The Factualist will find it hard to describe why the relevant cases strike us in this way. Under a conception of moral facts as an explanatory part of the world, why can it be a problem when people do not do what they think they ought to be doing? How, then, can we understand the problematic cases as deviations? In *Freedom and Reason*, Hare suggests two main manoeuvres.

(1) Persons doing what they think they ought not (or: Not doing what they think they ought), do not make moral judgments in the *proper* sense. In Hare's picture, moral judgment's overriding imperatives may fail to motivate because of a deviant use of moral terms, either prescriptively or with regard to their universalizability. A person may be sincere in her moral judgment but may fail to realise that, on the universalizability requirement, it applies to her. She may not be really committed "to wanting anyone else placed in exactly or relevantly similar circumstances to do likewise" (F&R, 5.4, 71).[1]Alternatively, a person may be sincere in her moral judgment but use it non-prescriptively (F&R, 5.5). Such a person, for example, may merely acknowledge that people call a particular action X right: "X is required in order to conform to a standard which people of a particular community accept". This, Hare tells us, is a use of moral terms in inverted commas. The improper or deviant use of moral language may then account for some cases of apparently sincere but non-motivating moral judgment.

(2) What about persons who sincerely make moral judgments, fail to act upon them, yet use moral terms properly and not in inverted commas? Hare agrees with common sense that such cases exist. The psychological phenomenon of an inner struggle, often involved in these cases, leads Hare to speak summarily of weakness of

[1] This move is peculiar to Hare's theory of universalizability and may not be open to other overriding Internalists. For a detailed discussion of the universalizability constraint, see Chapter 4.2 below.

the will, as when Paul writes to the Romans: "... though the will to do the good is there, the deed is not. The good which I want to do, I fail to do; but what I do is the wrong which is against my will;" (Paul, *Romans* vii; F&R, 78). How can the possibility of moral weakness be explained on an account of moral judgment as overriding? Hare writes:

> Nobody in his senses would maintain that a person who assents to an imperative must (analytically) act on it even if he is unable to do so. (F&R, 79)

The notions of physical and psychological impossibility are then elucidated in the following way:

> ... 'physical' impossibility (and also such allied cases as impossibility due to lack of knowledge or skill) causes an [singular] imperative to be withdrawn altogether, as inconsistent with the admission of impossibility; ... in a similar case an 'ought' does not have to be withdrawn but only down-graded. It no longer carries prescriptive force in the particular case, though it may do so with regard to actions in similar circumstances (similar, except that the action is possible). (F&R, 80)

For moral weakness as psychological impossibility, the prescriptive force of the sincere judgment survives into the psychology of the inner struggle.

> The form of the prescription is preserved, ... (and this shows how reluctant we are to suppress it) in the curious metaphor of the divided personality which, ever since this subject was first discussed, has seemed so natural. One part of the personality is made to issue commands to the other, and to be angry or grieved when they are disobeyed; but the other part is said to be unable to obey, or to be so depraved as not to want to, and to be stronger than the part which commands. (F&R, 81)

Let us now turn to particular cases which may constitute counter-examples to Overriding Internalism and construct the available responses from Hare's material.

3.5.2 Counter-Examples to Internalism

(a) The amoral person

An amoral person may refuse to make moral judgments altogether. A vandal, say, may pass a tree and simply break it without premeditation, then announcing that everybody may or may not break trees as they please, this not being a moral matter. (cf. F&R, 101) Still, the vandal may recognize that people call him a vandal. He may be able to say sincerely but in inverted commas that it is wrong to break trees. On the other hand, one might not easily ascribe a set of values to such a person. His preferences might be so incoherent that only with difficulty could a set of norms be formulated that he actually seems governed by. If the vandal cannot be seen as holding a set of moral beliefs, he will not constitute a counter-example to Internalism.

(b) The immoral person

A different kind of vandal, perhaps, might subscribe to a coherent set of purposes. He may be attracted by whatever people call wrong and evil (accepting their judgments in inverted commas). For his part, he might be out to destroy what is dear and useful to those people. A vandal with such a coherent set of destructive values is unlikely, yet possible. To this case, Hare's reply must be that

> A man's moral principles ... are those which, in the end, he accepts to guide his life by ... (F&R, 169)

On this reading, the vandal's moral beliefs are still overriding. Immoralism, as a coherent set of purposes, turns into a morality in its own right. This move makes coherent immoralism into a conceptual impossibility, thus disarming the immoralist as a counter-example to Overriding Internalism.

(c) The hypocrite

A hypocrite might be an employer who publicly disowns racism, yet does not employ black or coloured people. If she is sincere in her public judgments she must be deceived about what she really believes. If she is insincere she becomes like the Sensible Knave of Hume's *Second Enquiry*. The Sensible Knave may recognise that what he does is wrong (in inverted commas). Still, this does not prevent him from pursuing his own agenda. Again, Hare would have to employ the interpretative strategy, turning the sincere hypocrite into a moralist in her own terms (F&R, 83).

(d) The morally weak

A person might not do what she believes she ought to do because she is in some sense morally weak: she may be grief-stricken, depressed, hypnotized, under the influence of alcohol, her attention may have slipped, or perhaps she simply succumbed to temptation. For this person, judgments which under normal circumstances are motivational fail to be so. In *The Language of Morals* already, Hare had remarked that such motivational failures are typically "accompanied by feelings of guilt" (LoM, 169). In *Freedom and Reason*, he writes again:

> The residual feelings of guilt have supplied the place of real prescriptiveness. (F&R, 83)

It may therefore seem open to Hare to specify a set of conditions for a conceptual connection between a moral judgment and action along the following lines.

> X sincerely believes he ought to do A, if and only if (i) X will do A under normal circumstances [not being grief-stricken/depressed/ hypnotized/ ...], and (ii) failure to do A will result in self-censuring feelings [of compunction/guilt/remorse/ ...].

This resembles a manoeuvre I performed in Chapter 1 in order to secure a notion of practicality for certain states of minds. To be thirsty, for example, motivates intrinsically the quenching of thirst. This does not require that each time you are thirsty you drink. You may have other things to do, you may be in a rush, your life might even be under threat. In Chapter 1, I concluded that it is defining for intrinsically motivating states that we would need such explanations in the absence of behavioural manifestations while, for contingent motivations, there is no need to construe counter-factuals of this kind. Do the conditions sketched above provide a possible set of counter-factual explanations for non-motivational moral judgment?

To make progress on this tricky question, consider a further apparent counter-example to Hare's Overriding Internalism.

(e) The person weighing goods

In the previous example, the absence of grief, depression, hypnosis etc. constitute somehow conditions of normality. The plausibility of Hare's response to the morally weak person rests on the idea that *under normal circumstances* a sincere moral judgment would be fully motivational. The person to be considered now is a person who, under otherwise normal circumstances, weighs moral against non-moral goods. Is such a person a coherent possibility? It seems so. Philippa Foot gives an example where considerations of etiquette operate against moral judgment:

> There is ... a distinct resistance to the idea that a host or hostess might refuse to serve any more drinks when the guests have had as much as is good for them given that they must drive home. In spite of the fact that they might kill or injure someone, which is surely a moral consideration, the host is not expected to close the bar and refuse to serve more alcohol as soon as this point has been reached. A strong rule of etiquette forbids such a course of action, and it is the rule of etiquette that takes precedence ...[1]

Another phenomenon, undoubtedly common, is partial hypocrisy. Under protection from public scrutiny, most of us bend or break moral norms we sincerely accept. Allan Gibbard cites a psychological study of a Methodist community publicly opposed to tobacco, liquor and card-playing. Yet, "A number of them secretly smoked, drank, or played cards ... each believing himself the only one who would think of doing so."[2] It would probably strain our charitable inclinations if we had to read these Methodists, occasional lapsing in their moral aspirations, as pursuing a consistent set of purposes; they are not Sensible Knaves (as in (c)). On the other hand, I also find it difficult to see that Methodist smokers, drinkers and card-players violate conditions of normality (as in (d)). What about Philippa Foot's hosts? Again only two options seem to be available to Hare. (d) obviously misses the point. It is not that the hosts find themselves in a drunken stupor, unable to resist any further temptation. In the clear light of the day, the hosts may have accepted the demands of etiquette and stocked the bar. They may also not be hypocritical about that. Thus it seems that

[1] "Are Moral Considerations Overriding?", in Foot, 1978, 184.

[2] Gibbard, 1990, 76, n.20 (Schank, 1932).

Hare's natural response must be as in (b): our moral principles are those which, in the end, we accept to guide our lives. The hosts seem indeed to accept as the overriding norm that the consumption of alcohol, though potentially dangerous, falls into each individual's own responsibility. Thus the original moral norm against drink-driving is modified by a proviso declaring "Hosts are not to police liquor".[1] Similarly, the smoking Methodist must be said to believe really that the moral norm forbidding the use of tobacco should be read as permitting occasional lapses as long as they remain secret. With extensive use of this interpretative strategy, Hare might be able to avoid the challenge presented by some of the persons doing what they think they ought not to be doing. The immoral person, the hypocrite and the person weighing goods cannot claim first-person authority about what moral beliefs they maintain; their self-conceptions turn out to be incoherent. This, however, is achieved at a cost. In order to account for people apparently doing what they think they ought not, Hare has to collapse, in a terminology introduced briefly in Chapter 1.5, narrow into broad morality. He cannot preserve a distinctive narrow moral desirability. In the end, it is a person's deeds that reveal her morality. There may be some Marxist truth in this, though one would not expect to hear it from this corner.

A fundamentally different response for the Internalist is to drop the requirement of morality's overridingness. This would naturally allow for morally weak persons but also for persons weighing goods thus tolerating amoralism of some kind. Still, will *prima facie* Internalism be able to preserve a distinctive (narrow) sense of the moral? If *prima facie* Internalism succeeds here it will have advantages over Hare's conception on both counts.

Prima facie Internalism can be represented through a simple addition to the set of conditions for an intrinsic connection between moral judgment and action given

[1] In her article, Foot admits as much: "Moral rules are not taught as rigid rules that it is sometimes right to ignore; rather we teach that it is sometimes *morally permissible* to tell lies (social lies), break promises (as e.g. when ill on the day of an appointment) and refuse help (where the cost of giving it would be, as we say, disproportionate). So we tend, in our teaching, to accommodate the exceptions *within* morality, and with this flexibility it is not surprising that morality can seem 'unconditional' and 'absolute'." (Foot, 1978, 186-7)

above. *The absence of countervailing motivations* should supplement Hare's conditions of normality.

> X sincerely believes he ought to do A, if and only if (i) X will do A under normal circumstances [not being grief-stricken/depressed/hypnotized/ ...], (ii) in the absence of stronger countervailing motivations [of prudence/etiquette/aesthetics/ ...], and (iii) failure to do A will result in self-censuring feelings [of compunction/guilt/remorse/ ...].

For non-moral motivation, this move is natural and obvious. Your not quenching your thirst because you are in a rush is interpreted as a case where more than one reason plays upon one resultant motivation to act. The Overriding Internalist may object that this will be too weak a set of conditions for a distinctive moral desirability. The amended set of conditions, he might insinuate, says no more than that it is sufficient for moral sincerity to act morally if nothing speaks against it – a ridiculously insubstantial condition. This, however, underestimates the force of the third condition to which Hare himself had drawn our attention.[1] If you don't quench your thirst because you are in a rush, you may regret it but you should not feel a stronger sense of blame. Moral distinctiveness is achieved by introducing specifically moral sentiments. In the first-person, typical sentiments of moral failure seem to be compunction, guilt, shame and remorse.

The violation of norms believed to be distinctly moral results in specific feelings of sanction in the first-person. This account, capturing most intuitions of common sense, carries welcome implications within my overall programme. In search for a

[1] Gardiner (1954, 44) also stresses the importance of censuring feelings for moral sincerity. "... if certain other factors are absent, e.g. signs of compunction or remorse, or resolutions to mend his ways, we may ... reach the conclusion that the man was insincere." C.C.W. Taylor (1980, 516) gives the following conditions for '*A* ranks the doing of *x* by him on this occasion higher than he ranks the doing of *y* by him on this occasion': "*A* does *x* spontaneously and unhesitatingly in preference to *y*; *A* feels pleased that he has done *x* in preference to *y*; *A* feels remorse that he has not done *x* in preference to *y*; *A* regards this as a typical case of choice between doing *x* and doing *y*, and admires people who in such cases do *x* in preference to *y*;"

coherent Internalism we arrived at a version of Sentimentalism. An account of moral beliefs as dispositional sentiments of a certain kind would rescue Internalism from making amoralism a conceptual impossibility while preserving a place for distinctive moral value.

Before concluding this chapter, I shall mention two further cases which may be thought to constitute counter-examples to Internalism.

(f) The weak-willed person

A weak-willed person in the technical Greek sense of *akrasia* is a person who does not do what she wants to do. She fails to enact her expressed resolve. She is motivated but her motivation does not translate into action. If, like for Hare, a person's (X) sincere moral judgment that X ought to do A becomes somehow assimilated to X wanting to do A, moral weakness may seem like weakness of the will. The morally weak and the weak-willed person, both sincerely believe that they ought to do an action, yet fail to do so. The "ought" of moral weakness and of weakness of the will, however, are very different; and where there is weakness of the will in a moral case, the thesis of Internalism remains unaffected. The *akratic* paradox may afflict all kinds of intentional explanation of action. It only arises because we grant that an agent is motivated. A motivation, typically expressed by an agent's reasons for acting, fails to issue in the action it normally explains. Moral weakness, as discussed in (d) is a failure to be motivated. Weakness of the will is a failure of motivation. Thus weakness of the will (in moral cases) presupposes that an agent is *not* morally weak. In the apparently paradoxical case where a morally strong agent is *akratic*, it is sufficient for Internalism that the agent was motivated in the relevant sense.[1]

[1] As Gosling puts it (1990, 154): "... the shape of one's disquiet about *akrasia* retains the same form whether or not the prescriptivist thesis is right."

(g) The moral philosopher

What about the moral philosopher? Some, like Railton, believe in external moral facts, some, like Sidgwick (1907, 489) and other classic proponents of utilitarianism, even believe that our best chance in acting rightly does not lie in following what we believe to be right. Are they insincere? We would hesitate to say so but we must maintain that they are incoherent. Regrettably, the externalist moral philosopher has to be treated like the amoralist and the moral cynic.

3.6 Résumé

It is now time to turn back and survey the six options I mapped out in the beginning of this chapter. We found, I believe, conclusive arguments against three of the positions. Non-Cognitive Externalism (2)(ii), Cognitive Externalism (2)(i) as well as Non-Cognitive Overriding Internalism (1)(ii)(a), all fall out of the picture. Of the remaining positions, two are internalistic *prima facie* theories (Cognitive (1)(i)(b) and Non-Cognitive (1)(ii)(b)). Of the surviving overriding position (1)(i)(a) we may ask, do the objections against Hare's theory in *Freedom and Reason* apply equally to the corresponding cognitive theory? To me, this seems to be the case though we know little about how Cognitive Internalists actually would react to the problematic cases I discussed. Cognitive Internalist theories are typically embedded into a wider theory of moral reasoning and its cognitive credentials. In his more recent book *Moral Thinking*, Hare for the non-cognitivists has undertaken a similar move. He departs from the claim that all moral judgments are conceptually overriding. On what he calls the intuitive level, he adopts a sentimentalistic account not far away from the position urged above:

> If I have been well brought up, I shall, when I break the promise, experience this feeling of compunction (no doubt 'remorse' would be too strong a word in this case), which could certainly be described, *in a sense*, as 'thinking that I ought not be doing what I am doing, namely breaking a promise'. (MT, 30)

Overridingness is upheld but on the so-called critical level. Once we adopt a certain method of moral thinking and justification, we come to see that there cannot be a conflict of motivations. Hare's constraints for consistency of motivations are derived from his controversial notion of the universalizability of moral judgments which allows him (as he sees it) to drive a wedge between moral principles and moral sentiments (MT, 38). The Cognitive Overriding Internalist as Hare's opponent must rely on the sentimental response itself to provide a notion of adequacy and justification. This debate will surface in the next chapter but leaves the present discussion behind.

Summarily, I conclude that some version of Internalism must be true. Most at ease with the problematic cases of amoralism, moral weakness etc. are the *prima facie* versions giving an important role to certain moral sentiments. Overriding accounts will have to come up with a more comprehensive account of moral reasoning and justification. Russell wrote:

> We have, in fact, two kinds of morality side by side: one which we preach but do not practise, and another which we practise but seldom preach.[1]

On one perhaps uncharitable reading Externalists can be seen as adding a third kind of morality. To the morality we preach and the morality we practise comes morality as a matter of fact. On my account, there is the narrow morality we preach, and the broad morality we live, but Internalism as I defined it provides a logical link between the two spheres. Thus I speak in favour of one morality.

[1] Russell, 1928, 103.

4. SENTIMENTAL CAUSES

In Chapter 1, I claimed that the analysis of sentiments in terms of causes, symptoms and actions introduced in effect an explanatory chain. My fear, say, of dark alley-ways entails (1) that I avoid dark alley-ways *because* I am afraid, and (2) that I am afraid *because* (I believe) dark alley-ways are dangerous. The first *because* clause explains an action with reference to a state of mind, the second *because* clause explains (and perhaps justifies) that state of mind. The particular character of the first explanation, in my view, implies an internalistic reading of the underlying state of mind. Fear of dark alley-ways is intrinsically motivational.

In our prolonged discussion of Internalism about moral belief in Chapter 3, we ended up with a modified *prima facie* interpretation of the motivating features of moral belief. How should we construct the explanatory chain for moral belief? Consider the case where I apologise because I believe I offended her. The underlying state of mind, I said, is intrinsically motivational in that, in the absence of behavioural manifestations, a counter-factual explanation is required if we still are to ascribe a moral belief in sincerity. If I sincerely believe I had offended her and do not apologise it must be that I was somehow out of my mind, so to speak, or had better countervailing reasons, and in any case that I experience sanctions for not apologising, like feelings of compunction, guilt, shame or remorse. Sentimentalism, as I presented it, claims that moral beliefs are dispositional sentiments of sanction: of compunction, guilt, shame or remorse in the first-person, of anger, resentment, blame or indignation in the second- or third-person. To judge that I ought to apologise is to feel, say, dispositionally guilty. Like the fear of dark alley-ways, the guilt (or other feelings of self-censure) entails the motivation to perform a certain action; fear of dark alley-ways will be explanatory for the avoidance of dark alley-ways, guilt will be explanatory for the action whose absence would license the sanction of guilt, in this case the apology. With moral sentiments, we find ourselves immediately drawn to the second *because* in the explanatory chain. Dispositional guilt is explanatory for a particular action because there is a sentimental cause believed to license the occurrence of guilt. For fear, the respective sentimental cause can be readily identified as a dangerous circumstance; where my avoidance of dark alley-ways is explanatorily entailed by my fear, my fear is explained by a potential danger. This is what fear means if we properly understand that concept.

For moral sentiments, the situation is rather more complicated. Why does giving offence call for an apology? In which way does not apologising license sentiments of sanction? Is a guilty response implied by the competent use of the concepts of offence and apology? Is an adequate response subject to constraints of consistency among responses towards the perceived causes of guilt? Is there a pragmatic rationale licensing the occurrence of guilt in particular cases? These are keyquestions I shall pursue in this chapter.

The mentalistic language in which these questions are cast should not confuse the reader. We deal with a familiar and central problem of moral theory: the epistemological status of moral claims. If the internalistic moral psychology I outlined is broadly right, the dispute of moral realism vs. epistemological non-cognitivism will be settled by the corrective resources available to modify our dispositional responses – since dispositional responses *are* what our moral claims express.

In the following, I identify three substantially different argumentative strains in defence of the claim that there is one uniquely qualified answer to any given moral question (in my view a useful first approximation to moral realism or epistemological cognitivism):

(4.1) It is a requirement of the *competent use of central moral concepts* that there are uniquely qualified moral responses. (This view is more commonly labelled as Secondary Quality Moral Realism, or perhaps less friendly as Intuitionism.)

(4.2) It is a requirement of *consistency* that there are uniquely qualified moral responses. (This variety of Kantianism is most prominently held by Richard Hare's Universal Prescriptivism.)

(4.3) It is a requirement of *instrumental rationality* that there are uniquely qualified moral (or at least normative) responses. (This is the central idea of Contractualism.)

4.1 Moral Sentiments As Appropriate Responses

Sentimentalism, as it has been stated, constitutes one particular dispositional (or response-) theory of value. Following a recent debate, we may first ask how we should understand the peculiar response-dependence of moral concepts, before turning to the further question whether the responses peculiarly tied to moral concepts should be seen as epistemologically cognitive or non-cognitive.

4.1.1 Response-Dependence

If the causes or objects of psychological states like desires, attitudes, sentiments or beliefs[1] can be identified and assessed *in*dependently of these psychological states we may talk of those objects and their properties as being response-*in*dependent. On a traditional Lockean view, objects characterized by their primary qualities are response-independent in this way. Perhaps none of the properties we ascribe conform to primary qualities in the traditional sense. Kant and the later Wittgenstein are two prominent philosophers often credited with such a thesis of, we might say, global

[1] By casting moral responses and attitudes as dispositional sentiments, I turned abstract terms into recognisable psychological phenomena. This, in my view, is one of the attractions of Sentimentalism. In order to avoid excessive terminological rewriting, I now often attend to the more established but underdefined uses of responses and attitudes. On attitude, compare e.g. Honderich's more phenomenological characterization as "*an evaluative thought of something, feelingful and bound up with desire* ... where feeling is somehow akin to sensation but unlocalized... [An attitude] involves less of what can be called excitement or bodily commotion than typical emotions" (1988, 14) with Blackburn (1984), who defines attitudes initially negatively against "judgements, beliefs, assertions, or propositions - which have genuine truth-conditions" (167), though Blackburn is aware that this contrast "may look very different if we think of beliefs in pragmatic or instrumental terms rather in terms of correspondence with facts" (147).

response-dependence.[1] Be that as it may. Even within global response-dependence, one might still be able to draw a contrast between properties which are more or less response-dependent in a local sense. The question is then whether an area of discourse apparently involving the ascription of properties of a certain kind can be interpreted response-independently in a way apt to sustain truth, knowledge, objectivity or perhaps realism.[2] Local areas for which the question of response-dependence has been asked include religious discourse, causes, conditionals, generalizations, other minds and even science. Where should we locate moral properties and moral discourse?

In a well-known passage in Plato's *Euthyphro* (10a-11b), Socrates and Euthyphro discuss whether something is pious (or holy [*hosios*]) because the gods love it, or whether the gods love it because it is pious.[3] Socrates insists that the gods' love of pious acts is in fact explained by the acts' being pious, and not *vice versa.*

Similarly, Aristotle has suggested we should be able to give a somehow independent account of what it is to seem good:

> We desire the object because it seems good to us, rather than the object's seeming good to us because we desire it. (*Metaphysics*, 1072a29)

In contrast, we have Hume's claim that

> To have the sense of virtue, is nothing but to *feel* a satisfaction of a particular kind from the contemplation of a character. The very *feeling* constitutes our praise or admiration. (T. 471)

Hume, however, immediately modifies:

[1] Putnam (1981, 61-2) writes: "Locke's own treatment of secondary qualities ... was to say that (as properties of the physical object) we can only conceive of them as Powers, as properties – *nature unspecified* – which enable the object to affect *us* in a certain way... I suggest that ... the way to read Kant is as saying that what Locke said about secondary qualities is true of *all* qualities – the simple ones, the primary ones, the secondary ones alike (indeed, there is little point of distinguishing them)." McDowell voices frequently the Post-Wittgensteinian view that there is no Archimedean point "from which a comparison could be set up between particular representations of the world and the world itself" (e.g. 1983, 13).

[2] I am aware that some of the participants in the debate about realism and response dependence, notably Crispin Wright, would disapprove of this way of putting it. For Wright, a discourse that is apt to sustain a notion of minimal truth need not be interpreted realistically (Wright, 1987; 1992). My formulation employs a pretheoretical notion of substantial truth.

[3] Johnston (1989, 171), Pettit (1991, 614), Wright (1992, ch. 3, Appendix) all refer to this passage.

> We do not infer a character to be virtuous, because it pleases: But in feeling that it pleases after such a particular manner, we in effect feel that it is virtuous. (ibid.)

Whatever this view precisely amounts to, Hume thought it to be illuminating to draw a now infamous analogy to Locke's account of secondary qualities. In a letter to Francis Hutcheson he wrote on 16 March 1740:

> I must consult you in a point of prudence. I have concluded a reasoning with these two sentences: *When you pronounce any action or character to be vicious, you mean nothing but that from the particular constitution of your nature you have a feeling or sentiment of blame from the contemplation of it. Vice and virtue, therefore, may be compared to sounds, colours, heat and cold, which, according to modern philosophy, are not qualities in objects but perceptions in the mind: And this discovery in morals, like that other in physics, is to be regarded as a mighty advancement of the speculative sciences; though, like that too, it has little or no influence on practice.* Is not this laid a little too strong. I desire your opinion of it, though I cannot entirely promise to conform myself to it.[1]

Common to all these views is that they appear to constitute different interpretations of a biconditional of the following form:

$$x \text{ is } P \Leftrightarrow x \text{ is such as to produce a } P \text{ response in subjects } S.$$

In Mark Johnston's catchy (but as we shall see misleading) phrase, if for an area of discourse explanations of properties go from left to right the properties are *discovered*, if the explanations go from right to left the properties are *projected*.[2] Socrates and Aristotle take the left hand of the biconditional to explain the right hand side. Socrates holds that the gods love pious acts because they are pious – Aristotle that our finding an object good explains our desiring it. Hume, on the first quote, seems to hold that the explanatory relation goes from right to left; a character is virtuous because it arouses a certain sentiment of approval in us. In the next but one sentence, I quoted then, this view immediately is retracted into something strangely tangled. "In feeling

[1] Greig (ed.), 1932, letter no.16. In the event, the "two sentences" appeared almost unmodified in Book III of the *Treatise* (T. 469).

[2] Johnston, 1993, 122.

that it pleases" do we discover or project that a character is virtuous? Are the observer's responses irreducible, and if they are, do they explain, and what and how? Can human responses play an irreducible role within an account of moral truth, objectivity, knowledge or reality?

With no quick answers at hand, no wonder commentators seized on Hume's secondary quality analogy where already much philosophical ingenuity had gone into putting the biconditional to work by introducing conditions of normality. On what is still the standard account

x is red ⇔ x looks red to normal observers under normal conditions.

On one natural interpretation, an object's being red is both discovered and projected. x is red because x looks red to normal observers under normal conditions, and x looks red to normal observers under normal conditions because x is red. If we are to believe some commentators, the same should be said for value. The biconditional is said to explain equally in both directions. In Wiggins' words, property and response are "made for one another" (TIMoL, 107).[1] The property explains the response, the response explains the property. The 'because' holds "both ways round" (TIMoL, 106).

> Circularity as such is no objection ... provided that the offending formulation is also *true*. But what use (I shall be asked) is such a circular formulation? My answer is that, by tracing out such a circle, the subjectivist hopes to elucidate the concept ... (SS?, 189)

In our context, we need not be too concerned about the success or failure of the biconditional as an analysis of secondary quality judgments. We want to know: Is the biconditional really elucidating for value concepts? Can we distinguish cases where a particular interpretation of the biconditional of property and response sustains something like truth, knowledge, objectivity or reality from cases where it does not?

[1] See also SS?, 198 and 199. In this chapter, page numbers preceded by 'TIMoL' refer to Wiggins, 1987 (1976) "Truth, Invention, and the Meaning of Life", numbers preceded by 'SS?' refer to Wiggins, 1987 "A Sensible Subjectivism?". McDowell (1987, 12) talks of "pairs of sentiments and features reciprocally related – siblings rather than parents and children".

Consider four predicates – 'red', 'funny', 'U' and 'good' – and examine what the respective biconditionals reveal about the status of the associated properties redness, the comic, U-ness and the good.

4.1.2 Red, Funny, U and Good

Take one conception of secondary qualities as a paradigm case where we treat "psychological states and their objects as equal and reciprocal partners" (TIMoL, 106). On this view, colour properties, say, cannot be reductively analysed in terms of a subject's responses. There are no "purely phenomenological" or "purely introspective" responses which would allow us to identify secondary quality psychological states independently from the properties under which the states subsume their objects (cf. TIMoL, 106 and SS?, 195). The dispositional account (expressed by the biconditional) allows that an object could have been red (in that it would occasion red responses under certain conditions) if standard observers or standard conditions had been different, or even if there were no standard observers and no standard conditions.[1] The property of redness as analysed by the biconditional is thus a genuine property: We look to objects in order to determine whether they are red; the predication of redness can be true or false; and redness explains why a thing looks red. Still the property of redness is response-dependent in that it would never manifest itself in a world devoid of standard (human) observers.

What about 'funny'? Again, one might say, we look both ways – to property and response: When we are in doubt whether a joke is really funny we may look to funny making features such as a particular ambiguity of meaning, or the timing of the delivery, on the other hand "there is no saying what exactly the funny is without reference to laughter or amusement or kindred reactions" (SS?, 195). Of a missed joke we might say we didn't see a feature – we didn't understand – but equally, after

[1] For a quick sketch of colour-properties along these lines, see e.g. Blackburn, 1993b, 376.

repeated failed attempts to improve our understanding and to raise a smile, we might resign to the fact that we simply do not share the speaker's sense of humour. It might not be possible to specify a set of conditions under which a joke has to "crack" for an observer with a standard sense of humour. In consequence, the explanatory force of comic properties is perhaps limited to observers with substantially shared responses. This suggests that the property/response biconditional for the comic may not sustain a suitable notion of truth, knowledge, objectivity or reality. To any given joke there might be more than one adequate response. Though we look to features of a joke, say, in order to determine whether it is funny, the predication of funny may not be substantially true, and comic features seem to explain why something is funny only in a restricted sense.

Next, consider the predicate 'U'. Short for "upper class" it was coined by Nancy Mitford in her notorious guide to social etiquette *Noblesse Oblige: An Enquiry into the Identifiable Characteristics of the English Aristocracy* (1956). Nowadays 'U' apparently stands for a property that manifests itself in circles around London's Sloane Square. Philip Pettit in a paper "Realism and Response-Dependence"[1] brought it to the attention of a wider philosophical audience. 'U' is a predicate whose extension is in constant motion; today laying cloth napkins is U, tomorrow it may be non-U; today you must have plastic flowers, tomorrow your own herb-garden; and so on and so on. The activity of Sloanies is characterized by their constant endeavour not to fall behind in the game. To be exposed as doing non-U things amounts to a kind of excommunication from Sloane Square. (Similar behavioural mechanisms can be found in other fashionable scenes: the clothes code of music-clubs or the acceptance of philosophical submissions by virtue of, say, "being a MIND-paper".)

Pettit has suggested that in order to determine whether something is U, a Sloany does not look to properties (as they may be represented on the left hand side of the biconditional) but to his responses. If I am a true Sloany and judge something to be U

[1] Pettit, 1991.

it *is* (analytically) U. Cloth napkins have no feature that makes them U apart from Sloanies finding them U. Thus a Sloany's judgment is entirely immune from "ignorance and error" (op. cit., 611). This seems wrong. To be sure, there exist predicates whose extensions are determined by a single authority. The pope's use of 'catholic' (in the Latin sense of "accepted basis of faith and order") may be such a case. With respect to the use of 'catholic', the pope appears to be analytically immune from ignorance and error. The pope, however, would deny that. At most he will admit that he is contingently infallible. Strenuously he will defend that 'catholic' stands for real properties. His judgments, he will declare, arise from the left hand of the biconditional, they are "epistemically servile" (op. cit., 612) judgments of discovery. For the pope, it might not be easy to identify the features to which he purports to respond. In Sloane Square, however, an answer is ready at hand. The property of a thing that is U is that it occasions similar responses in a significant number of Sloanies. About that property, any Sloany can be wrong.[1]

U-ness is distinct from redness in that it resists an easy reading of the *standard* conditions. A thing is red, we assumed, iff it looks red to standard observers under standard conditions, and *vice versa* a thing looks red to standard observers under standard conditions iff it is red. For 'U', the standard is set by nothing but actual collective responses which cannot be further specified. (In this, as we shall see, 'U' resembles Wiggins' 'good'.) 'U' also differs from 'funny' in interesting ways. While for 'funny', once we have exhausted the possibilities of an improved understanding of what makes a thing funny, a competent speaker might say, "I do not find that funny" and excuse herself with her divergent sense of humour, for 'U' there is no such escape route. Once we know the collective responses of the Sloanies, there is only one

[1] In this, I reject Pettit's test according to which we ask "whether something evokes the U/red-response in normal subjects because it is U/red or whether it is U/red because it evokes the U/red-response" (op. cit., 614). Failing the former formulation U-ness is said to be marked as a projected property. For redness, Pettit claims, affirmative answers can be given to both questions. Judgments of red are attuned to an "independent authority" (op. cit., 612). Redness is discovered, thus posing no threat to realism. I hold that in no biconditional representing a disposition an "order of determination" can be found.

answer to the question whether something is U. (As always, it might be vague, but it is not discretionary.)

Again, I conclude, we can offer a workable biconditional for a property without making progress on the question under what reading the biconditional may sustain a substantial notion of truth, knowledge, objectivity or reality. On the dispositional account of 'funny', it is ultimately up to you (your sense of humour) whether something is funny. On the dispositional account of 'U', it depends on the actual responses of a particular social group whether something is U. For 'red', the dispositional account seems to allow for objectivity because we can specify standards human observers have to meet if they are to count as normal. *It is not by virtue of a left to right or a right to left reading of the biconditional that the objective credentials of the discussed properties differ.*

Turn now to 'good' and Wiggins' response-dependent analysis of value-concepts in "A Sensible Subjectivism?". What are the grounds for Wiggins optimism that for 'good', the biconditional representing observers' dispositions may sustain moral objectivity of some kind?[1] I have already quoted Wiggins' opinion that a dispositional account of a property does not import vicious circularity if we can read the biconditional representing the disposition as true on both sides of the '⇔' functor. While a property's being explained by relevant responses introduces subjectivism, the responses' being explained by corresponding properties is said to resurrect something like truth, knowledge or objectivity, thereby turning wild, relativistic subjectivism into the sensible subjectivism the title of the paper wisely questioned. After what we have heard about 'red', 'funny' and 'U', this seems to be a misleading description of the situation. For we have been able to identify mutually explanatory property/response pairs for 'funny' and 'U' without feeling compelled to grant substantial objectivity to

[1] Put in terms of responses, Internalism does not import any new problems. The fact that judgments of value are motivational while judgments of colour may not be so, is covered by a non-representationalist reading of response. For value, the motivating thought is "finding that x deserves a response" (Wiggins, 1990, 83) where the response includes that we are party to an attitude. Similarly, judgments of the comic tend to make you laugh, judgments of U-ness may induce you to performing U-acts.

the properties these predicates stand for. One way to meet this objection is to question the characterizations of the predicates sketched above. With respect to 'funny', Wiggins can be seen as seeking to preempt the charge of relativism by rejecting a description of the comic as discretionary. "A feeble jest or infantile practical joke does not deserve to be grouped with the class of things that a true judge would find genuinely funny." (SS?, 193). Once we have explained what is funny by certain properties and refined our sense of humour we are expected to converge on what can count as funny and what not. This rigidifies the predication of 'funny' since (under the supposition that you possess a sense of humour) it cannot remain discretionary whether something is funny or not.

But does this commit ourselves to being objectivist or realist about the comic? It seems not. As we have seen with 'U', the extension of a predicate may not be discretionary yet 'U' falls well short of picking out a property that may license talk of realism, objectivity or substantial truth. The ascription of 'U' may be true in some minimal sense but always relative to responses around Sloane Square.[1] Thus we retain a difficulty equally affecting value predicates. How can we distinguish predicates with acceptable (universal) from predicates with unacceptable (relative) convergence in extension? Here, Wiggins appears simply to point to the phenomenon that there is a remarkable consensus about what is "*genuinely* [funny/appalling/shocking/consoling/reassuring/disgusting/pleasant/delightful/...]" (SS?, 199). Approvingly, Wiggins quotes from Hume's essay "Of the Standard of Taste" (SS?, 198):[2]

> There are certain terms in every language which import blame, and others praise; and all men who use the same tongue must agree in their application of them.

[1] There is a related suspicion that Wiggins' convergence-criterion for truth in "Truth As Predicated Of Moral Judgements" may sustain no more than a minimal notion of truth. For convergence and minimal truth, see Wright, 1992, esp. 88-9. Compare also Williams, 1985b, 143-5.

[2] For more on this revisionist reading of Hume, the my Appendix below.

The problem about what is good, right and beautiful is not that it is somehow discretionary in evaluation and/or relative in extension. The problem is simply a certain vagueness in application. This I find difficult to understand for general value concepts like good, right and beautiful for it is often the point of asking what is good, right or beautiful that *we don't know* in a more fundamental sense. Normative inquiry cannot be conceptually restricted to inquiry into the application of given norms or evaluative predicates. It is revealing that Wiggins feels himself drawn to less general evaluative predicates like 'funny', 'appalling', 'shocking', 'consoling', etc..

> Approving of some *x* had better not be a barely determinable state, approving *tout court.* It had better be approving of *x* as a good *g*, as a good *f*, or as a good *fg* ... (SS?, 212, longer note 19)

Is it plausible with less general concepts to expect "univocity" (SS?, 198)? Is there a substantial consensus hidden among the considerations which would make general or thin approval determinable?

4.1.3 Thick Concepts

The tendency to operate with less general evaluative predicates puts Wiggins interestingly close to philosophers defending the moral relevance of so-called thick concepts. McDowell, another prominent proponent of a sensibility or response-dependence theory of moral concepts, is here more explicit. For McDowell, the virtuous man is a man who conceives of his circumstances in terms of thick concepts. The virtuous agent's "conception of the situation, properly understood, suffices to show us the favourable light in which his action appeared to him" (1978, 16). In a terminology invented by Williams, evaluative concepts tied to specific circumstances are thick. The predication of thick concepts evaluates a situation as having a certain property while thin concepts are apt to evaluate without committing the competent speaker to any properties of the evaluated situation.

As paradigmatic thick concepts, Williams gives *treachery*, *promise*, *brutality*, and *courage.*[1] These concepts we may contrast with all-purpose thin evaluative terms like *good*, *right*, *beautiful* and *ought*. The idea is clear enough. As Philippa Foot put it, certain concepts (then not yet called thick) we cannot consistently employ without adopting an attitude. In her paper "Moral Beliefs"[2], Foot claims that we cannot (by rules of meaning) call a circumstance dangerous and deny that we have reason to avoid it. Similarly, for the moral term 'rude', we cannot identify an action as causing offence by lack of respect without condemning it. Foot's paper reacted to one familiar move analysing any given evaluative judgment into logically separable conjunctions of descriptive and evaluative elements. The evaluative element might be, for example, Hare's *ought* (understood purely prescriptively), while the descriptive element (once conceptually isolated) would be treated like all other *is* representations of the world.[3]

What are the arguments claiming to show that descriptive and evaluative functions of thick terms are inextricably linked? If we could always disentangle the evaluative and the descriptive, McDowell says[4], any thick concept could be mastered by an outsider to the community the concept is taken from. But an outsider cannot reliably classify new cases employing thick concepts. Therefore thick concepts cannot be disentangled. I do not find this argument convincing. It might be true that for us outsiders, the only way to pick out the property of an action the predicate 'U' stands for is to think of what responses an action would evoke in a typical inhabitant of

[1] Williams, 1985b, 129.

[2] Foot, 1958b.

[3] A two component analysis is also suggested in Gibbard, 1990, 112-7. Thick judgments (judgments including thick concepts) are tied both to circumstances they "naturally represent" (113) and to normative governance: "where normative judgment naturally represents something, a plainly non-normative judgment could naturally represent the same thing" (114). Gibbard's 1992 paper "Thick Concepts and Warrant for Feeling" enriches this account by a third element: judgment of warrant. In judging something to be "lewd", say, we have a description of the circumstances judged to be lewd ("open display of sexuality"), we have an attitude expressed ("L[ewd]-censoriousness") and we have the acceptance of a presupposition warranting the attitude ("the general importance of limiting sexual displays", 280). This, to my mind, is the best account of thick concepts we have.

[4] McDowell, 1981, 144.

Sloane Square. But, as I have argued, the inhabitant of Sloane Square is, in this respect, in no better situation than any outsider.

There is no need to rehearse in detail the arguments for and against an irreducible union of the evaluative and the descriptive. A simple reflection will ensure that even if thick concepts cannot be disentangled no suitable epistemic capital can be secured for the moral realist, cognitivist or objectivist. The fact that a rule of meaning is said to require the adoption of an attitude as part of the competent use of 'rude', 'cruel', 'dangerous', 'lewd', 'U', 'funny', 'consoling', etc., does not make thick concepts favour objectivity. Judgments containing thick concepts can be criticized. And the thicker the concepts are the more likely it is that they are relative to specific cultural set-ups. Often, they arise from openly relativistic circumstances. The man using 'U' may not speak my tongue, to take up Hume's phrase, still I might not want to evaluate the world in 'U'-terms. We need a specification (and authorization) of the language whose predicates pick out the moral. For Foot's term 'dangerous' nobody would disown the attitude to avoid what threatens harm (all other things being equal) because nobody likes to suffer harm. Still, not everybody wants to become a member of Sloane Square. Not everyone views the world through the same thick glasses.[1]

Ending this excursion, let us return to Wiggins. The closest Wiggins comes to offering an (authoritative) warrant that licenses particular thick concepts is this:

> What is wrong with cruelty is ... that it is not such as to call forth liking given our *actual* collectively scrutinized responses. (SS?, 210)

Again, this may also be true of 'U'. Something is "U" if it calls forth liking given our actual collectively scrutinized responses under the supposition that we are Sloanies. What guarantees that 'cruel' is not relative to a social group's responses as 'U' is relative to circles around Sloane Square, 'uppity' to responses of the old American South and 'lewd' to responses of the prudes?[2]

[1] Blackburn (1992a, 285) argues that "attitude is much more typically, and flexibly, carried by other aspects of utterance than lexical ones". Intonation, not lexical meaning, Blackburn claims, may do most of the evaluative work.

[2] 'Uppity' and 'lewd' are examples of Gibbard (1990 & 1992).

With cruelty (as with danger), we may be fairly confident that acts described as inflicting arbitrary harm draw a derogatory reaction universally. Here, meaning may coincide with warrant. There are, however, not many such basic human responses and not even all of them call for inclusion in a theory of the right and good. It is, for example, a fairly universal human response to meet experienced harm with desires for revenge, or to laugh at certain kinds of misfortune (*schadenfreude*). Wiggins is therefore right, I believe, to reject an account of our acceptable responses as those constituting the smallest common human response denominator under normal conditions. (Such an account seems to be the unintended consequence of some of Hume's talk about "an entire or a considerable uniformity of sentiment" if the aesthetic or moral "organs" of mankind are in a "sound state" (*Of the Standard of Taste*; for Wiggins' critique, see SS?, 190-2).) Once we leave behind the most basic human moral responses under conditions of normality there will be less than automatic convergence. If Wiggins intends to assimilate judgments of value to judgments of colour and not to judgments of "U", we need to know more about what is to replace the conditions of normality on which notions of truth, knowledge, objectivity or reality depend for secondary quality predicates and the properties they stand for. Even for a dispositional account of colour, it is often thought that we need a fairly detailed story of how initial basic red sensations people find similar can generate reliable colour judgment in less than ideal conditions, e.g. at dusk or under fluorescent light. Here, normality is made good by practices of correction.[1] Equally, I say, the epistemological status of moral belief and judgment must depend on the resources we can draw on to correct our responses. Epistemology, properly understood, is not about meaning (as the proponents of thick concepts might have it) but warrant and authority. Here, a moral cognitivist defending a substantial notion of

[1] Cf. Philip Pettit's account of colour stability as resting on two assumptions. "The intrapersonal assumption is that something is amiss if I find myself reliably inclined to make different judgments at different times – in particular, judgments different by my own lights – without any justifying difference in collateral beliefs or whatever. The interpersonal assumption is that something is amiss if you and I find that we are reliably inclined to make different judgments – again, judgments different by our lights – without any such justifying difference." (Pettit, 1991, 600-1; see also Pettit, 1990)

truth faces the same difficulties as a projectivist who aims to license a use of true and false-predications within a non-cognitivist framework.[1] A sentimentalist like myself need not align himself with either of them but he should be able to accommodate any corrective resources cognitivists and projectivists might employ.

4.1.4 Blackburn's Projectivism

Wiggins, as we have seen, does not offer us much of a story how and on which basis we are to improve our moral sensibility and avoid the relativistic pitfalls threatened by other response-dependent properties like U-ness. Projectivism similarly holds that moral judgments are judgments expressing sensibilities. Blackburn, inventor of Projectivism, writes:

> A moral *sensibility* ... is defined by a function from *input* of belief [or more generally, awareness of certain features, as Blackburn adds in a footnote] to *output* of attitude. (192)[2]

Features of the world are met with an attitude or response, loaded with value and projected back onto the world. The attitude or response (as output) is logically supervenient on other features (the input), known in other ways. In these slightly mechanistic terms, again, the projectivist account must apply across the range of response-dependent properties. Judgments of the comic, judgments of U, judgments of redness as well as moral judgments respond to properties as gilded (in Hume's words) by a sense of humour, a sense of U, a sense of colour and perhaps a moral sense. As in the biconditional representing dispositions, there are features and responses. With a joke, we may have the timing of the delivery which is met with laughter and then loaded with comic value; for judgments of U, we have plastic

[1] I agree here with A. Price (1986) that the projectivist and response-dependent cognitivist (or "affective anthropocentrist" as Price has it, 219) sit in much the same boat. Unlike Price, I do not take this to speak in favour of response-dependent cognitivism. For another critique of Projectivism as a detour, see Wright, 1985.

[2] In this section, references to Blackburn's *Spreading the Word* (1984) are revealed by page-numbers only.

flowers, say, which are responded to favourably by a significant number of Sloanies and then considered to be the thing to have in virtue of being U, and so on.

> we speak and think as though there were a property of things which our sayings describe, which we can reason about, know about, be wrong about ... (171)

Under this picture, there is no real contrast between an analysis of moral beliefs as attitudes (answering to certain features which are then projected back, loaded with value) and an analysis of moral beliefs as dispositions to respond to certain features as moral features straightaway. What is more, the projectivist as well as the response-dependent cognitivist may only have a natural right to talk of properties *relative* to one or more senses of humour/U/colour/morals. The fact that attitudes correspond to certain properties we can reason about, know about, be wrong about does not secure that these properties are in any substantial sense objective. The cognitivist has to show (and has so far failed to do so) how talk of moral value can be substantially truth-apt; Blackburn's projectivist voluntarily undertakes the same task.

> [Projectivism] seeks to explain, and justify, the realistic-seeming nature of our talk of evaluations – the way we think we can be wrong about them, that there is a truth to be found, and so on. (180)

We may therefore hope that what the projectivist has to say about moral truth covers some of the grounds on which we are to correct and improve moral responses (attitudes or sentiments) – material that Wiggins withheld. In fact, Blackburn's offer is surprisingly similar to what Wiggins has hinted at. Wiggins wrote (SS?, 196):

> One may surmise that at any stage in the process some <property, response> pairs will and some will not prove susceptible of refinement, amplification and extension. One may imagine that some candidate pairs do and some do not relate in a reinforceable, satisfying way to the subjectivity of human life at a given time. Some pairs are such that refinement of response leads to refinement of perception and *vice versa*. Others are not. Some are and some are not capable of serving in the process of interpersonal education, instruction and mutual enlightenment.

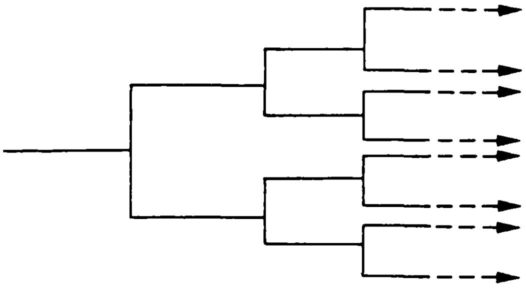
So what sense does the projectivist give to "notions of improvement, refinement, and progress towards correct opinion" (RF&MR, 176)?[1] The challenge is to show

[1] RF&MR is short for Blackburn, 1981, "Reply: Rule-Following and Moral Realism".

> how, given attitudes, given constraints upon them, given a notion of improvement and of possible fault in any sensibility including our own, we can construct a notion of truth. (198)

I detect several distinct strains of argument. Blackburn's treatment of moral truth is best set up by introducing his tree-diagram (199):

> We might imagine a tree... Here each node (point at which there is branching) marks a place where equally admirable but diverging opinion is possible. And there is no unique *M** [best possible set of attitudes] on which the progress of opinion is sighted. So there is no truth, since the definition [A moral judgment is true iff it expresses an attitude which is a member of the best possible set of attitudes] lapses. More precisely, truth would shrink to only those commitments [blanket term for belief with the proviso that beliefs might turn out to be best theorized about as something else, e.g. attitudes[1]] which are shared by all the diverging systems: truth belongs to the trunk.



As Blackburn sees it, moral truth is hard to obtain not because we (who competently disagree about what is good in any given case) sit on different trees but because we might find ourselves on different branches of the same tree. It is not that we disagree about basic values, "a core of attitude which we regard as beyond discussion" (RF&MR, 177), but that we don't know what they require in more complex cases. In order to arrive at a notion of truth sustaining substantial moral objectivity, however, Blackburn would need arguments against relativity in both respects – different branches *and* different trees. There might be equally admirable sensibilities on the same tree and there might be two or more incompatible trees of sensibilities.

Blackburn seeks to illustrate how we are to overcome the relativism imported by the tree-structure with an example from literary criticism. We are faced with the question: Is Ovid a better writer than Tacitus or is Tacitus better than Ovid? Hume (in

[1] Cf. Blackburn, 1993b, 365.

Of the Standard of Taste, 1985 (1741-77), 244) had argued that the answer is somehow *discretionary*.

> A young man, whose passions are warm, will be more sensibly touched with amorous and tender images [Ovid], than a man more advanced in years, who takes pleasure in wise, philosophical reflections concerning the conduct of life and moderation of the passions [Tacitus] ... We choose our favourite author as we do our friend, from a conformity of humour and disposition.

Blackburn admits that it might well be the case that different ages prefer different authors but he argues that this difference in preference must not be allowed to be a difference in value-judgment. If I claim outright that Ovid is the better writer I miss that Tacitus is indeed the better writer from another perspective. The improved sensibility is likely to hold both poets in equal esteem. On this account,

> an evaluative system should contain the resources to *transcend the tree structure*: evidence that there is a node *itself* implies that it is wrong to maintain either of the conflicting commitments. (201)

Once I acknowledge that there are features of a situation which appeal differently to different people I am constrained "to argue and practise as though the truth is single" (ibid.). What kind of constraint is this? Why are we to think and act as if there was only one best set of attitudes, responses or sentiments? As a conceptual claim internal to "serious evaluative practice" (RF&MR, 181) it amounts to no more than a restatement of the old intuition that the content of moral judgment is somehow not up to us. Moral judgment must not allow for individual variation in attitude. "The correct opinion ... is not necessarily the one we happen to have, nor is our having an opinion or not the kind of thing which makes for correctness" (219). Since what is right is not up to us, we have to take account of divergent opinions we consider to be competent. As Blackburn presents it, this claim is merely repeated not defended. Blackburn needs to argue why I should not be happy to accept that different things appear good from different angles and to different people? It may even be that the apparent impersonality of morals is best theorized about as something else, to use Blackburn's phrase. (Morals, a Marxist might say, are interests in disguise: The Gulf-war was fought not to liberate a small country (impersonal rule) but to ensure continuing access to important oil-fields.)

A bit further on in *Spreading the Word*, we find an altogether different argument which seems to provide a pragmatic rationale for excluding value-judgments as discretionary (i.e. representing equally admirable but rival sensibilities on the same tree). In his defence of the principle of bivalence in the construction of truth, Blackburn explains why in many practical cases we cannot shelter behind incompleteness.

> Because D*p* answers nothing. [D*p* symbolizes the judgement that it is discretionary that *p*]. We *need* to know (for instance) whether the contract is valid, and one party to pay the other, or invalid, and vice versa. For this pragmatic reason a judge must *think and argue as though* there is one proper verdict ... (207)

There is of course harmless incompleteness: a question that is to be decided might not be a moral issue at all. As a test for this I suggest: no sanction or reward would be called for if the decision went either way. But then again, as Blackburn rightly notes, "a wish not to discuss a choice in moral terms is itself describable as a moral attitude" (204). Either you treat the case as moral or you don't; discretion again would answer nothing.

This seems to me to be a much better argument but it is not an argument internal to "serious evaluative practice" if that means "moral practice". I may be pragmatically prevented from finding two things morally equal if I have to base a decision on that judgment. The same constraint, however, applies across practical decision making. There is a classic fable in which an old dog sets out to welcome his unexpectedly returning long-lost master but is called back, in his stride, by his present mentor. Caught in between, the dog finally sits down and falls asleep. Such behaviour, we agree, is no good guide to life.[1] It might not be a good guide morally speaking but equally, it is certainly not a good guide, say, prudentially speaking. The trunk of the

[1] This *Anekdote*, I believe, is to be found in the writings of the German pre-romantic poet Kleist. I did however not succeed in tracing the reference.

Hare may be seen (though he disowns it, claiming merely linguistic or logical support) as giving a pragmatic justification for introducing his *critical level* of moral thinking: "when we are not able to rely on our intuitions, either because they conflict in a particular case, or because we are uncertain what are the right intuitions to cultivate, we have to do the best we can [i.e. reason critically]" (Hare, 1989 (1986), 112). The "have to" of the last clause may be a pragmatic one: we cannot live from conflicting moral principles. For further discussion, see section 4.2 below.

prudential tree might not be identical with the trunk of the moral tree. What is more, many people would hold that there is more than one prudential tree. The core of what is prudential to pursue might be relative to what you want. Similarly, as far as morals go, the pragmatic argument only shows that you have to make a decision between sensibilities *you* consider equally admirable. These sensibilities might be moral, prudential, aesthetical, etc., and there might be more than one tree of sensibilities in each case.

Blackburn suggests a method to arrive at one answer; we find the candidate for evaluative truth by putting the onus on one side: "unless there is a proof of p, the proof of an absence of one *counts* as a proof of $\neg p$" (209). If you cannot cite features of Tacitus' writing that appeal to a competent reader advanced in years I can be said to be justified in my considered opinion that Ovid is the better author. Suppose you are the sole juror of a literary prize which cannot be split. Short-listed are Ovid and Tacitus. Pragmatically, you are constrained to come up with a verdict (you cannot fall asleep for good). The combination of pragmatic argument and conceptual intuition then only shows that, internal to serious evaluative (in this case probably aesthetic) practise, you have to take account of what other admirable attitudes towards features of Tacitus and Ovid might consist in. *You are constrained to argue to the trunk of one tree though we do not know which one it will be.*

We may best assess the situation by reflecting again on the response-dependent predicates and the contrasting properties they stand for. So far, Blackburn may have successfully established that 'good' cannot be like 'funny'. It cannot be discretionary whether something is good as it may be discretionary whether something is funny. This is why amusement is not often treated as a serious evaluative practise.[1] 'Good', however, might still be like 'U'. It might be relative to the responses of a particular social group. This would satisfy the intuition that it is not up to me to determine the

[1] Some theorists of laughter might want to deny this. I do not need to commit myself here. I only say, if finding something funny is discretionary and not a serious evaluative practice (as it may be), then finding something good differs in important respects.

best set of attitudes (Blackburn's first argument); it would also bow to the pragmatic constraint that truth is located in the trunk (Blackburn's second argument). The tree however might represent the sensibilities of a particular social group. Moral truth, on this account, comes out as insubstantial or relative.

Though Blackburn claims that "the deep problem of *relativism*" (199) can be defused with the arguments above, he in fact offers some arguments why there cannot be competing trees each representing the best possible set of attitudes. The first argument is again a conceptual argument – or perhaps better, the statement of an intuition. It is the argument from second-order attitudes. If we ask people which attitudes they find admirable they come up with similar answers. There is a contingent consensus among mankind about "the general character of the wise man" (RF&MR, 175-6). And similarly there is, Blackburn claims, surprising unanimity about attitudes we should not desire to have. We may not want to admire things "because of propensities we regard as inferior: insensitivities, fears, blind traditions, failures of knowledge, imagination, sympathy" (ibid.).

Suppose now a number of people are of the opinion that it would be right to kick dogs. If what is good depends on a contingent consensus there are two related dangers. Either moral value comes out as *relative* to the attitudes of a particular social group (like U-ness), or it is substantially objective but depends on the contingent responses of a *majority* of mankind. The attitudes the dog-haters express may satisfy Blackburn's two conditions above. Kicking dogs is not up to each individual and there also is a verdict: "It is right to kick dogs". Dog-haters in the majority endorse a certain sensibility: "one which lets information about what people feel dictate its attitude to kicking dogs" (218). According to Blackburn, we need *not* endorse this trunk of a tree of sensibilities. Why? Because "nice people do not endorse such a sensibility" (ibid.). There is second-order attitude agreement that "niceness" is violated by the first-order majority of dog-haters. It is definitive of 'nice' that it excludes inflicting arbitrary harm. Cruelty is unacceptable to nice people.

Unfortunately it is possible to extend the dog-haters' manoeuvre to second-order attitudes. Being nice for those people may include inflicting harm. Wanton cruelty may not pass the test but, as I mentioned against Wiggins, the thirst for vengeance seems to be a pretty universal tendency upon mankind, as is sympathy (which Blackburn listed above among the human tendencies deserving second-order approval). Already now, it seems possible to conceive of at least three competing trees of moral sensibilities: the tree with sympathy in the trunk, the tree with revenge in the trunk, and perhaps one more – Blackburn's nice tree. Moral judgments derived from sympathy express an attitude which is a member of the best possible set of attitudes (i.e. are true); a set of attitudes which is to be a candidate for moral truth gives a prominent place to revenge; the sensibilities we must endorse are the sensibilities of nice people. Each tree of sensibilities would represent quite a different set of first-order judgments. Sympathy-ethics may extend moral obligations to non-human beings; an ethics with a central place for vengeance might be deontological, incorporating a strict honour code; as for Blackburn's ethic of niceness, we have little idea what it would demand of us.

> Just as the senses constrain what we can believe about the empirical world, so our natures and desires, needs and pleasures, constrain much of what we can admire and commend, tolerate and work for. There are not so many liveable, unfragmented, developed, consistent, and coherent systems of attitude. (197)

Thus was Blackburn's – as we now see premature – optimism. Our contingent moral natures constrain less than we need for moral truth.

We should not see Blackburn as resting with this rather disappointing result. From other material, we may again construct a second, better argument. It builds formal constraints on pragmatic roots, taking up the notions of consistency and coherence. If there is more than one tree of sensibilities as possible candidate for moral truth, things might easily get out of hand. Sensibilities might change; what I find good today I might not find good tomorrow because I may have jumped moral trees, so to speak. Yesterday, I was poor and defended the welfare state, today I am inheriting a fortune and reject taxation – both on entirely general grounds, thus satisfying all of Blackburn's previous arguments.

Blackburn calls jumping trees a "fickle function – one which has an apparently random element through time, or across similar cases" (RF&MR, 180). What is wrong with fickle sensibilities? They violate a requirement of moral consistency – the requirement to respect the supervenience of the moral on the natural. Supervenience is first introduced as a conceptual claim:

> ... it seems conceptually or logically necessary that if two things share a total basis of natural properties, then they have the same moral properties. (184)

In the case of a judgment about kicking dogs, for example, there is some moral feature, "cruelty", which is supervenient on the natural property of "pain to the animal", which yields "disapproval and indignation as the output" (218). (Remember, on Blackburn's definition a sensibility is the function from input of belief, or awareness, to output of attitude.) If, in the lights of our sensibility, "cruel" is supervenient on "painful" we may not deny other situations of inflicting pain the title of cruelty.[1] The pragmatic rationale subsequently given is that non-supervenient sensibilities offer "no guide to practical decision-making" (186).

> Our purpose in projecting value predicates may demand that we respect supervenience. If we allowed ourselves a system (shmoralizing) which was like ordinary evaluative practice, but subject to no such constraint, then it would allow us to treat naturally identical cases in morally different ways. This could be good shmoralizing. But that would unfit shmoralizing from being any kind of guide to practical decision-making (a thing could be properly deemed shbetter than another although it shared with it all the features relevant to choice or desirability). (ibid.)

If I don't know *why* I evaluate situations and prospects the way I do, my ranking of preferences and attitudes will become confused. This in turn is likely to result in the joint frustration of my attitudes and preferences. Yesterday, say, I ranked procrastination high; today I marvel at the prospect of becoming a chief executive. Today, that is, I might not see having a lazy and agreeable life as a feature relevant to

[1] Supervenience does not constrain you to denounce kicking dogs. It is only claimed that *if* you call inflicting arbitrary pain cruel you are committed (by pains of some kind of inconsistency) to denounce other cases of inflicting arbitrary harm too. There are difficulties surrounding the notion of *other* cases sharing the *same* basis of natural properties. It is commonly argued against supervenience as a tool in moral discussion, that no two cases are identical. Blackburn's pragmatic support of the demand for supervenience seems to rely on a notion of relevant similarity between two cases. This is a logically weaker but substantially stronger claim since it seeks to exclude "fickle functions" over time, something the logical thesis does not. For more on these difficulties, see the discussion of Hare below.

choice and desirability. Today, it is power that counts. It may not be always a bad thing to change one's mind but if I do it too often I shall have neither an agreeable nor a powerful life.

In his earlier paper "Rule-Following and Moral Realism", Blackburn suggests a slightly different pragmatic rationale. Disrespect for supervenience will limit, Blackburn claims, my possibilities of communication:

> A fickle sensibility is going to be difficult to teach, and since it matters to me that others can come to share and endorse my moral outlook, I shall seek to render it consistent. (RF&MR, 180)

In continuation of the example above one might say, if I do not feel able to speak over time in favour of procrastination (or achievement, respectively) I will find it difficult to communicate, teach and seek endorsement of my attitudes towards procrastination (or achievement). Nothing speaks against combining the two claims from *Spreading the Word* and "Rule-Following and Moral Realism". Together they make a powerful pragmatic case against random variation of attitudes.

What are the implications for Blackburn's account of moral truth? The second, combined, pragmatic argument, though sounding similar to the first, is more complex. On the first pragmatic argument, to be a guide to decision-making only required that we arrive in the trunk of one tree at any given time. If I want to base a decision on a judgment, be it moral, aesthetical or prudential, I will have to *make* a judgment and not retire into discretion. The second pragmatic constraint limits what can count as a *successful* set of judgments over time and similar circumstances. If we want to communicate and teach our practical attitudes, as well as avoid confusion in ranking our own preferences, we are prevented from jumping trees, and generally should accept only those trees as candidates for moral truth that respect supervenience.

I believe it is right to search for pragmatic groundings to formal constraints like supervenience. Still, it may just be that humankind's contingent will to live from unfrustrated preferences and to communicate does not impose sufficient restrictions on choosing between trees of attitudes, and *a fortiori* not even construct a notion of moral consistency as strong as supervenience. Blackburn, officially, is always happy with the first implication. Repeatedly he pours scorn on philosophers who endeavour to portray moral failing as a violation of formal constraints (more often than not of rationality).[1] Blackburn may be right that "it takes a value to make a value" (1993b, 370). There may be no pragmatic value from which we can derive a suitable constraint of consistency that would allow us to decide between competing sets of attitudes. On the other hand, this leaves Blackburn dangerously short of options. We have to judge Blackburn as having failed to earn a notion of moral truth outside such pragmatically based formal requirements. He owes us the underlying value that "makes value" that can be the object of true judgment.

How good would requirements of moral consistency have to be in order to support decisions between competing sets of moral responses, sensibilities or attitudes? Richard Hare's writing has been in the forefront of this discussion.

4.2 Moral Consistency

Hare is quite rude about anything that looks remotely like Response-Dependent Cognitivism or Secondary-Quality Realism. In the absence of explicit modes of correction, philosophers defending, say, an analogy of colour- and value-judgments are summarily labelled as "intuitionists". What do we do, Hare may be seen as challenging those philosophers, when two disputants disagree about the morally appropriate responses to a given situation? Are their responses or attitudes located on the same tree of sensibilities, representing apparently equally admirable branches? Are they found on different, morally incompatible trees? In any case, how do we decide between conflicting responses, attitudes or dispositions, be they equally admirable or not? With redness, we know that we have to arrive in the trunk of one and only one tree – but with wrongness?

[1] e.g. *Spreading the Word*, 222, n. 6.3: "This is the permanent chimaera, the holy grail of moral philosophy, the knock-down argument that people who are nasty and unpleasant and motivated by the wrong things are above all *unreasonable*..."

> If ... a dispute arose about the redness of some object, then we should have to say that one of the disputants was either colour-blind, or mistaken in his use of the word 'red'. But in the case of wrongness the intuitionist will not say (because he is not a naturalist) that the dispute between the two people is a verbal one; he will say that it is a difference between the moral reactions that they respectively have. He is required by his theory, therefore, to say that one of them (though he has not told us how to say which) is 'morally colour-blind', that is, that he has the wrong moral reactions ... (1989 (1986), 105)

Hare assumes that a genuine moral dispute is never directly verbal (there is no coherent descriptive, naturalistic interpretation of 'good' independently of human dispositions and attitudes; here response-dependent cognitivists and projectivists will agree); moral dispute is also not indirectly verbal in the guise of a disagreement about the application of less general predicates like 'cruel', 'nice', 'industrious' or 'lazy'. Thick moral terms are, in Hare's words, "secondarily evaluative" (MT, 17).[1] They only mistakenly give the impression "that our conceptual scheme, and the very meanings of our words, from which we cannot escape, commit us to the adoption of certain norms of conduct." (ibid.)

In terms of Blackburn's account of moral truth, Hare explicitly rules *in* that morally disagreeing parties might find themselves on different trees. Our attitudes, inter- or intrapersonally, may clash in a fundamental way.[2] Hare, however, believes that in contrast to intuitionists (be they response-dependent or projective) he commands the corrective resources to arbitrate any clash of moral opinion and attitude. Considerations of moral consistency will take us all the way. There is a unique set of best attitudes within our reach – the attitudes that conform to utilitarian critical thinking. In some sense, though Hare himself remains quietist, Harean critical moral judgments come out as true in that they express an attitude, disposition,

[1] See also F&R 2.7, 10.1 ff.; LoM 7.5. As in Chapter 3, Hare's main works are revealed by the following abbreviations. *The Language of Morals* (1952) = LoM; *Freedom and Reason* (1963) = F&R; *Moral Thinking* (1981) = MT.

[2] Hare writes typically: "we are bound to find ourselves in situations in which [intuitive principles or dispositions] conflict and in which, therefore, some other, non-intuitive kind of thinking is called for, to resolve the conflict" (MT, 40). This is stronger than Blackburn's conflict between apparently equally admirable attitudes.

preference, response which is member of a unique set of best possible attitudes, dispositions, preferences, responses.[1]

4.2.1 Moral Consistency As Consistently Realizable Attitudes

A popular approach to the notion of moral consistency is to read it as the requirement to make like judgments in like cases. (This was the burden of Blackburn's argument from supervenience.) It seems true enough that we are here in the grip of strong intuitions. Statements like this example of Hare's are likely to be met with some kind of incomprehension:

> 'Jack did just the same as Jim, in just the same circumstances, and they are just the same sort of people, but Jack did what he ought and Jim did what he ought not'. (MT, 81)

On closer examination, however, the notion of moral consistency is quite elusive. For non-normative beliefs there is a well-established account of consistency linked to the construction of truth. Roughly, two beliefs are inconsistent iff their propositional contents cannot be true together. If I believe both that the milk is boiling over and that the milk is not boiling over I am forced by pains of inconsistency to drop one of my beliefs. What makes the requirement of consistency so compelling for non-normative beliefs? The most obvious pragmatic rationale seems to be that if I judge *p* both to be the case and not to be the case I shall not be able to act upon my judgment. Should I take the milk off the stove or should I not?

Now a moral cognitivist may be tempted to transfer this account straightforwardly to moral beliefs. If I both judge morally that *p* and that $\neg p$ my judgments cannot be true together. If I believe (with truth-value) that kicking dogs is good and its denial, consistency seems to demand that one of the propositions must be

[1] In order to locate Hare's project within our present discussion, I use attitude, response, reaction, disposition and preference interchangeably as picking out motivational states of minds. This is not as good a philosophical practice as I should wish but again excessive terminological rewriting would be more confusing than enlightening. Hare himself dithers between motivations, reactions, desires and dispositions (while officially being committed to moral judgments as expressing *preferences*).

false. At this stage in the discussion, however, this move is not open to the moral cognitivist. For we are in the process of *questioning* whether a notion of moral truth can be earned on the assumption that moral beliefs are sentiments i.e. express something like responses, attitudes, preferences, etc. (and we have already considered and rejected several arguments to that effect).[1]

Though, within our programme, we cannot rely on a notion of truth to construct a requirement of moral consistency, there can be, within moral discourse, a constraint very similar to non-normative consistency. What is inconsistent in taking both the attitude that I ought to remove the milk from the stove and that I ought not? Naturally, we might say, not both attitudes can be realized together.[2] Again, there is a clear pragmatic rationale for constraints of consistency in this sense. It is not possible to act upon attitudes, preferences, etc., that exert conflicting demands on action at any one time. (This may remind us of Blackburn's first pragmatic argument [page 99 above]).

If a set of normative judgments is consistent in that it expresses consistently realizable practical attitudes (1), this evidently constrains little which set of judgments we are talking about. It may be a set of judgments that favours taking boiling milk off the stove or it may be the opposite; it may even be that today my consistent set of judgments tells me to remove the milk, tomorrow to leave it burning. Consistency in the sense of the last paragraphs simply excludes conflicting normative demands *at any one time*, be they moral, prudential, aesthetical or demands of etiquette.

[1] Philosophers leaning toward a notion of truth as corresponding to hard facts that are part of the fabric of the world tend to put the point even stronger. On such a view, the world cannot conform with inconsistent beliefs. That is why p or $\neg p$ must be false. For moral beliefs, there is no world they can be said to conform to. Two inconsistent moral beliefs, therefore, may be equally good. See for example Williams (1966): "Consistency and Realism".

[2] There have been several attempts to formulate a logic of attitudes in this vain. See e.g. Blackburn, 1984, 189-196; 1993a (1988), 182-197; 1992b, 947. Gibbard (1990, 98) says that "normative statements rule each other out if their representations have no factual-normative world in common". A proposal of this kind may then be used to preserve sameness of meaning for normative statements in unasserted contexts. This would circumvent the so-called Frege-Geach objection I discuss in Chapter 2.2 above.

4.2.2 Moral Consistency As Supervenience

The next more substantial notion of consistency (2) is precisely designed to prevent apparently arbitrary variation of judgment across time and place (Blackburn's "fickle sensibility", [p. 103 above]). To be morally consistent in this stronger sense is to respect the so-called supervenience of moral judgment and belief. As Blackburn introduced it:

> It seems conceptually impossible to suppose that if two things are identical in every other respect, one is better than the other. Such a difference *could* only arise if there were other differences between them. (183)

What precisely is this conceptual impossibility? Hare's notion of universalizability in *Moral Thinking* is here elucidating.

> Moral judgements are, I claim, universalizable in only one sense, namely that they entail identical judgements about all cases identical in their universal properties. (MT, 108)

The conceptual impossibility derives from a peculiar contradiction or inconsistency (2):

> if we make different moral judgements about situations which we admit to be identical in their universal descriptive properties, we contradict ourselves. (MT, 21)

Were these different judgments made about the *same* situation, "they would be inconsistent with one another" (ibid.). Thus, moral inconsistency (2) in the stronger sense of disrespect for supervenience is defined with reference to consistency (1) as the requirement of consistent realizability at any one time.

Before we assess why moral consistency (2) or the respect for supervenience should matter to us, let us try to make more precise what that requirement would commit us to. Suppose you believe that a bicycle ought to be moved so that you can park your car (another of Hare's characteristically dry examples). First, respect for supervenience requires that you talk of the situation in terms of its universal properties so that it may become possible to establish whether there be like judgment in like circumstances. Yesterday, say, you yourself had been approached by a car-owner to move your bicycle and you had refused to do so. Morally it will not do, so the story goes, that you are you and he is him, that today is today, and yesterday was yesterday, London is London and Oxford is Oxford. Swap all indexical terms and references to time and place with unversalizable descriptions.

As many of Hare's critics have remarked, this will not make much of an arbiter in situations of conflict since no two situations are identical. It is a metaphysical truth that no two situations are one. So it should always be possible to supply universalizable descriptions of the "natural base" of a moral judgments that fit your particular case and interests. Instead of saying, "I drive a BMW and you don't, bugger off", you might say, "A man driving a car from a near-Alpine country where the people drink lots of beer ought to be given way to". Fortunately (you think) in yesterday's incident you had not been confronted by a BMW but a Rover. If it would be pointed out to you that Rover has been bought up by BMW, you just add another feature, so the objection goes. Respect for supervenience is no good arbiter between conflicting judgments.[1]

Hare meets the charge by introducing hypothetical thinking supported by what has aptly been called the Conditional Reflection Principle (Gibbard, 1988, 60). Hypothetical thinking first:

> ... we may imagine hypothetical cases, absolutely identical in their universal properties, but with the roles reversed, and look for a universal principle covering such cases that we can accept. The fact that there cannot be identical actual cases is no bar to this. (Hare (Seanor & Fotion), 1988, 211; see also MT, ch. 2.4)

The principle of Conditional Reflection then reads:

[1] Many of Hare's critics, amongst them Jonathan Dancy, have objected to the notion of universalizability on the assumption that it means: "a person who makes a moral judgement is committed to making the same judgement in any relevantly similar situation" (Dancy, 1993, 80). (This is the most obvious reading from *Freedom and Reason*, ch. 2.) The strategy is then to attack the idea of *relevant* similarity. Does it include all properties of a situation that were a person's reasons for his judgment? (Counter-argument: "... in a new case there may be a strong reason against the judgement which was not present in the first case - a defeater as we may call it", 80). Does it include all reasons in favour as well as reasons against? (Suppose somebody giving flowers in order to seduce instead of expressing regret. "There are just too many potential defeaters for the absence of each one to count among our original reasons ...", 81). Does it include any natural properties? (Then universalizability as supervenience becomes trivial since no two cases are identical). Hare, however, confesses to be happy with this reduction. "It is commonly thought that I have changed my understanding of the thesis of universalizability. This is not so." (Hare (Seanor & Fotion), 1988, 203)

> ... I cannot know the extent and quality of others' sufferings and, in general, motivations and preferences without having equal motivations with regard to what should happen to me, were I in their places, with their motivations and preferences. (MT, 99)

To draw moral conclusions from supervenience, we only need to test the *one* situation we are judging about. Thus the problem does not arise how to get from one case to any differing in its universal properties. The one case exactly similar to the case judged about, is the imagined case with role-reversal. In the case of the bicycle and the car, you hypothetically put yourself onto the "receiving" end of your judgment. Would you judge the same, you ask yourself, if you were owner of the bicycle and not of the BMW? The Conditional Reflection Principle ensures that in thinking about the other person's perspective, you acquire his motivations and preferences as if they were your own. As Hare puts it, if asked "'How do you feel about being put yourself forthwith in that position with his preferences?', I shall reply that it would be *me*, I do now have the same aversion to having it done as he now has" (MT, 98).

Much is unclear about that principle, its status and scope. In hypothetically changing roles, will I acquire (a) the preferences I would have were I put in his situation or, much stronger, will I acquire (b) his preferences as mine? Since both options seem possible given our psychological make-up, which one should we adopt and how would each option be apt to support decisions between competing sets of moral responses, attitudes or preferences? And more generally, will hypothetical thinking analytically carry any of these implications? Hare believes that demands of consistency will ultimately lead us to take up the second option which implies giving equal weight to other persons' preferences as to your own in any given case; that is, in any given case, consistency will require that you act upon a utilitarian maxim. There will be no problem with how to get from case to case, since each case on its own, thought through hypothetically, will imply decisions in accordance with the same maxim: giving equal weight to the preferences of all affected parties.[1]

[1] If successful, this strategy should rebut e.g. Dancy, 1993, 260.

So far, we may grant that there is a psychological capacity for thinking hypothetically about any given situation.[1] There seem to be at least two ways in which one could exercise this psychological capacity, be consistent (2) in that one respects supervenience (now the demand for hypothetically putting oneself in the others' shoes) and still not reach convergence about what is the right course of action. First, in any exercise of the capacity of hypothetical thinking, there is a limit over which this capacity will not be exerted. Some people may seek to enter the "mind" of a butterfly who is cruelly deprived of his wings, other people find it difficult to think hypothetically of any preferences outside their own family, class, country, religion, race or species. Considerations of consistency (2) will not guarantee that two disputants agree about the range of beings they are prepared to include in their conditional reflection. (A pragmatic rationale may restrict the circle to people you want to communicate with). A BMW-driver may refuse to think of a cyclist under the Conditional Reflection Principle. He may be happy to apply his judgment to hypothetical, role-reversed situations (just as the respect for supervenience demands) but he may simply not take any notice of how he would feel as a cyclist. He thinks it impossible for him to be like a cyclist. If he were a cyclist, he admits, his present judgment implied that he would have to remove his bicycle but he considers this hypothetical situation as far-fetched as his becoming a butterfly. A BMW-driver, he is confident, simply will not be a butterfly, never mind a cyclist. Here, for disagreeing parties, the Conditional Reflection Principle may not extend over the same range of cases.[2]

The second case in which we may not reach convergence about the right course of action allows that there is prior agreement about the circle whose members are seen

[1] This psychological capacity will reappear in section 4.3.5 below as part of "sympathy".

[2] Hare's reply to this objection is that any criterion restricting the circle of beings conditionally reflected upon may be turned against the restrictor. (See Hare on Singer in Hare (Seanor & Fotion), 1988, 273-4.) This does not solve the problem since whatever your criterion there must be a cut-off point somewhere. The circle can not expand indefinitely, to use Singer's phrase (Singer, 1981). No sensible account of conditional reflection will extend to cockroaches, say. Since there must be a cut-off somewhere, the point might be at different places for different people.

as adequate objects of hypothetical thinking under the Conditional Reflection Principle. Suppose again that the BMW-driver submits to the constraints of supervenience. Under the formula of hypothetical thinking, that means the BMW-driver reflects whether he could endorse his judgment even if he found himself on the receiving end of it. This time, he finds it possible to imagine himself as a cyclist and thus acquires all the preferences (a) *he* would have as a cyclist. He comes to the conclusion that, as a cyclist, he still would prefer to have the bicycle removed. Indeed, he comes to the conclusion, at least so he says, that bicycles always should make way for cars. Thus he seems to be a model of consistency (2). The cyclist, however, might not agree. After applying the Conditional Reflection Principle himself (under reading (a) of the respective preference), he finds that, as a BMW-driver, *he* would allow the bicycle to stand. What is more, yesterday, we had imagined, the BMW-driver was cycling himself and, after being confronted by a Rover, refused to move his bicycle so that the Rover could park. If the test for supervenience has to be conducted through hypothetical thinking (since no two cases are identical), the BMW-driver-cum-cyclist may in full consistency (2) hold yesterday that the bicycle ought to stay and today that the bicycle ought to be moved. Thus Hare might be caught on the horns of a dilemma. Either he accepts consistency as demanding that the same universal properties may be met by the same judgment (with the consequence that always a different universal property may be found, since no two cases are identical), or he accepts that consistency demands hypothetical thinking in accordance with the principle of Conditional Reflection (then, if such thinking does not yield convergence in judgment (as it might not), the requirement of consistency permits fickle judgment across time and place). In each case, consistency (2) would be a bad arbiter of moral conflict.

4.2.3 Moral Consistency As Taking Equal Account of Different Preferences

Hare believes he can escape the dilemma by arguing that constraints of consistency (2) (as tested in hypothetical thinking) yield consistency (3), that is convergence on one unitary principle. On Hare's reading, consistency (3) demands that we take equal account of different preferences. How is this to be done?

According to Hare, BMW-driver's and Cyclist's preferences may have been tested for consistency (2), still they remain preferences. As preferences (even universalized ones in the sense of consistency (2)), they command no special moral authority. We need some non-intuitive critical way of assessing conflicting preferences. ("To insist on the *prior* authority of the moral intuitions that one starts with is simply to refuse to think critically." MT, 179). Given the nature of moral beliefs as universalizable preferences, only one critical method is available: to give equal weight to all preferences.

In which sense am I inconsistent in giving more weight to my preferences than other people's? I am inconsistent (3) as if I would discount one of my own preferences, but in contrast to consistency (2) the preferences (a) are not ones I (the BMW-driver) would have as a cyclist, they are preferences (b) the cyclist has as a cyclist.

> ... if I fully represent to myself his situation, including his motivations, I shall myself acquire a corresponding motivation, which would be expressed in the prescription that the same thing *not be* done to me, were I to be forthwith in just that situation. But this prescription [expressing a preference] is inconsistent with my original 'ought'-statement, if that was, as we have been assuming, prescriptive. (MT, 109)

In short, the inconsistency is derived on the assumption that the only critical way of assessing preferences involves full representation of each affected person's situation enabling the acquisition of preferences I would otherwise not share. Inconsistency (3) is then like inconsistency (2) apart from a different mode of acquiring the preferences to be tested for consistency (1).

As we have seen from the dilemma stated above, if Hare's move from consistency (2) to consistency (3) does not succeed, the conception of consistency (2) itself may be threatened. I contend now that on the assumption that moral judgments express preferences, attitudes (or any other motivational state) several other ways of critical thinking may be available. I shall give just two. One is to reflect critically on the natural tendencies most of us share, another to reflect critically on the mutual benefits of cooperation.

The first may be seen as an extension of Blackburn's argument from second-order attitudes. Our morally acceptable attitudes or preferences are those we desire to have. This certainly is a way to think critically about moral judgments: Are the attitudes or preferences expressed by my judgment the ones I would choose to attain and cultivate (perhaps given certain other conditions, e.g. viewed over a longer period of my life, avoiding self-deception, etc.)? To think critically about our moral judgments in this way does not guarantee that we converge on one set of attitudes or preferences we all agree we ought to cultivate, but it constitutes a rival mode of corrective thinking to Hare's suggestion.[1]

The second mode of critical thinking that came to mind will be familiar from the contractualistic tradition: Are the attitudes or preferences expressed by my judgment conducive to achieving or sustaining mutually beneficial cooperation? ("If I break my promise today will I be able to rely on yours tomorrow?") There may be doubts whether judgments assessed in this way can count as genuinely moral; there may also be doubts whether, under contractualistic critical thinking, there will be convergence on a unique best set of attitudes. Still, it is a way to think critically about moral judgment.[2]

Returning to Hare, what justifies his confidence that giving equal weight to rival preferences is the only mode of critical thinking applicable to moral judgment? Hare

[1] Hume introduces sympathy as a tendency most of us share and approve of. (See his account of the natural virtues, *Treatise* III iii 1. I discuss Hume in the Appendix.) For a conception of value as what we desire to desire see for example David Lewis' "Dispositional Theories of Value" (1989, 113-4): "*x* is a value iff we would be disposed to value *x* under conditions of the fullest imaginative acquaintance with *x*." The classic introduction to second-order desires is Harry Frankfurt's "Freedom of the Will and the Concept of a Person" (1971). Brandt's full-information account of rationality trades on related ideas. According to Brandt, desires count as rational if they would survive a process of "cognitive psychotherapy", that is full exposure to all relevant facts (Brandt, 1979, 11, 113).

[2] Classical exponents of contractualistic theory are of course Hobbes (1651) and later Hume in his account of the artificial virtues (*Treatise* Book III part ii). Modern versions are suggested by among others David Gauthier and John Mackie. For more about modern contractualism, see section 4.3 below. Hume's account of the artificial virtues is discussed in the Appendix.

believes moral judgments to be *universalizable* preferences, that is, preferences that bow to certain constraints of consistency. Given that moral judgments must be (as conceptual truth) consistent, the natural way of thinking critically is to extend consistency from consistency (1), excluding conflicting preferences at any one time, over consistency (2) as respect for supervenience (which in hypothetical thinking supported by a weak reading (a) of the Conditional Reflection Principle again amounts to excluding conflicting preferences at any one time), to consistency (3) or the demand (supported by the strong reading (b) of the Conditional Reflection Principle) to give equal weight to different preferences, be they yours or others.

Why should moral consistency matter so much, given that other modes of critical thinking are available to us? Hare makes us believe that he relies on only one notion of moral consistency which is extended by introducing hypothetical thinking under the principle of Conditional Reflection. Harean moral consistency is defined simply as the demand to overcome conflicting preferences at any one time – preferences conflicting in that they cannot be jointly realized. If I prefer to move the bicycle and prefer not to move the bicycle, consistency demands that this contradiction in preference has to be resolved. As I said, there are good pragmatic reasons why this should matter to us. If I want to act upon any preference I have to make up my mind.

But why should it matter if today I prefer to move the bicycle and tomorrow I don't? A moral judgment expressing preferences or attitudes of this kind seems to disrespect supervenience, the demand to judge alike in like situations. But why should inconsistency (2) as disrespect for supervenience matter? By introducing hypothetical thinking and the principle of Conditional Reflection, Hare seeks to show that inconsistency (2) must matter to us in precisely the same way as consistency (1). For preferences we think about under the Conditional Reflection Principle become our own. Inconsistency (2) then matters since, by implication, we cannot jointly realize our *own* preferences. This, however, is evidently false. There may be reasons why respect for supervenience should matter to us. Blackburn has given two [page 103-4 above]: If I judge today that I ought to procrastinate and tomorrow that I ought to achieve I may find it difficult to communicate and teach my attitude (which matters to me); secondly, if I change my attitudes too often I may become confused about what I really prefer. I miss features relevant to choice and desirability. If I judge today that I prefer to procrastinate and tomorrow that I want to become a chief executive it is likely that both preferences will become frustrated over a period of time.

What is more, as long as my attitudes do not become too fickle, disrespect for supervenience might not be a bad thing. It is perfectly possible to communicate, teach and live by slowly shifting attitudes (we all do). Blackburn appears to have missed the point; Hare implausibly holds that practical guidance and teaching presupposes principles applying regardless of time and place: "it is the function of moral principles to provide universal guidance for actions in all situations of a certain *kind* ... and one of the most important functions of singular moral judgements is to make clear what our principles are (e.g. in teaching them ...)" (MT, 88). To insist on perfect consistency (2) may in fact prevent that shifting of attitudes that is often necessary to achieve a moral consensus. Too much respect for supervenience may drive us into fanaticism. Instead of conceding I want you to move the bicycle and you want me to move the car one starts slamming each other with universal principles: "All cyclist always ought to give way to cars!" and *vice versa*.[1]

Since the pragmatic groundings for consistency (1) and consistency (2) differ so clearly we should be suspicious about Hare's suggestion that they are the same. What about consistency (3), the demand to treat different preferences equally? Pragmatic grounds for utilitarian principles are notoriously absent. Other people may matter to us, but not all people matter equally. Hare therefore adopts a pragmatic rationale for a method of justification, not for the result. It does not pay to stay outside morality ("Those who do not love their fellow men are less successful in living happily among

[1] The danger is obvious from Hare's life-long preoccupation with examples like the Nazi who is prepared to end in a concentration camp once it is discovered that he has a Jewish grandmother. Cf. F&R, ch. 9; MT, ch. 10.

them"; MT, 197); and once we feel a need to justify our own moral concerns against others who do not share them, or against ourselves in situations of conflict, we are driven into utilitarianism (see e.g. MT, ch. 11).

Because of its systematic and ambitious nature, Hare's project is threatened at many stages. I concentrated on his defence of the demands of moral consistency. If, as I hold, moral beliefs are sentiments, i.e. if moral judgments express motivational states of a particular nature, must they be subject to constraints of consistency? Clearly, Hare's notions of consistency (1) - (3) have different groundings, though they present themselves, under the veil of the Principle of Conditional Reflection, as the same fault of normative logic: contradictions among our present preferences. Of the groundings, I fully accept only the rationale for (1), the demand to resolve inconsistencies among our preferences at any one time. Consistency (1), however, is no good tool in normative debate. To hold that the bicycle ought to be moved is consistent (1) if I don't take any conflicting attitudes at that time. On the other hand, consistency (2) (or respect for supervenience) is a tool in normative debate. It requires us to treat like cases alike, say, the car and the bicycle yesterday and today. Since no two cases are identical, Hare is led into a complicated story of hypothetical thinking. This raises a number of internal difficulties (any Conditional Reflection reaches a cut off point; also Hare runs into a dilemma if he cannot support consistency (2) by consistency (3), the utilitarian demand to treat all preferences equally). Apart from being an insufficient normative tool, consistency (2) as the demand for supervenience lacks a compelling pragmatic rationale; it is not clear why consistency (2) should matter to us. Utilitarian consistency (3) as treating all preferences equally presupposes consistency (2) the respect for supervenience, which is not available in its strong form as a logical thesis if my criticisms are right. Additionally, the pragmatic methodological motivation behind the transition from (2) to (3) can be met by at least two rival modes of critical thinking.

Summarily I judge that demands of consistency (1) and (2) (in some weaker, non-logical, sense) are valid but make no good arbiter in moral conflict. At best, consistency may play the role of a test: whom do we accept as object of hypothetical thinking, and how far have we come in normative discussion? At worst, the demand for consistency may drive us into fanaticism.

4.3 Instrumental Rationality

We could understand the discussion of the previous section as pursuing the question which attitude we should take towards consistency. I have been insisting that nothing forces us to accept anything like Hare's compound notion as constitutive for rationality and that it pays to look carefully for the respective underlying rationales. I supported the conclusion that we should accept two weaker versions of moral consistency: We ought not to adopt contradictory attitudes at any one time; there also are good reasons not to shift our attitudes too quickly. We are, in my view, *normatively compelled* to modify attitudes that violate these requirements. Still, moral consistency in these senses restricts only weakly which attitudes we can take. I may consistently hold that a bicycle ought to be moved while you think it ought to stay, provided that we are both able to cite some general grounds for our respective beliefs, and we do not express violently fickle attitudes in the immediate context of the situation. Does this open the door to a rather random moral relativism? Under the supposition that moral judgments express attitudes (or, in my official terminology, dispositions to experience certain sentiments), we should hope to do better.

There are other ways in which we may feel drawn to accept or reject attitudes we find ourselves with. Just as it seemed coherent to ask *why* we should reject inconsistent attitudes we may ask whether we should approve of some attitudes as revealing natural tendencies most of us share (as perhaps Blackburn and Wiggins suggested), or - perhaps more compelling - whether we should adopt attitudes that enable and sustain mutually beneficial cooperation. In the latter case, the contractualistic tradition thought these constraints to be of *instrumental rationality.*[1] If I like to live an unharmed life should I agree not to harm others? And then, if I have committed myself not to harm others, should I not harm others when the opportunity arises without incurring costs? These questions we will have to examine now in some detail.

4.3.1 Theory of Choice

Like consistency, instrumental rationality is sometimes seen as an *a priori* constraint so that it seems as unintelligible to ask "I am thirsty but why should I drink?" as to exclaim "I should drink and not drink" (taking inconsistent attitudes at any one time). If you are thirsty and it is within your means to relieve your longing, you are, *ceteris paribus*, normatively compelled to do so. What is the force of this idea?

Under the standard theory of choice underlying much of recent economics and the social sciences, people are supposed to be rational in that their preferences form an ordering such that they always do what they most prefer. The theory attaches so-called expected utilities to consequences of actions, so that people can be seen as

[1] Under "contractualistic" I understand the tradition that models and perhaps justifies normative demands as the outcome of decisions under the constraints of instrumental rationality from our *actual* aims, desires or purposes. This is true of Hobbes, but neither of Locke, Rousseau or Kant. They defend for various reasons a *hypothetical* social contract, and stand in direct line to a prominent recent theory: Rawls' *Theory of Justice*. For "contractarians" (as I call this latter breed) the question is not if we actually have instrumental reasons to submit to given norms, rather a normative system is qualified as having been instrumentally chosen in a hypothetical situation. The characterization of the hypothetical choosing situation determines which set of norms it is rational to choose. Such an "Original Situation" therefore merely *reflects* the norms the choosers already have, the choosing situation itself is not the outcome of instrumental deliberation. Honderich rightly critisizes Rawls' contractarian label as misleading. The support the hypothetical contract gives to a normative system is the support of an "Ordinary Argument" (Honderich, 1975, 70) not of instrumental thinking. It brings merely into the open principles we already have – or may acquire on reflection.

Rawls might even agree with this description: "According to the provisional aim of moral philosophy, one might say that justice as fairness is the hypothesis that the principles which would be chosen in the original position are identical with those that match our considered judgments, and so these principles describe our sense of justice... When a person is presented with an intuitively appealing account of his sense of justice ... he may well revise his judgments to conform to its principles even though the theory does not fit his existing judgments exactly." (Rawls, 1971, 48)

Within my framework, Rawls' enterprise stands somehow between theories placing emphasis on basic intuitions most people share (e.g. Wiggins, McDowell, Blackburn; Ch. 4.1) and Hare's rigid consistency constraints (Ch. 4.2).

adopting what constitutes the best means to satisfy their preferences. This is an *explanatory* claim. People's behaviour is interpreted under a constraint of instrumental rationality which *reveals* their preferences. If you reach for a glass of water and not a piece of bread you are taken to be thirsty, having adopted the means to maximize expected utility. No gap remains between preference and choice. It is not that you chose the glass of water because you had a preference for doing so. We cannot ask why a preference should be satisfied. A gap between preference and choice may only appear if you count as irrational, i.e. if no maximizing interpretation can be given to your actions. For example, you may insist that you preferred the bread though you chose the water. To admit widespread irrationality of that kind, however, is bad news for the explanatory claims of the theory of choice itself since it would fail to describe behaviour it set out to describe. In so far as you are party to the interactions of the socio-economic world you have, as interpreted subject, no authority about what your preferences are. You prefer what you choose. Thus the question does not arise what you *ought to* choose, and whether you have indeed adopted the best means to satisfy your preference. The only *normative* element is interpretation under the constraint of maximization.[1]

In the greater number of so-called parametric cases, i.e. situations of choice in which a person takes her actions to be the sole variable in a fixed environment, the theory of choice appears to describe and predict socio-economic behaviour quite successfully.[2] In situations of strategic choice, however, where the outcome of one's actions depends on the actions of one or more other persons, the theory runs into a

[1] This abstract of rather familiar ideas about instrumental rationality derives from Ramsey and his axioms about preference ordering and subjective probability (1931). Pareto (1972 (1927)) pioneered the use of so-called utility indices as mathematical representations of the total of an agent's motivations; under the interpretative constraints of instrumental rationality expected utilities are attached to consequences of actions. The theory of choice was canonisized in Savage's four axioms (1972 (1954)).

[2] Even for decisions in fixed enviroments, puzzles arise. In a Russian roulette case, the removal of one of four bullets in a six-chamber-revolver reduces the probability of killing yourself more than the removal of only one remaining bullet, yet most people would be prepared to pay more for the removal of the last bullet, thus violating Savage's axioms. (Cf. Kahneman & Tversky, 1979)

number of deeper troubles. Most prominently, it struggles to explain basic cases of reciprocity. The behaviour of cooperating parties often resists an interpretation under the maximizing constraint as adopting the best means to given preferences.

4.3.2 Problems: Reciprocity

Imagine driving down a narrow country-lane when suddenly a car approaches from the opposite direction. Without previous conventions about behaviour on roads, should you keep to the left or to the right? Obviously you would do best to keep to the right if the other car does so too, but equally you'd better keep to the left if that is what the other does. The trouble is that under the maximizing interpretation the other thinks exactly the same. Without additional information there is no successful way to coordinate your approaches. In practise, this is of course what you would seek. You would try to provide each other with indications on which side you intend to pass. You might hold sharply to the left and see what the other does.

If coordination seems problematic even in cases where both parties pursue a *common goal* they can only reach together (not to collide, say) what about reciprocity where both parties have *competing interests* they can only satisfy through cooperative activity? The problem has been acutely formalized in the so-called Prisoner's Dilemma. Two rogues have committed a crime. Now in prison each is faced with the options to confess thus incriminating his partner or to keep mum. If only one confesses he is let off lightly for turning in State's Evidence while the other receives the maximum penalty (say 10 years). If both confess, they receive 5 years each (mitigating circumstances). If both remain silent they receive a much reduced sentence for lack of evidence (1 year each).

Predicting their behaviour under the constraints of instrumental rationality seems to suggest that both rogues will confess, thus paradoxically revealing preferences for serving 5 years in prison, though both sought to be let off as lightly as possible. A must have thought: if B confesses, I do better to confess; if B keeps mum, I do better to confess – and *vice versa.*

A variation of the Prisoner's Dilemma is the more general case of promising. A and B, foreseeing the situation, have promised each other to remain silent. But under the maximizing interpretation the parties will never successfully establish the institution of promising since the second party would always have reason to abort coordinated activity after it received the benefits advanced by the first party.

Now most of us are to some degree trustworthy, we keep many promises, we drive on the right (that is in Great Britain: left) side. This is part of our socio-economic behaviour and – as one might want to insist – one of the more reasonable parts. Admitting the existence of reciprocity, the theory of choice may move into two directions. Either the endeavour must be to show that people in successful coordination conform in some unexpected way to the constraints of instrumental rationality or the theory has to accept that people characteristically do not act rationally. In the latter case, the theory of choice gives up its explanatory claims and may therefore have to rethink the foundations of the maximizing conception. This is no easy undertaking since interpretations under the constraints of instrumental rationality were convincing in the first place *because* of its explanatory credentials. Once the gap between preference and choice reopens, we may ask again more fundamental normative questions: why should we maximize? Why should we take the second-order attitude to act out certain preferences and not others? I shall pursue this line of thought in time. But first let us return to the first response.

Might not a utility-maximizing explanation of reciprocity be available? Could it not be ultimately rational for the two rogues to keep mum and, in a wider sense, for most of us to keep promises? David Gauthier has taken this route, most detailed in Chapter Six of his *Morals by Agreement.*[1] There he argues that a rational person (in a

[1] Gauthier 1986. Gauthier claims that *Morals by Agreement* grew out of a deliberation of the Prisoner's Dilemma (v). My cursory treatment of Gauthier does not do justice to the intricate argument of his book. Still, the central difficulties facing normative systems derived from instrumental rationality can be brought out clearly from Gauthier's discussion. References to *Morals by Agreement* in this section

strategic setting) "chooses on utility-maximizing grounds not to make further choices on those grounds" (158). The rational person will do better overall if she disposes herself to become trustworthy.

> The disposition to keep one's agreement, given sufficient security, without appealing to directly utility-maximizing considerations, makes one an eligible partner in beneficial co-operation, and so is in itself beneficial. (162)

This would solve Prisoner's-Dilemma-type situations since A could count on B to perform after he (A) has undertaken the first move. So the explanatory claim is that a maximizing interpretation can be given to the tendency to be stably disposed not to abort cooperation single-handedly.

> we do not purport to give a utility-maximizing justification for specific choices of adherence to a joint strategy. Rather we explain those choices by a general disposition to choose fair, optimizing actions whenever possible, and this tendency is given a utility-maximizing justification. (189)

There are in effect two distinct actions. The action to dispose oneself to become trustworthy and the further action of making specific choices in strategic settings. While the former can be given a maximizing interpretation under the constraints of instrumental rationality, the second defies this attempt. An example invented by Parfit elucidates the point. While it may be rational to dispose myself to become a threatfulfiller (since anybody threatened by me will then be more likely to comply with my demands) it may not be rational to carry out my threat after it has been ignored (say, blow up the aircraft). Thus one may reject the claim that

> If it is rational for someone to make himself believe that it is rational for him to act in some way, it *is* rational for him to act in this way. (Parfit, 1984, 23)

The difference between the explanations of disposing oneself to become trustworthy and actually performing specific acts in strategic Prisoner's Dilemma situations can be brought out clearly. For each single strategic choice the party will do best that would only pretend to be disposed to be a restrained maximizer but abort cooperation after it received the advanced benefits.[1]

are revealed by page-numbers alone. Luce & Raiffa (1985 (1957), 94) attribute the Prisoner's Dilemma to A.W. Tucker.

[1] Cf. my discussion of Hume's Sensible Knave in Chapter 3.4.3.

Gauthier tries to convince us that this is not the case. Partially transparent as we are (Gauthier uses the term "translucent", 174) we cannot expect deceptive motivations to remain hidden. Thus Gauthier insists that there is good reason why we should become trustworthy, that is dispose ourselves to constrained maximization. This is a normative claim which explicitly presupposes a gap between preference and choice. We prefer in the first place to maximize but, after deliberation, choose not to do so.

> ... the capacity to make such choices [among dispositions] is itself an essential part of human rationality ... At the core of our rational capacity is the ability to engage in self-critical reflection. The fully rational being is able to reflect on his standard of deliberation, and to change that standard in the light of reflection. Thus we suppose it possible for persons, who may initially assume that it is rational to extend straightforward maximization from parametric to strategic contexts, to reflect on the implications of this extension, and to reject it in favour of constrained maximization. (183-4)

If Gauthier had given primacy to maximization as an explanatory theory a way out might have been to ascribe to cooperating parties a different set of preferences. In performing reciprocal actions, they reveal preferences, say, for trustworthiness which again can be given a maximizing interpretation under the constraints of instrumental rationality.[1]

Instead, Gauthier opts for the second, normative route. Though people do not always act ideally rationally, if they reflected critically they would. In fact they *ought* to. Gauthier still insists that the only critical considerations available to rational agents are maximizing reasons. Gauthier would have to argue why only maximizing reasons offer a "sure grounding" (17) since maximization, on our reading of Gauthier's exposition, just had failed to explain reciprocity. If the normative problem is which attitudes and dispositions survive critical reflection, we have to set the frame wider

[1] The Prisoner's Dilemma, however, will remain unsolved if the utilities are as they have been stated above (cf. Binmore, 1993). Purely self-seeking parties under the constraints of instrumental rationality will fail to cooperate successfully. This may be sad for moral theory but not necessarily bad news for the explanatory credentials of the theory of choice. Again it has been questioned whether the axioms of the theory of choice can allow motivations like "trustworthiness". Instrumental reasons are forward-looking reasons where utility is only attached to consequences. "Trust" or "Having promised", on this account, appear to be backward looking utilities referring to the history of the situation. (See Hollis & Sugden, 1993, 27 ff.)

and make from the start psychological assumptions about what kind of motivations people have. Curiously, Gauthier does precisely that. He adopts the maximizing constraint of instrumental rationality subject to Hobbes' material conditional "that each seeks above all his own preservation" (159). Thus a person's rational choices, the choices she ought to make, are those that further her interest - and I must stress again: 'interest' is here not coextensive with the technical term 'utility' in rational choice theory since people may attach utility to acting altruistically, adopting the best means to that end. Gauthier introduces the technical sense:

> Let us suppose it is agreed that there is a connection between reason and interest - or advantage, benefit, preference, satisfaction, or individual utility, since the differences among these, important in other contexts, do not affect the present discussion. (6)

– only to continue with interest in the sense of self-interest:

> Morality, we have insisted, is traditionally understood to involve an impartial constraint on the pursuit of individual interest. (7)

This terminological uncertainty may account for the confusion between the normative and explanatory role of interpretations under the constraint of instrumental rationality we have encountered.[1] If the debate was conducted under the supposition that people's motivations are purely self-seeking we might reassess some of the results while skipping the maximizing apparatus of the theory of choice. What Gauthier has perhaps shown is that purely prudentially motivated persons cannot cooperate successfully and that people who are disposed to be trustworthy can.

In identifying purely prudential motivation as a kind of smallest common denominator of people's possible choices, we don't say that all people are self-seeking but ask whether people with such minimally conceived ends could be party to mutually beneficial cooperation. The problems of the two tier structure of dispositions

[1] The confusion continues into other parts of Gauthier's book. In chapter one "Reason and Value" Gauthier locates the normative element in standard economic theory correctly as interpretative constraint "expressed by the single injunction, 'Maximize!' To say that one should maximize utility as a measure of preference adds nothing, since utility is simply identified with whatever one's behaviour may be interpreted as maximizing." (27) While in the introduction economic theory is described as part of normative inquiry itself. "... the role of economics in formulating and evaluating policy alternatives should leave us in no doubt about the deeply prescriptive and critical character of the science." (3)

and single decisions besetting maximizing explanations of reciprocity reappear now in much the same guise for normative expectations: People see a gap between their preferences and choices. They ask themselves which choice they ought to make, and again they may find they would do best (given Hobbesian motivations) to pretend to be trustworthy and break rank whenever they can do so at reasonable cost. The dominant strategy is only to appear to have a certain disposition. This would undermine norms of reciprocity as prudential norms.

Under these gloomy motivational assumptions, it remains strange that in daily life so many cases of reciprocity are available. We cooperate successfully in many ways and even keep promises. So people may not be only self-seeking after all. But is there a way to show that we are *normatively compelled* to enter mutually beneficial cooperations? And does instrumental thinking over minimally conceived ends support one unique set of norms? Maybe not. The most we might feel able to do is point out more precisely which motivations we would have to have in order to overcome the contractualistic dilemma. This seems to me a worthwhile task, the more since the justification of reciprocity is not the only challenge to contractualism. It may be theoretically the most fundamental one (since it targets the possibility of rational agreement itself) but practically speaking other consequences of instrumental thinking based on a Hobbesian psychology are much more dramatic. I shall briefly sketch two of the central scenarios: asymmetric and exclusive agreements.

4.3.3 Problems: Asymmetry

Suppose for now that the problem of reciprocity found a solution of some kind, be it political, Gauthierian or otherwise.[1] Two or more parties find trusting possible and

[1] Of course a solution might look different for two-person and many-person dilemmas, for one-off and repeated "multiple" situations. One is to make it physically impossible for the collaborators of a cooperative venture to defect (like Ulysses let himself chain to the mast before facing the sirens; cf. Jon Elster, 1979, *Ulysses and the Sirens*). A similar course is recommended by Hobbes himself where the "prisoners" of the state of nature agree to install an absolute sovereign whose prerogative of punishment increases the costs of defecting significantly (Hobbes, 1991 (1651), esp. chs. 13 - 17).

agree to enter mutually beneficial cooperation. Are they rationally bound to come up with one unique set of norms? And will it be something like a "just" normative system? The intuitive answer to both questions is "no". What is prudentially rational to agree between parties with purely Hobbesian motivations depends on the various strengths of the participants. An extreme case is considered by Gauthier as the Master-Slave-Parable (190 ff.). In this story, masters and slaves are naturally distinguished by their power. Once upon a time, the masters, by physical or intelligent means, were more successful in their predatory activity. For the weak underclass therefore it became advantageous to enter the master's society as slaves, be it from fear of their lives or other expected benefits. On this description, the asymmetric norms of a master-slave-society must count as justified, as long as the slaves' benefits outweigh certain costs – and what could on a Hobbesian psychology be more costly than one's lives? Now Gauthier rejects this suggestion:

> The masters employ coercion to keep the slaves obedient. Coercion is costly to both. Masters and slaves would both benefit were coercion removed and the slaves continued to serve voluntarily. But ex-slaves would not comply with an agreement to this effect. The slaves provide their services because the costs of their resistance exceed the benefits it would afford them, given their masters' power. But only the maintenance of this power rationally induces them to continue their services. Without coercion, ex-slaves might accept and adhere to some form of co-operation, but not one based on the outcome of coercive interaction. (195)

These political solutions have the disadvantage of being inflexible ("each promise needs a witness") as well as costly since part of the mutual benefit of the cooperative scheme is used up by the administration of the regulating authority. More flexible options include attempts to show directly that compliance is prudentially rational. We discussed Gauthier's suggestion. Already in the classic "Luce & Raiffa" the possibility of so-called tit-for-tat solutions to repeated Prisoner's Dilemmas is mentioned (1985 (1957), 101-2). The idea is that if two non-cooperative players do not know how often they will find themselves in similar circumstances, facing each other, they might do best to shadow each others moves, establishing finally a kind of collusion. If successful, the tit-for-tat strategy would leave the axioms of decision theory intact. Other attempts in this vain include the notion of expectation of match. Since the players are in relevant respects similar, A might expect B to act like he does, thus ultimately recommending keeping mum. (Cf. Nozick, 1969, "Newcomb's problem and two principles of choice" and Lewis, 1979, "Prisoner's Dilemma is a Newcomb Problem".) Apart from problems how precisely to formalize "tit-for-tat" equilibria, common practical objections are that, as prudential solutions, "tit-for-tat" and "expectation of match" are non-moral and as such potentially instable. Whenever there is a low-cost chance to defect, a purely prudentially motivated party will abort cooperation. Gauthier believes that his notion of *constrained* maximization or trustworthiness preserves an irreducibly moral element though it is itself subject to a prudential interpretation: if we become moral in Gauthier's sense we do better even in prudential terms.

Gauthier's argument presents itself two-fold. On the one hand, the benefits of master-slave-societies are sub-optimal (and therefore not instrumentally rational even to the masters) since coercion is costly; on the other hand, it seems implied that under non-coercive conditions, less than fair agreements cannot command stable acceptance (cf. 230). This notion of fairness is in danger of importing moral premises into the conditions of rational agreement. Why should the less empowered not voluntarily accept asymmetrical arrangements, thus not requiring excessive costs of coercion? To be sure, extremely unjust social situations lend themselves to social unrest. They may not be in the interest of the masters. But there seems little support for the idea that only agreements under the Lockean proviso *that nobody took advantage before negotiations began* (e.g. 192, and 200 ff.) command rational compliance. The fact that an agreement may be considered unfair need not make mutual compliance less advantageous. Gauthier admits as much for technological asymmetry among negotiating parties:

> A superior technology enables its possessors rationally to maintain, and requires others rationally to acquiesce in, arrangements that rest on differential rights in clear violation of the proviso. (231)

As an example Gauthier mentions the small number of Spaniards who dominated the Indian civilisations of the Americas thanks to their guns. It is hard not to see this as a typical event in the history of rational compliance.

To sum up this sketchy excursion into the problems of asymmetrical agreements it seems likely that instrumental thinking from minimally conceived ends (like survival and individual well-being) does not render one unique system of norms, nor need it cover core normative beliefs like procedural or substantive fairness. Again we may ask what kind of motivations would provide a remedy? Trustworthiness was a prominent candidate to solve the problems of reciprocity but it will do little work on asymmetry, the opposite! Before we study other possibilities let us turn to the third central problem of contractualism.

4.3.4 Problems: Limits of Scope

Morals by agreement may exclude beings who bring nothing to the bargaining table. The old, the weak, children, the handicapped, future generations and animals will not be able to bring their ends to bear since they do not pose a threat nor can they offer *mutual* advantages. For Gauthier these groups fall simply "beyond the pale of a morality tied to mutuality" (268; for future generations, see 298 ff.).

There have been attempts not to leave contractualism with this unsatisfactory conclusion. One natural reply is to consider some of the apparently excluded classes as "ones that individuals move into and out of" (Mackie, 1977, 193). A further suggestion is to extend an individual's interest to include one's children, friends, neighbours, pet-animals etc. on whose behalf one may negotiate.[1] This would make some of the contractualistic conclusions more palatable but does not solve the central difficulty of the limited scope of mutuality: one who has no advocate at the table will have their interests ignored.

Mackie had admitted earlier (on the problem of asymmetry) that "Rational bargaining can result in exploitation" (1977, 119). He also accepts the need for a moral psychology. The strategy is to widen the conception of one's well-being by including various not strictly self-seeking motivations. Mackie speaks of a "humane" disposition that "naturally manifests itself in hostility to and disgust at cruelty and in sympathy with pain and suffering wherever they occur" (1977, 194). Moral motivations like "disgust at cruelty" is then given a prudential justification (a move reminiscent of Gauthier's treatment of the disposition of trustworthiness): "the man who represents the extremes of injustice is psychotic, his soul is a chaos of internal strife" (1977, 191). In exercising cruelty towards permanently handicapped human beings, orphans or animals, say, our own well-being will suffer, so we have

[1] See Mackie, 1977, 170: "... for any individual a good life will be made up largely of the effective pursuit of activities that he finds worthwhile, either intrinsically, or because they are directly beneficial to others about whom he cares, or because he knows them to be instrumental in providing the means of well-being for himself and those closely connected to him."

instrumental reasons to grant rights beyond the ties of mutuality. Mackie concludes that "nearly all of us do have moral feelings and do tend to think in characteristically moral ways, and that these help to determine our real interests and well-being" (1977, 191-2).

We may ask whether Mackie's premises allow him this move that is "not quite equivalent to a contractual one" (1977, 193). Gauthier for one rejects appeals to moral dispositions as circular: "an affective capacity for morality presupposes a prior conception of morality; one cannot be moved by a sense of duty unless one antecedently believes some action to be one's duty" (328), and again: "we do not want to weaken the position we must defeat, straightforward maximization, by supposing that persons are emotionally indisposed to follow it" (188). It would seem that Mackie's official *error theory* (as developed in part one of his *Ethics)* should force him to agree with this verdict. If our moral sentiments imply an erroneous ontology they cannot support enlightened contractualism. One may therefore doubt whether Mackie has succeeded in resolving the "tension between the moral reason and the morality of self-interest, between any recommendation that we could defend as moral and the advice that anyone could be given about his own well-being" (1077, 190).

Be that as it may. Is there an alternative framework within which moral dispositions come out as non-erroneous? Let us take stock. In reflecting instrumentally whether we should keep promises (1), free slaves (2) or protect the sick (3) we find that from minimally conceived ends we are not compelled to pursue either of these courses of action. Neither norms of reciprocity (1), fairness (2) nor altruism (3) can be instrumentally justified. Still, if we were to acquire the motivation of trustworthiness, Gauthier maintains, the dilemma of reciprocity (1) would be solved. Mackie, another prominent contractualist, holds rather more tentatively that humane dispositions like "disgust at cruelty" might overcome problems (2) and (3), the threat of asymmetrical norms of restricted scope – but already moralized motivations may have no place in a contractualistic system. I suspect this to be true for a theory which exposes our fundamental moral attitudes as erroneous. But must it hold for any approach based on reciprocity, mutual benefit and instrumental thinking?

4.3.5 Remedies

I have argued that it is in the nature of instrumental justifications that they rely on certain, however minimally conceived, psychological assumptions. Insofar as we seek not to describe but to justify norms we have to decide where the reasoning starts from. Note that even Gauthier's minimalistic Hobbesian psychology is put to work under substantial assumptions. Persons with destructive preferences cannot be party to rational bargaining, Gauthier declares.[1] Contractualistic norms place no obligations on me who, in Hume's famous words, may prefer "the destruction of the whole world to the scratching of my finger" (T. 416). No axiom of instrumental rationality precludes this preference. Is it not *unreasonable* to seek self-destruction? One is tempted to agree. It may be unreasonable to most of us who foster other ends. This is not a foundational but a relative consideration. So even from the standpoint of pure (and otherwise successful, say) Gauthierian contractualism we should give up the idea of a final justification: the idea that there are certain norms that place obligations on everyone, the idea that we find in instrumental rationality a "sure grounding" (Gauthier, 1986, 17) of morals. This might clear the way for a different, more modest line of thoughts. Under what psychological assumptions are we compelled to accept some norms and not others? And then, why should we approve or disapprove of certain psychological features we have? One answer to the first question might be that somebody with destructive preferences need not "be contented with so much liberty against other men, as he would allow other men against himself" (to quote one of the contractualistic articles of faith, Hobbes, 1991 (1651), ch. 14); a response to the second question might be any of the following: "Most people have other ends than

[1] See for example Gauthier 1993 , 189: "My defence of the rationality of morality must be limited to those persons whose overarching life-plans make them welcome participants in society."

destruction", "You would not want that happening to yourself", "Arbitrary harm is universally despised", "At some point you must have felt sympathy", ... Disagreeing parties are invited to take up a second-order stance reflecting their own motivations, preferences, dispositions, attitudes etc.

In sections 4.1.2 - 4.1.4 (on among others Wiggins and Blackburn) I warned not to expect too much from reflections on natural tendencies most of us are said to share. We do not only care about the well-being of ourselves and those close to us, we do not only exert pressures towards consistency, we do not only keep many promises, we do not only sympathize with the distressed, we also laugh at misfortune (*schadenfreude*), we also seek actively revenge – pretty destructive dispositions after all. But it seems to me that we are now in a position to arbitrate between such conflicting suggestions on reasonable grounds. *That disposition wins that on the weakest motivational grounds leads to wide acceptance of a normative system.* All we have to do is to find which motivations minimally would overcome the perils of instrumental thinking, since thinking from ends to means, and from ends to other ends comes closer than either thinking from shared responses fixed by meaning (4.1) or consistency (4.2).

Consider again the three central dilemmas of a normative system derived from instrumental rationality. Most of the attention in the literature has been focused on the first, the problem of reciprocity. We saw Gauthier favouring trustworthiness as a solution. Jon Elster mentions altruism as the more efficient "because it is *not* derived from calculated self-interest" (1979, 145). Parfit talks of four possible moral solutions to Prisoner's Dilemmas [E = more egoistic acts, A = more altruistic acts]:

> We might become *trustworthy*. Each might then promise to do A on condition that the others make the same promise.
>
> We might become *reluctant to be 'free-riders'*. If each believes that many others will do A, he may then prefer to do his share.
>
> We might become *Kantians*. Each would then do only what he could rationally will everyone to do. None could rationally will that all do E. Each would therefore do A.
>
> We might become *more altruistic*. Given sufficient altruism, each would do A. (Parfit, 1984, 64)

Mackie, finally, quotes Warnock (1971) approvingly who appeals to the distinctly moral dispositions of "non-maleficence, fairness, beneficence, and non-deception" (Mackie, 1977, 114). These virtues are said to stabilize the "house of cards" (1977, 113) of Hobbesian reciprocal agreements.

Look first at non-deception or trustworthiness, Gauthier's favoured candidate. Instrumentally, I said, we have little reason to become fully trustworthy. We should not break agreements and promises too often, and certainly, we should not be seen to do so. But *if* we became trustworthy, would that not solve the problems of contractualism? It would certainly solve the dilemma of reciprocity (1) under the Gauthierian proviso of non-destructive preferences. On the other hand, such a disposition also would ensure that asymmetrical agreements are kept, thereby lessening the chance to arrive at norms of fairness (2). As to the third central difficulty, trustworthiness has little to contribute. If I find myself outside the range (3) of mutual agreements I shall care little whether they are kept or not. The motive of trustworthiness by itself, I conclude, is an insufficient remedy to the perils of instrumental thinking.

A second, surprising candidate is presented in Parfit's suggestion that *Kantian* motivations might solve the problems of reciprocity. The Kantian test requires that you only do what you rationally will everyone to do. This maps in part with the demand for moral consistency discussed in the previous section (4.2). There, moral consistency (in the second sense) was defined as the requirement that there be like judgment in like circumstances. You replace indexical terms with universalizable descriptions and then hypothetically change your position. Hare asked the driver to test if he could endorse his judgment to remove the bicycle even if he were the bicyclist. This strategy should produce results for dilemmas of reciprocity. Prisoner A, judging he ought to confess, obviously would not endorse that judgment were he prisoner B. In discussing Hare, I argued that the demand for moral consistency of this kind stands on doubtful pragmatic (or instrumental) grounds. This, however, is not at issue now. The question is: *if* we could get ourselves to become morally consistent in the Kantian sense, would that bridge the gaps in instrumental thinking? Clearly it would guarantee norms of reciprocity (1); it would induce us to solve Prisoner's Dilemmas, to keep promises etc. But how about the problems of asymmetry (2) and limited range (3)? Kantians can expect to do well on the former (2). No master would endorse his behaviour were he a slave. On the latter (3), Kantians, too, must face a cut-off point beyond which the test ceases to apply. Kantians don't ask the gardener if she could endorse her actions were she a snail. But what is the limiting case? "Fellow rational beings" sounds like a rather circular and little less worrying suggestion than Gauthier's "pale of a morality tied to mutuality" (268).

Apart from limits of scope (excluding e.g. animals and perhaps handicapped humans) there are some more technical problems (how would a world look in which everybody only does what he will that all do – going on holiday in August, say?) and finally the difficulty that Kantian motivations are not easy to acquire. Hardly anybody finds it psychologically possible to become fully consistent in a Kantian sense. So even if Kantian motivations were beneficial (and perhaps they are not) they may not be the best way to achieve widespread normative agreement.

"Reluctance to be a free-rider" is a further interesting but slightly mysterious motivation. At first sight, it seems to incorporate a kind of moderated Kantianism. Consider the case of over-fishing, a multiple-persons-dilemma of reciprocity. Everybody would do better were each to limit his catch. Individually most successful, however, is the strategy to have others limit their catch yet not do so oneself. In this dilemma, pure Kantian motivation would forbid you to pursue a course of action you could not accept were you one of the others, while the motivation of reluctance consists perhaps but in a desire not to do differently than most. It is part of the reality of the human psyche, that we often respond to such pressures, be they beneficial or not. The motivation under discussion, however, lends itself as easily to blind military obedience as to the establishing of conventions. So "reluctance to be a free-rider" must be a more substantial disposition than a tendency towards conformism in order to save instrumental thinking – perhaps something closer to Warnock's "non-maleficence"?

A non-maleficent person may be reluctant to take advantage of others and he may not pursue merely destructive goals. It is not only a reluctance to do differently than others, it is chiefly a substantial aversion to do harm for harm's sake. Thus the disposition should square with Gauthier's demand to exclude destructive motivations. Understood in this way, non-maleficence together with trustworthiness seems to be a sound remedy for (1), but unlike Kantian motivations, it cannot do the work by itself. As to the problem of asymmetry, non-maleficence may not prevent mutually beneficial but asymmetrical agreements (2), nor would it transcend limits of scope (3). The non-maleficent person may not actively do harm to beings she expects no benefit from, but she would not protect them. By itself, non-maleficence is too weak a disposition to secure either norms of reciprocity, fairness or altruism.

Perhaps we need something more active: beneficence, altruism or sympathy. Now there is more than one sense to this motivation. On one end of an imagined scale we may find random affection, on the other a general love of mankind or even all sentient creatures. The former is clearly inadequate since unreliable, the latter would seem to make norms of mutuality superfluous in the first place. If your goal is universal happiness you have achieved that rare thing: a utilitarian disposition.[1] Again, like Kantian motivation, such an attitude is not within the ordinary psychological scope of human beings. How should we describe a minimal, widespread and psychologically possible form of sympathy? Following Hume, I believe there are two elements to it. One is a natural tendency to share what we take to be the feeling of others, or as Hume puts it, "that propensity we have to sympathize with others, and to receive by communication their inclinations and sentiments, however different from, or even contrary to our own." (T. 316) The other is a

[1] Parfit (1984, 66) suggests that pure altruists "may face analogues of the Prisoner's Dilemma. It can be true that, if all rather than none do what is certain to be better for others, this will be worse for everyone".

substantial motivation to wish people well, to help where it is possible at reasonable cost. This substantial motivation is biased towards those close to us: family, friends, neighbours, compatriots... Again Hume gives this tendency a catchy label: "confin'd" (or limited) generosity (T. 494, 495, 586). Are these "sympathies" two facets of one disposition or in fact two distinct psychological capacities? It appears that under some conditions, they may come apart. You may imagine how it feels like (1) to be laughed at, and still don't sympathize (2) when it happens to an enemy. Vice versa, though, it seems difficult to feel sympathy (2) with a victim of *schadenfreude* and not comprehend (1) what he goes through. When there is substantial sympathy (2), or the wish to relieve suffering, it may be explained by a more fundamental disposition (1) which leads us to partake in the feelings of fellow beings. Sympathy as a principle of communication allows us to feel substantial sympathy and makes it psychologically possible, at least potentially, to expand the circle of beings we wish well.

Can two-fold sympathy be an effective remedy for the perils of rational agreement? Consider first a two-person-dilemma of reciprocity. Prisoners A and B reflect whether they (each in turn) ought to confess, thereby implicating the other and be let off with a reduced sentence. Beforehand, they have promised each other to keep mum, knowing that it would be to their disadvantage if both confessed. Now, sympathy as a principle of communication (1) seems to enable each to form beliefs about how the other might feel: "He doesn't like it in his cell, he wants to be let off as lightly as possible." Still in some cases we may not expect A and B to feel substantial sympathy (2) for each other. Their cooperation may have been built solely on mutual advantage; there was no room for unreturned favours. If that is indeed a correct description of their relationship, sympathy does not solve this two-person-dilemma. A and B may have the psychological capacity for substantial sympathy but not towards each other. A then knows that he does better to confess whatever B does: if B keeps mum, the better for A; if B confesses, A still does better to confess. A and B are rationally compelled, given their only narrowly sympathetic motivations, to choose a course of action that, in the end, is worse for both. The fact that a promise has been given does not change that since no motivation of honesty or trustworthiness is presupposed. (I argued above that there is no such thing as universal honesty. Pure honesty as a motivation is psychologically much less available than sympathy).[1] I conclude that norms of reciprocity may not be valid for two-person-cases where the two players consider each other as enemies and do therefore not substantially sympathize with each other though they are endowed with a capacity for sympathy.

Is this bad news for my claim that sympathy may be the disposition that on the weakest motivational ground leads to widespread agreement? I do not think so. Two enemies may not be normatively compelled on grounds of sympathy and instrumental rationality to adopt norms of reciprocity. The real practical dilemmas of mutuality, however, are multi-person-cases, and here the situation may be typically described as in the case of over-fishing.

Though it may be instrumentally rational within a Hobbesian psychology for each of the fishermen only to pretend to restrict his catch, under slightly wider motivations stable reciprocal agreement should be possible. Not living in complete isolation, each fisherman will come in contact and form bonds with others, helping out and enjoying the benefits of being helped. I do not say that fishermen are rationally compelled to form these bonds, I only say that most find it psychologically possible to do so. Most fishermen will wish their colleagues well and will not seek actively to take advantage – though some bending of rules in one's favour is to be expected. These fishermen now agree to limit their catch; they set up some not too expensive system of public control and together with a weak disposition not to take advantage at every opportunity, stocks might recover.

There is, of course, the danger that two *groups* of people may find themselves in a stand-off situation just like the two rogues in the classic Prisoner's Dilemma. They have formed no previous bonds, they do not wish each other well, they don't sympathize. This scenario is unfortunately part of political reality. What I can offer as

[1] This is the plausible psychological assumption behind the economists' claim that promises are "cheap talk". Cf. Hollis/Sugden, 1993, 17-19.

a reply in defence of sympathy is that it seems psychologically possible to sympathize with each member of a hostile group; from literature, theatre and film we know that our sympathies always depend on the perspective from which a story is narrated. I also suggest that it might be possible to organize societies in a way that interlock potentially opposing groups, so that a cut-off point is avoided beyond which no bonds can be formed. Remember again that I do not deny that people, and especially groups of people, often hate each other, seek revenge etc. I defend the more modest position that *if* we can put the psychologically available disposition of sympathy to work, we *then* may be normatively compelled to establish (fair) norms of mutuality. Therefore it is reasonable, in some non-foundational sense, to cultivate in us, our children and fellow men and women the disposition of sympathy.

Let us turn to the other two core situations where instrumental thinking led us into apparently unacceptable conclusions. May it not be instrumentally rational, even from a substantial disposition to sympathize, to accept mutually beneficial but unfair Master-Slave-Societies? After all, small fishermen face a much larger threat to their livelihood for agreeing to limit their catch than boats operating on an industrial scale. Are small fishermen not "voluntary slaves", so to speak? In my view, a small fisherman, even with a sympathetic disposition, has good reasons to break the rules where he can, so that the costs of policing the agreement increase to a level where it becomes worthwhile for all to compensate small boats for their losses. The same result may be achieved by political protests. The capacity for substantial sympathy may then facilitate negotiations of fairer settlements. Forced agreement, as in the setting up of Gauthier's Master-Slave-Society, conflicts directly with a disposition to sympathize.

Turning to the third problem of limited scope, it seems plain that sympathy operates beyond the pale of mutuality. It may do so in different ways. Some societies protect animals, others don't; most care for orphans, the elderly and handicapped; few for future generations. Initially, however, there is nothing in the nature of sympathy that determines a point beyond which it becomes psychologically impossible to function.

The proposal which emerged in the course of this last section may now appear to share some structural similarities to an account by Peter Railton, discussed and criticized in Chapter 3.4 above. Railton urged us to adopt an analysis of moral belief as answering to facts of what is instrumentally rational from a social point of view (cf. p. 57). I just suggested that we should conceive of moral beliefs (as dispositional sentiments of sanction) as being grounded on sympathetic motivations that would allow and sustain norms of mutuality. It may be a *matter of fact* that motivations that sympathetically include others in some way are liable to sustain a system of mutual norms. Again such a system should also be beneficial to society as a whole, promoting Railtonian social goods.

In Chapter 3.4 I argued that the onus must be on externalist accounts of moral belief (such as Railton's) to show how factual beliefs about social goods can be normatively compelling for individuals as we take moral beliefs to be. I might say "I apologised *because* I had offended her" where a moral belief (about giving offence) explains an action (the apology). Yet what is the force of facts of social rationality? Is it explanatory for a specific action to say "I apologised because it is instrumentally rational for a society as a whole to censure giving offence and promote the institution of apologising"? Now my analysis of moral belief as dispositional sentiments of sanction will face similar questions. Sentimentalism naturally allows for explanatory entailments such as "I apologised because I felt guilty". But in examining the causes of sentimental sanctions (e.g. of guilt) I came up with a structure that may sound externalist. The following might be an answer to why I apologised in a given circumstance: "I felt guilty; and I felt guilty because I had given offence; and giving offence is not licensed by mutual norms as derived from sympathetic motivations".

For Railton, we have reason to respect what is instrumentally rational from a social point of view because of our *contingent epistemological empathy* with good explanations. We find that the institutions of apology, promise, and so on are instrumentally rational for a society to have – they make a society flourish. This explains why apologising for giving offence, the keeping of promises etc. are established practices, and finally, Railton hopes, such facts from "social history and historical biology" would explain a more equal distribution of resources when it occurs. Why was Railton's response to the externalist problem deemed to be unsuccessful?

First, one may doubt how good an explanation Railton offers. I would suggest that unjust societies with an unequal distribution of resources can flourish. It also doesn't seem to speak against the objective interests of a society to deny rights altogether to certain groups of beings such as the elderly, the disabled and animals. (A roughly just society, however, could be explained by sympathetic dispositions of its constituting individuals. This I sketched in sections 4.3.2 - 4.3.5 above.)

Secondly, Railton's conception of facts explaining the social good presupposes itself a conception of norms, that is, patterns of intrinsically motivating states of mind among the constituents of a society. The institution of promising, for example, relies on the conceptually motivating power of sincere utterances of promise. This leaves Railton's distinction wanting between observing that something is of value (as a superior explanation of social goods with a contingent epistemological commendation) and valuing something (as intrinsically motivating "things to do"). That there is such a distinction is the burden of the externalist thesis, while my internalist project seeks to diminish this contrast. One might say, I state psychological conditions under which there would be a deliberative route from valuing something to seeing something of value. Sound deliberation here makes for normative recommendation. Let me recapitulate this in more detail.

4.4 Résumé

At the end of Chapter 3 we found that a conception of sentiments of sanction captures flexibly our internalistic intuitions about moral belief: Believing that I ought to ϕ is conceptually linked to my ϕ-ing. If I don't do what I sincerely believe I ought to do we need a counterfactual explanation why I didn't. This explanation includes that failure to ϕ results in the experience of sentimental sanctions such as guilt. The belief that I ought to ϕ thus expresses a disposition to experience these sentiments.

An Internalism of this form is logically prior to the investigations of Chapter 4 and may be seen as constituting an analysis of valuing. In this chapter I asked how it can be that we think of the sentiments of sanctions as appropriate or misguided. I assessed if and how our sentiments to any given act, character or situation might converge – where qualified reasonable convergence mark sentimental responses as appropriate. This then also constitutes an analysis of observing that something is of value.

How do instances of valuing become instances of value? I concentrated on the notion of a sentimental cause that might license feelings of sanction. The most straightforward idea is that these sentiments respond to a kind of reality. I am justified in feeling guilty, say, if I *really* did something wrong. In section 4.1 I rejected this suggestion. We cannot extract a sufficient idea of objectivity or truth neither from so-called response-dependent or thick concepts, nor from Blackburnian trees of attitudes. Next I turned to Hare and the suggestion that demands of consistency among our attitudes might compel us to reject certain sentiments and license others (4.2). I found appeals to consistency pragmatically doubtful and in any case insufficient to single out one unique set of attitudes. Finally I discussed the contractualistic tradition (4.3). It seeks to extract not a true or objective (in a traditional reading of these concepts) but a uniquely compelling system of norms from instrumental thinking upon minimally conceived ends. I concluded that without substantial psychological assumption there will not be convergence on one qualified set of attitudes, but at the same time, I saw some hope in a natural disposition most humans share: sympathy. Still, my Sentimentalism is in no way equivalent to sympathy ethics, since it stresses the importance of mutuality above any natural disposition. How securely we may ground norms of mutuality on sympathy may finally be a question of empirical inquiry.

CONCLUSIONS

Let us bring together the more important results of a conception of moral beliefs as dispositional sentiments of sanction.

The cognitive analysis of sentiments developed in Chapter 1.2 opens a two-tier-structure for the explanation and justification of normative demands. To the question "why ought I to ϕ?" an answer might be roughly: because I would experience a sentimental sanction, e.g. of guilt, if I didn't. To the follow-up question "but why should I feel guilty?", our cognitive sentimental analysis identified a so-called sentimental cause implied by that feeling.[1] To begin with, this is best understood in analogy to simpler cases such as fear or jealousy. Here an answer to the question "why should I feel afraid of X?" is naturally "because X is dangerous"; equally an answer to "why should I feel jealous of X" might be "because X is a rival to my rightful claims".

In the case of fear, the sentimental cause of danger seems to mark the end of the chain of normative reasoning. Either X is dangerous or it is not. As always, there might be vagueness in application, although for us, as competent users of the concept of fear, there can be no intelligent *normative* disagreement. We, analytically, ought to avoid what threatens to harm us.

For jealousy, the situation is already more complex. Again, there might be little sustainable disagreement about whether X is a rival or not. But what about that "rightful claim" that seems to be part of the concept of jealousy? Analytically, I cannot be jealous about something I have no claim to. I might be *envious* of X in virtue of his beautiful house but I cannot be *jealous* of X in that respect. Between partners, for example, is jealousy based on religious blessing, a contract of law, consent, a convention? Depending on the respective answers, jealousy may or may not be appropriate in a given situation. Here remains ample room for fundamental

[1] 'Ought' and 'should' are here used as general terms of normative recommendation. It is not supposed that the former prescribes from a moral point of view while the latter endorses a prudential motivation.

normative disagreement. Even more so with the narrowly moral sentiments which occupied us for most of this dissertation. Why should I feel specific sanctions in given circumstances? In Chapter 4, three answers were rejected:

(1) The cause of a sentimental sanction (of guilt, say) is not answering to moral reality, objectivity or substantive truth, be it of a "secondary quality" kind, specified by "thick" evaluative concepts or revealing basic moral intuitions (e.g. Blackburn's "ethic of niceness"). (Ch. 4.1)

(2) The cause of a sentimental sanction (of guilt, say) is not determined by constraints of moral consistency. (Ch. 4.2)

(3) The cause of a sentimental sanction (of guilt, say) cannot be fully characterized as instrumental constraints of mutuality. (Ch. 4.3)

Finally, a modification of (3) was suggested. If we extended the motivational basis of mutuality such as to include sympathetic dispositions of some kind, a widely acceptable answer to the sentimental cause of guilt might become available.

We can now make the two-tier structure of moral belief explicit. What is the content of the judgment that I ought to ϕ?

(4) First, "I ought to ϕ" entails that I would feel guilty, say, were I not to ϕ (this is the requirement of Internalism) where my guilt must be responsive to modes of critical reflection (this is a condition of propriety: I can't feel guilty about just anything since guilt is a cognitive sentiment).

(5) Secondly, what makes my guilt appropriate are certain convictions as sentimental cause. Critical reflection here favours a modification of (3): sympathetic motivations would allow and sustain norms of mutuality.

The *formal* sentiment of sanction is in part supported by a *substantive* sentimental disposition to sympathize. This constitutes not a foundational consideration; rather a recognition. It is merely recognised that *if* I had the disposition to sympathize it would license specific sanctions of guilt.

Two powerful arguments might be brought to bear against this now emerging structure of Sentimentalism. One is the *argument from redundancy*, the other the *argument from the psychologically impoverished.* The *argument from redundancy* is very simple; it claims "tier-one" of the proposed analysis of moral belief to be redundant. To explain moral belief with reference to sentimental sanctions (tier-one) explains nothing since sentimental sanctions themselves (on the cognitive analysis of Ch. 1.4) presuppose a conception of what ought to be done, the morally right and good (tier-two). It is never an acceptable explanation for ϕ-ing that "I would feel guilty if I were not to ϕ". Normative discussion always directly addresses what has been called sentimental causes.

According to this methodological hierarchy, only Chapter 4 of this dissertation is really central to moral theory, the status and justification of moral claims. The natural place for an account of the sentiments of sanction would be in an appendix on moral psychology, of how we should manipulate our psychological capacities to achieve the (independently understood) morally best outcome.

What then are our reasons for subordinating moral theory to moral psychology, as we have been insisting? Crucially, our approach depends on a distinction between practical and non-practical states of minds. Chapter 3 defends an internalistic conception of moral belief in detail. Taking over some results of Chapter 1, we arrived at a clear definition of what it is for a mental state to be practical:

(6) A mental state (S) is practical iff in the absence of a behavioural manifestation (B) of S we would need a counter-factual explanation why B did not occur.

In the particular case of moral belief, the definition takes the following shape (Ch. 3.5):

> (7) A moral belief (M) is practical in that the course of action (A) recommended by M is pursued
>
> (i) under normal conditions [not being grief-stricken/depressed/hypnotized/ ...]
>
> (ii) in the absence of stronger counter-vailing motivations [of prudence/etiquette/aesthetics/ ...], and
>
> (iii) failure to pursue A will result in feelings of sanction, either of self-censure [compunction/guilt/shame/remorse/ ...] or of blame [outrage/resentment/indignation/ ...]

This introduces a certain, I believe virtuous, circularity. The practicality of moral belief (which is in question) can only be defended with reference to sentiments of sanction. While the analysis of sentiments offers a functioning model of practicality (Ch. 1.2), this model appears to possess natural explanatory value only in so far as moral beliefs can be shown to be intrinsically practical. *If* we need an account of the practicality of moral belief, the moral sentiments are apt to provide one. Yet, we cannot say how moral beliefs are intrinsically practical without introducing the idea of sentimental sanctions.

It is not uncommon in philosophy that two sets of concepts turn out to be interdependent in this way. All we need to do in order to reject the argument from redundancy is to show the overwhelming implausibility of a picture of the mind excluding intrinsically practical states. In Chapter 3.4.3 ("Norms and States of Mind") I hope to have done enough in this vein:

(8) Any successful mutually coordinated activity relies on motivational states becoming attached *in a systematic way* to linguistic utterances (and the mental states they express).

It is more than a coincidence that, if we arrange to meet for a coffee tomorrow afternoon, we often do so successfully. Admittedly, we sometimes fail. But the definition given in (7) is precisely designed to allow this flexibility while preserving a systematic tie between some mental states and action. If under roughly normal conditions sincere coordinated activity fails sentimental sanctions appear. Moral sentiments are the regulators of mutuality.

The *argument from the psychologically impoverished* accepts this result. It only turns against the conception of moral beliefs as *dispositional* sentimental states. If moral judgments were to express dispositions to experience sentimental sanctions anybody who had lost (or never possessed) this psychological capacity could not make moral judgments. But clearly people make moral judgments who do not experience moral sentiments. Therefore moral judgment must express something other than dispositional sentiments. This is a neat argument. One way to avoid it is to maintain that moral judgments are about moral sentiments. More precisely, moral judgments would express that certain sentimental sanctions would make sense.[1] On this account, the psychologically impoverished judge that they should experience a moral sentiment though they do not. This presupposes a self-contained conception of the moral sentiments – a non-cognitive reading. The moral sentiment takes place in a mental sphere untouched by intentional states, beliefs, convictions. Just as we may describe the fear of a mouse fleeing the shadow of a bird of prey as a non-cognitive reaction we should be able to identify human sentimental sanctions without ascribing intentional states, e.g. the belief that a wrong had been done. In Chapter 2.1 I argue

[1] The argument from the psychologically impoverished I believe is due to Allan Gibbard. (Darwall/Gibbard/Railton, 1992, 149). Gibbard, 1990, develops the answer I sketch here.

that this non-cognitive conception fails for the moral sentiments: Excuses exculpate. We do not blame the blind for not seeing us. Apologies placate. Amends end hostility.

(9) If our view of a moral situation changes the moral sentiments change too.

It is typical of the moral sentiments that they fuse cognitive and non-cognitive mental elements in some way. If the view that a wrong had been done cannot be upheld the sentimental sanctions correct themselves, though sometimes reluctantly.

My answer to the *argument from the psychologically impoverished* is that moral beliefs indeed are tied to a (sentimentally cognitive, but epistemologically non-cognitive) capacity to experience and exert sanctions. I therefore accept that people who lack this capacity cannot entertain fully fledged moral beliefs. We cannot meaningfully call something morally right or wrong beyond the pale of mutual sanctions.

APPENDIX: A READING OF HUME

(a) Sceptic and cognitivist interpretations of Hume's moral philosophy

I am still owing a more explicit justification of the sub-title of this dissertation. Why is my version of Sentimentalism a *Humean* analysis of moral belief? As is well known, Hume held that

> The vice entirely escapes you, as long as you consider the object. You never can find it, till you turn your reflexion into your own breast, and find a sentiment of disapprobation, which arises in you, towards this action. (T. III i 1, 468-9)

And again;

> All morality depends upon our sentiments, and when any action, or quality of the mind, pleases us *after a certain manner*, we say it is virtuous; and when the neglect, or non-performance of it, displeases us *after a like manner*, we say that we lie under an obligation to perform it. (T. III ii 5, 517)

Two extreme interpretations have been given to these quotes: the one traditional and sceptical, the other revisionist and epistemologically cognitivist. On the former view, there is literally no normative content to a moral judgment. A moral judgment does not express a state of affairs or a relation between moral concepts – or at least not only. What makes a moral judgment *normative* is some psychological condition, our response to a given circumstance, character or action. In judging morally, we evince our sentiments of approval or disapproval, endorse perhaps this our reaction, are motivated to take certain actions and invite others to take a similar stance.[1]

The sceptical implications of this view are evident. If it depends on each individual's psychological condition whether something counts as right or good there is no standard against which it could be measured. The moral judgment cannot be substantially true. Furthermore, if the special normativity of the subjective condition is grounded in the fact that it possesses no normative *content*, the moral judgment

[1] This reading is close to A.J. Ayer's Hume inspired emotivism (cf. Ayer, 1946 (1936) ch. 6). For a later overview, see also Ayer, 1980.

cannot even be minimally true – say, as a judgment *about* the speaker's subjective condition or the values of the community she lives in.

This reading is said to be supported by two other familiar passages from the *Treatise*: *ought* propositions express a "new relation" that is often mistakenly but imperceptibly derived from *is* propositions (T. III i 1, 469) – a comment moulded into the slogan "No Ought from an Is!"; then a second passage from II iii 3, a specific view of motivating mental states. There, passions, emotions or sentiments are described as "original existences" lacking intentional content:

> When I am angry, I am actually possest with the passion, and in that emotion have no more a reference to any other object, than when I am thirsty, or thick, or more than five foot high. (T. II iii 3, 415)

On this account, it ultimately does not make sense to ask whether any motivating mental state is appropriate. It either happens to me that I am "possest" with a passion (as when I am more than five foot tall) or it doesn't: "... 'tis impossible, that reason and passion can ever oppose each other, or dispute for the government of the will and actions" (T. II iii 3, 416); "Reason is, and ought only to be the slave of the passions" (T. II iii 3, 415). In Book III i 1 ("Moral Distinctions not deriv'd from Reason"), Hume draws sweeping conclusions form this premise:

> Morals excite passions, and produce or prevent actions. Reason of itself is utterly impotent in this particular. The rules of morality, therefore, are not conclusions of our reason. (T. III i 1, 457)

This reading of Hume has the disadvantage of being philosophically inadequate, both in the conception of sentiments, emotions or passions it appears to presuppose as well as in the resultant analysis of moral belief. If moral sentiments just spring up, uncontrollably, there is little point to normative discussion or even reasonable solution of conflict. But clearly and to our benefit (a benefit that Hume is always keen to emphasize), acceptable arrangements are sometimes found. So was Hume really a moral sceptic?

A radically different interpretation of Hume's sentimentalism has more recently emerged. It denies that Hume is committed to treating all subjective responses to a given circumstance, action or character as being on an equal footing.[1] The main sources here are the *Second Enquiry* and the 1757 essay "Of the Standard of Taste" where Hume denounces the view he once appeared to hold, almost quoting himself:

> There is a species of philosophy, which cuts off all hopes of success in such an attempt [of reconciling the sentiments of men], and represents the impossibility of ever attaining any standard of taste. The difference, it is said, is very wide between judgment and sentiment. All sentiment is right; because sentiment has a reference to nothing beyond itself, and is always real, wherever a man is conscious of it. (Of the Standard of Taste, 229-30)

Hume then goes on to show how sentiments can be false by identifying conditions under which our aesthetic responses are to converge on a standard. Roughly, there are some basic values all men agree on; they have been "universally found to please in all countries and ages" (ST, 231) – or as Hume puts it in the conclusion of the *Second Enquiry*:

> The notion of morals implies some sentiment common to all mankind, which recommends the same object to general approbation, and makes every man, or most men, agree in the same opinion or decision concerning it. (E. 272)

In aesthetics, "Every voice is united in applauding elegance, propriety, simplicity, spirit in writing; and in blaming fustian, affectation, coldness, and false brilliancy" (ST, 227); in morals, we expect to concur in "the epithets of *vicious* or *odious* or *depraved*" (E. 272), in the repugnance of "tyrannical, insolent, or barbarous behaviour" (E. 273), and in the condemnations of "Celibacy, fasting, penance, mortification, self-denial, humility, silence, solitude, and the whole train of monkish virtues" (E. 270). These are qualities "whose tendency is pernicious to society" (E. 272), lacking "utility".[2] On the other hand there is considerable agreement on what we approve of.

[1] A revisionist interpretation of Hume's moral philosophy is most prominently defended by David Wiggins (e.g. 1987, ch. 5; 1991; 1992); related tendencies can be found in Norton (1982), Blackburn (1984, ch. 6) and A. Baier (1991).

[2] Even this revisionist reading does not make Hume a utilitarian prescribing that one ought to do whatever maximizes utility. It is not the correct calculation of utility that makes a disposition or action virtuous but Humean virtues happen to contribute to the public good; they fulfil a social function. A good discussion of the utilitarian aspects of Hume's theory may be found in Mackie (1980, 151-5).

> Besides *discretion*, *caution*, *enterprise*, *industry*, *assiduity*, *frugality*, *economy*, *good-sense*, *prudence*, *discernment*; besides these endowments, I say, whose very names force an avowal of their merit, there are many others, to which the most determined scepticism cannot for a moment refuse the tribute of praise and approbation. *Temperance*, *sobriety*, *patience*, *constancy*, *perseverance*, *forethought*, *considerateness*, *secrecy*, *order*, *insinuation*, *address*, *presence of mind*, *quickness of conception*, *facility of expression*; these, and a thousand more of the same kind, no man will ever deny to be excellencies and perfections. (E. 242-3)

If aesthetic agreement is wanting "we must choose with care a proper time and place, and bring the fancy to a suitable situation and disposition. A perfect serenity of mind, a recollection of thought, a due attention to the object" (ST, 232). Among the more specific conditions for a standard-setting aesthetic judge are his ability (or *delicacy of imagination*) to "perceive any ingredient in the composition" (ST, 235), years of *practice* (ST, 237), *freedom from prejudice* (ST, 239) and finally *good sense* to see the purpose to which a piece of art is calculated (ST, 240).

Similarly, a sound moral judge will appreciate "all the circumstances" (E. 290) of a case laid before him and "depart from his private and particular situation, and ... choose a point of view, common to him with others" (E. 272).[1]

Though there is some good textual evidence (parts of which I presented) this reading, too, ascribes to Hume a deeply problematic position. The standard of taste or virtue is defined with reference to the competent judge, while the competent judge again is qualified by his sound judgment. Thus there may be no way to arbitrate between two conflicting judgments, a prerequisite of any plausible cognitivist position. (Not surprisingly, this difficulty may remind us of the discussion of Wiggins and Blackburn in Chapter 4.1 above). One may attempt to break the circle by taking Hume's claim about basic values all humans share literally as an empirical claim. The procedure of moral arbitration would then proceed in the following curious way: First, there would be a check whether any self-proclaimed competent judge does not physically suffer from "some apparent defect or imperfection in the organ" (ST, 233). He may be subjected to paradigm cases (say, wanton cruelty), just as "the appearance

[1] This seems already close to Adam Smith's notion of the "impartial spectator" as qualified moral judge (cf. Smith, 1790 (1759)).

of objects in day-light, to the eye of a man in health, is denominated their true and real colour, even while colour is allowed to be merely a phantasm of the senses" (ST, 234). Once it is confirmed that the supposed judge is morally sane (say, responding with outrage, not sadistic pleasure) all that needs to be done is to determine the moral response data of mankind and see if they map with the judge's sentiments:

> The hypothesis which we embrace is plain. It maintains that morality is determined by sentiment. It defines virtue to be *whatever mental action or quality gives to a spectator the pleasing sentiment of approbation*; and vice the contrary. We then proceed to examine a plain matter of fact, to wit, what actions have this influence. (E. 289)

This would provide an external since empirical standard of correctness but an unsatisfactory one. What was right and good depended on majority responses one could no longer meaningfully criticize.[1]

There is little doubt that already the Hume of the *Treatise* wished to treat morals and aesthetics alike (e.g. T. III iii 1, 576-7, 589-90; T. III iii 6, 618) and that he hinted on some of the thoughts later developed in the *Second Enquiry* and "Of the Standard of Taste".[2] But the *Treatise* as a whole, in my view, offers a incomparably more sophisticated and plausible account of moral belief that is neither sceptical nor plainly cognitivist. To be sure, Hume typically insists (as quoted at the beginning of this Appendix) that an action, or quality of mind, has to please us "*after a certain manner*" before we can call it virtuous. (Cf. also T. III i 2, 471: "We do not infer a character to be virtuous, because it pleases: But in feeling that it pleases after such a particular manner, we in effect feel that it is virtuous"; and T. III i 2, 472: "Nor is every sentiment of pleasure or pain ... of that particular kind, which makes us praise and

[1] I discuss the dangers of "majoritarianism" in Chapter 4.1.3 above. Cf. also A. Price, 1986, 221.

[2] We find some evidence in III i 2 "Moral distinctions deriv'd from a moral sense": "there never was any nation of the world, nor any single person in any nation, who was not utterly depriv'd of them [the sentiments of vice and virtue], and who never, in any instance, shew'd the least approbation or dislike of manners. These sentiments are so rooted in our constitution and temper, that without entirely confounding the human mind by disease or madness, 'tis impossible to extirpate and destroy them." (T. 474). Compare also III iii 1, the brief account "Of the origin of the natural virtues and vices": "'tis impossible we cou'd ever converse together on any reasonable terms, were each of us to consider characters and persons, only as they appear from his peculiar point of view. In order, therefore, to prevent those continual *contradictions*, and arrive at a more stable judgment of things, we fix on some *steady* and *general* points of view" (T. 581-2).

condemn.") This appears to distance Hume from the sceptical view of II iii 3 that "all sentiment is right". But is there a non-circular reading of the Humean moral sentiment, and of the conditions under which specific sentiments may be appropriate?

(b) The indirect passions

First, I believe, we should read Humean moral sentiments as instances of indirect passions. It is often overlooked that Hume spends most of Book II of the *Treatise* ("Of the Passions") in developing an account of four basic modes of evaluation which is (not only tacitly) presupposed in the argument of Book III ("Of Morals").[1] These four evaluative passions are pride, humility, love and hatred, forming in Hume's words a "square" (T. 333): If I take a favourable attitude to myself, I feel pride; if I take an unfavourable attitude to myself, I feel humility (first-personal evaluations); if I take a favourable attitude to another person, I experience love; if I take un unfavourable attitude to another person, I experience hatred (second/third-personal evaluations). Note that Hume's names for these modes of evaluation are terms of art. Evaluative "pride", Hume says (T. 297), is not that arrogant feeling or character-trait people might consider as vice but thinking of oneself highly – something closer to "self-respect" or "self-esteem". Similarly, "humility" is best translated as "shame" or "guilt", "love" perhaps as "esteem", and "hatred" as "anger" or "indignation". These passions can be structurally discriminated from other passions as being "indirect", or as Hume puts it cryptically:

> nothing can produce any of these passions without bearing it a double relation, *viz.* of ideas to the object of the passion, and of sensation to the passion itself. (T. II ii 2, 333)

How is this indirect "double relation" to be understood? The passion Hume is most explicit about is pride. Consider the example of a man being proud of his beautiful house:

[1] Árdal (1966, 111) was probably the first of the modern commentators who saw the connection between moral approval/disapproval in Book III and the indirect passions of Book II.

> Here the object of the passion is himself, and the cause is the beautiful house: Which cause again is sub-divided into two parts, *viz.* the quality, which operates upon the passion, and the subject, in which the quality inheres. The quality is the beauty, and the subject is the house, consider'd as his property or contrivance. (T. II i 2, 279)

This structure imposes the following conditions: A man can only be proud if (1) what he is proud of has some close connection to himself ("the object is himself") and if (2) the value of what he is proud of can, he believes, be independently characterized (e.g. as "beauty consider'd as his property or contrivance"). It would be impossible to be proud of "A beautiful fish in the ocean, an animal in a desart" (T. II 1 9, 303), equally the cause of one's pride must be "very discernible and obvious, and that not only to ourselves, but to others also" (T. II i 6, 292).[1]

As Hume presents the indirect passions, we can ask for any given instance of pride, love, humility or hatred whether the passion is appropriate, e.g. "Why do you feel proud of this house?" To which a Humean might answer "Because I own it, and because it is generally considered to be beautiful". This distinguishes the indirect passions from the direct passions which are "perfectly unaccountable" (T. II iii 9, 439) arising "immediately ... from pain or pleasure" (T. II i 1, 276). Here reason giving may be inappropriate. Paradigm direct passions are "hunger, lust, and a few other bodily appetites" (T. II iii 9, 439) but Hume also includes phenomena that should be at the least doubtful under my interpretations: "the desire of punishment to our enemies, and of happiness to our friends" (ibid.), "desire, aversion, grief, joy, hope, fear, despair and security" (T. II i 1, 277). (Compare my analyses of fear and grief in Chapter 1.2 above).

[1] Hume's presentation of the causes of pride is not very economical. In T. II 1 6, he identifies altogether five "limitations" which I believe boil down to the two conditions I give. Davidson (1980 (1976), 277) goes even further in interpreting what Hume "should have meant". According to Davidson, entertaining the emotion of pride is only intelligible if the causes of pride can be constructed as "judgements that logically imply the judgement that is identical with pride" (284). This requires that among the causes of pride is a judgment that is *universal* in form. Justified pride of my beautiful house implies the judgments that (i) I own a beautiful, that (ii) *all* who own a beautiful house are praiseworthy (in so far as they own beautiful houses), so (iii) I am praiseworthy (in so far as I own a beautiful house). Judgment (iii) is equivalent to the emotion of pride. I agree with G. Taylor (1985) that condition (ii) is too strong.

I shall now briefly give the key quotes from Book III that suggest that Hume indeed understood the moral sentiments as indirect passions.

> Pride and humility, love and hatred are excited, when there is any thing presented to us, that both bears a relation to the object of the passion, and produces a separate sensation related to the sensation of the passion. Now virtue and vice are attended with these circumstances. They must necessarily be plac'd either in ourselves or others, and excite either pleasure or uneasiness; and therefore must give rise to one of these four passions; which clearly distinguishes them from the pleasure and pain arising from inanimate objects, that often bear no relation to us: (T. III i 2, 473)

Or more plainly:

> Now since every quality in ourselves or others, which gives pleasure, always causes pride or love; as every one, that produces uneasiness, excites humility or hatred: It follows that these two particulars are to be consider'd as equivalent, with regard to our mental qualities, *virtue* and the power of producing love or pride, *vice* and the power of producing humility or hatred. In every case, therefore, we must judge of the one by the other; and may pronounce any *quality* of the mind virtuous, which causes love or pride; and any one vicious, which causes hatred or humility. (T. III iii 1, 575)

If the moral sentiments of approbation or blame are "nothing but a fainter and more imperceptible love or hatred" (T. III iii 5, 614) why did Hume appear to rely in III i 1 on the account of the passions as "original existences" from II iii 3 – with its sketched sceptical implications? Most plausibly we should assume that at times Hume is simply overstating his anti-rationalist case. Already in II iii 3, in the small-print, "reason" turns out to be a pretty powerful slave.

> Since a passion can never, in any sense, be call'd unreasonable, *but when* [my emphasis] founded on a false supposition, or when it chuses means insufficient for the design'd end, 'tis impossible, that reason and passion can ever oppose each other, or dispute for the government of the will and actions. (T. II iii 3, 416)

If a passion relies on false beliefs and "we perceive the falsehood of any supposition, or the insufficiency of any means our passions yield to our reason without any opposition" (ibid.). For the direct passions, this indeed may be the correct account. I desire an apple (an immediately identifiable "bodily appetite") and only then I discover that it is poisoned ("falsehood of any supposition") or beyond my reach ("insufficiency of any means"). Some other passions however – the moral sentiments belong to that "indirect" species – can only be identified with regard to their causes (suppositions or implied beliefs).

What then are the causes of the sentiments of approval and disapproval as moralized versions of the indirect passions of first-personal "pride" or "self-respect", second/third-personal "love" or "esteem", first-personal "humility" or "shame" and second/third-personal "hatred" or "indignation"?

(c) Explanation and justification

Hume offers a *genealogical* explanation of our approval of virtues classified as natural or artificial respectively. Natural virtues, in Hume words, are "Meekness, beneficence, charity, generosity, clemency, moderation, equity" (T. III iii 1, 578). Of these dispositions we approve because of a natural tendency to sympathize: "The minds of all man are similar in their feelings and operations, nor can any one be actuated by any affection, of which all others are not, in some degree susceptible" (T. III iii 1, 575-6). This defines "sympathy" as a principle of communication (cf. T. II i 11, 316). When we are confronted with (or think of) people in need we become motivated to relieve a discomfort we sympathetically acquire ourselves, and *vice versa* we take pleasure in character-traits and actions that tend to ensure the well-being of those close to us. Humans, however, are not endowed with *extensive* generosity or a "love of mankind, merely as such, independent of personal qualities, of services, or of relation to ourself" (T. III ii 1, 481), so mankind came to approve of artificial virtues which keep our self-centred inclinations in check. The representative artificial virtue is justice which divides into respect for property, promises and political allegiance. Justice is the enlightened product of "human contrivance" since (T. III ii 2, 492) "there is no one, who has not reason to fear" from the instability of possession, broken premises, or disrespect for government and law – in short, the threat of a Hobbesian state of nature. The passion of self-interest therefore "restrains" itself by entering conventions which focus on "common interest":

> When this common sense of interest is mutually express'd, and is known to both, it produces a suitable resolution and behaviour. And this may properly enough be call'd a convention or agreement betwixt us, tho' without the interposition of a promise; since the actions of each of us have a reference to

> those of the other, and are perform'd upon the supposition, that something is to be perform'd on the other part. Two men, who pull the oars of a boat, do it by an agreement or convention, tho' they have never given promises to each other. (T. III ii 2, 490)

Crucial is Hume's insistence that no prior sense of duty, no already moralized institution of promising is presupposed in the approval of justice. Each individual merely acquires a "*natural* obligation" (T. III ii 2, 498) from his contingent interests to enter conventions of justice. But Hume recognises that once a person has received the benefits of an orderly society he may lose this natural obligation:

> [Men] are at first mov'd only by a regard to interest; and this motive, on the first formation of society, is sufficiently strong and forcible. But when society has become numerous, and has encreas'd to a tribe or nation, this interest is more remote; nor do men so readily perceive, that disorder and confusion follow upon every breach of these rules, as in a more narrow and contracted society. (T. III ii 2, 499)

Natural obligation thus does not seem to solve the contractualistic dilemmas of reciprocity I discussed in Chapter 4.3. As an individual that person does best who, after entering a convention, reaps the benefit and avoids to contribute his share. Hume's way out is to broaden the motivational basis for "maintaining order". Moral obligation is more than natural obligation arising from the deliberation of enlightened self-interest because of a further natural disposition – sympathy again: "we naturally *sympathize* with others in the sentiments they entertain of us" (T. III ii 2, 499). If I defect from an agreement the party I leave behind (and any spectator) would feel resentment and indignation. I take over these sentiments by formal sympathy as a principle of communication and both parties finally develop a substantial "sympathy *with public interest* [which] *is the source of the* moral approbation" (T. III ii 2, 499-500). This is why parents "inculcate on their children, from their earliest infancy, the principles of probity, and teach them to regard the observance of those rules, by which society is maintain'd, as worthy and honourable, and their violation as base and infamous" (T. III ii 2, 500).

Now this seems a curious position. Hume does not analyse the idea of moral approval (or the meaning of moral judgments) nor does he state conditions for our rational acceptance of certain norms, rather he tells a *causal* story why we approve of some actions and character-traits as virtuous and not others. His clauses typically start with 'when' not 'if'. But do we not need to know whether parents have *good reasons* to inculcate probity in their children, not only that – when they do it – they do it on the broader motivational basis of enlightened self-interest supported by sympathy? After all, reasonable parents might equally cultivate the cunning disposition of the "sensible knave" of the *Second Enquiry* who "may think that an act of iniquity or infidelity will make a considerable addition to his fortune, without causing any considerable breach in the social union and confederacy" (E. 282).

Hume's response to this challenge is best understood within his general philosophical framework. The ambitious project of the *Treatise* is nothing less than a complete explanation of human mind and behaviour (the 18th century "moral subjects" as opposed to the "natural subjects" of the physical world less human *cognoscenti*). The sceptical thrust of some of Hume's philosophy arises because he discovers that central elements of the mind resist explanations in terms he deems acceptable. However, *if* a problematic idea (or as we might better say: judgments that make use of a problematic idea or concept) can be explained *in the right way* the explanation counts as justification. Among the ideas Hume investigates are space and time, causal necessity (or more generally: inductive inference), the external world, personal identity, free will, moral and political distinctions, aesthetic distinctions and finally religious concepts (such as miracles). Hume approaches these problematic ideas in a surprisingly uniform way (in this Hume was decidedly a systematic thinker, not an inventor of unconnected philosophical puzzles). A first question typically asks: Is a given idea either the product of our senses or of deductive reason? – these being the two apparently sound human epistemic faculties. Here Hume's answer is always negative; in fact it is this negative answer that makes a given idea problematic. A second question then arises: If we can't acquire this idea by a traditionally sound epistemic route, how did we get it? The explanations Hume offers are extremely varied. They range from complex psychological claims to an almost evolutionary anthropology. Hume's most telling own label is perhaps "natural history", as in his *The Natural History of Religion.*

It is evident that Hume seeks to derive normative conclusions from these genealogical explanations. For such ideas as have been explained in the right way are confidently reinstated as epistemically sound. For example, the successful explanation of causal necessity in Book I Part iii ("Of knowledge and probability") ends in Section 15 with "Rules by which to judge of causes and effects", rules by which the "wise men" are guided – hardly a sign of causal scepticism![1] On the other hand, there are some ideas whose epistemic status does not survive a thorough investigation of their origin. Notorious victims are some central notions of the Christian religion, such as the concept of "miracle".[2] An explanation of one specific idea may also irresolvably conflict with other explanations. This is the case with "the continued and distinct existence of body".[3] Again for other ideas, such as the idea of personal identity, Hume thinks he should be able to give a satisfying explanation but confesses finally that he has failed to do so.[4]

There is one influential charge against the Humean approach – the charge of psychologism. It goes as follows: you may well explain why we have the ideas we

[1] Among them feature the following: "where several different objects produce the same effect, it must be by means of some quality, which we discover to be common amongst them", or conversely: "The difference in the effects of two resembling objects proceed from the particular in which they differ" (T. 174).

[2] Hume deleted the section "Of miracles" at a late stage from the draft of the *Treatise* in order not to prejudice the reception of the book. It would have been tellingly placed between Sections 13 and 14 of Book I Part iii, just before the successful explanation "Of the idea of necessary connexion". A later reworking of this piece can be found in the *First Enquiry.*

[3] Belief in causal inferences is saved by the postulation of external objects (As A. Baier puts it: "Hearing the door opening, without seeing its movement, does not count as breach of regularity"; 1991, 6). Yet causal thinking discovers that both primary and secondary qualities are mere perceptions, not representing a continued and distinct external world.

[4] There is considerable scholarly disagreement why precisely Hume retracts in the Appendix from the explanation he offered in I iv 6. ("all my hopes vanish, when I come to explain the principles, that unite our successive perceptions in our thought and consciousness", T. App. 635-6). It is however evident that Hume thinks he needs a successful explanation, for in Book II ("Of the passions") he unashamedly makes use of the notion of "self" and states: "the idea of ourselves is always intimately present to us" (T. II ii 4, 354).

have and make the judgments we do, but the explanations must take necessarily the same form for acceptable and unacceptable ideas, for true as for false judgments. Empirical explanations therefore can never show that some judgments are better than others; they miss the central epistemological question.[1]

How then does Hume as the causal theorist of human nature earn the right to draw normative conclusions? There is certainly a *form of explanation* that is immune to the charge of psychologism because it explains the charge itself on empirical terms. Most clearly this can be demonstrated for Hume's treatment of causal necessity. Hume tells a causal story about causality: Whenever (R) a constant conjunction between two events X and Y has been observed (and the regularity has been tested by Hume's rules of the wise men, say, conforming examples have been found but no decisive counter-

[1] Kant may be seen as offering such an argument for moral judgments: "*Empirische Prinzipien* taugen überall nicht dazu, um moralische Gesetze darauf zu gründen. Denn die Allgemeinheit, mit der sie für alle vernünftigen Wesen ohne Unterschied gelten sollen, die unbedingte praktische Notwendigkeit, die ihnen dadurch auferlegt wird, fällt weg, wenn der Grund derselben von der *besonderen Einrichtung der menschlichen Natur*, oder den zufälligen Umständen hergenommen wird, darin sie gesetzt ist. ... [Diese Prinzipien, besonders das Prinzip der Glückseligkeit, zernichten die Tugend] "indem sie die Bewegursachen zur Tugend mit denen zum Laster in eine Klasse stellen und nur den Kalkül besser ziehen lehren, den spezifischen Unterschied beider aber ganz und gar auslöschen. [Kant adds in a footnote with reference to Hutcheson] Ich rechne das Prinzip des moralischen Gefühls zu dem der Glückseligkeit" (*Grundlegung zur Metaphysik der Sitten (Groundwork of the Metaphysic of Morals)* BA 90). ("*Empirical principles* are always unfitted to serve as a ground for moral laws. The universality with which these laws should hold for all rational beings without exception – the unconditional practical necessity which they thus impose – falls away if their basis is taken from the *special constitution of human nature* or from the accidental circumstances in which it is placed. [These principles, especially the principle of personal happiness, undermine morality] inasmuch as the motives of virtue are put in the same class as those of vice and we are instructed only to become better at calculation, the specific difference between virtue and vice being completely wiped out. [footnote with reference to Hutcheson] I class the principle of moral feeling with that of happiness.")

In recent analytic philosophy, the charge of psychologism probably originates in Frege who argues that logical validity cannot be described as the rules of inference people in fact apply because "die Gesetze des Wahrseins" ("the laws of being true") prescribe how we *ought* to think ("Der Gedanke", 1967 (1918)). Frege ("Über Sinn und Bedeutung", 1980 (1892)) also gives a second anti-psychological argument from the content of belief and judgment. It claims that psychological states are "private" while more than one person may refer to the same thought or content. A judgment's content therefore cannot be explained psychologically. This second charge may be less serious. Though private psychological introspection is sometimes taken to be the source of Hume's claims (e.g. about the principles of association), it is in fact no essential part of the Humean project. To some degree, Hume was even aware of the problems of this method: introspective "reflection and premeditation would so disturb the operation of my natural principles, as must render it impossible to form any just conclusion from the phaenomenon" (T. Intro, xix). In fact, Hume claims his data are collected from the "cautious observation of human life", of "men's behaviour in company, in affairs, and in their pleasures" (ibid.).

examples) and an X is remembered or actually observed, we will come to believe (B) that a Y must have occurred or must occur.

To the objection that it is the justification of this inference that is in question the reply is that on the meta-causal level we can repeat the same move. Whenever a regularity between mental occurrences (R) and (B) is observed (and tested) we will come to believe that (R) caused (B), in other words, we will come to believe Hume's explanation.[1]

For moral distinctions, the situation is even more complicated. There is not the same inevitability in calling treacherous acts vicious as in calling tested regularities causally necessary. In consequence, Hume explains moral distinctions as inevitable only from certain *normative* premises, that is: the motivational dispositions most people have or would want to have. The idea of treachery as vice (or the judgment that treachery is vicious) originates in enlightened self-interest supported by further dispositions to sympathize. To the challenge that it is the *justification* of judgments like "you shalt not betray" that is in question the Humean answer is: When you are not a traitor, it is because you have certain minimal dispositions from which your acceptance of the norm not to betray is inevitable; these dispositions, at the same time, set out your reasons for not being a traitor. (In Chapter 4.3 I suggested this inevitability is one of instrumental rationality, i.e. an instrumentally rational person will come to believe Hume's theory). A person who does not accept the judgment "you shalt not betray" is likely to be inexplicable i.e. irrational. In rare cases, this person may have altogether different dispositions; he may be destructive or systematically cunning. Such a person is an ethical possibility and makes Hume's theory non-cognitive. Still, for most of us there is a sound epistemic route to moral conclusions.

One last exegetical worry may remain. If we explain specific moral approvals as the result of *rational* thinking from dispositions of self-interest and sympathy (and

[1] A variation of this argument appears in Stroud, 1977, 92. For more on this and the overall structure of the *Treatise*, see Kretschmer, 1993 (bound in at the end of this dissertation as supplementary material).

therefore justify those approvals for people with precisely such a motivational basis), do we not violate another of Hume's central maxims? Hume after all was the anti-rationalist who set out to destroy the epistemic credentials of reason and famously pronounced: "we ... employ our reason only because we cannot help it" (T. Abstract, 657). How then can reason reclaim its place among the trusty mental faculties?

I mentioned earlier that Hume accepts two traditional human cognitive faculties as sound: deductive reason if it is valid and the "testimony" of the senses if it is reliable. The anti-rationalist (some may say: sceptical) thrust of Hume's philosophy derives from the discovery that our most central epistemic ideas cannot be explained from these faculties. Hume then goes on to explain those problematic ideas by postulating further mental faculties: imagination, sympathy, comparison, fancy, instinct, habit, custom, superstition, enthusiasm, and so on. Now it has to be admitted that Hume's terminological consistency in this area is unforgivably loose. Sometimes imagination is the sound faculty for making inferences on the basis of evidence ("the extending of custom and reasoning beyond the perceptions"; T. I iv 2, 198), at other times it is the cause of severe epistemic confusion – the "occult quality" that "appears in children, by the desire of beating the stones, which hurt them" (T. I iv 3, 224). Equally, "comparison" in aesthetics is a prerequisite of good judgment ("By comparison alone we fix the epithets of praise or blame, and learn how to assign the due degree of each"; ST, 238), in ethics it is related to jealousy and thus "directly contrary to sympathy" (T. III iii 2, 593).

Hume, in an enlightened moment, recognises this problem. All his faculties are defined as *functions* – that is the Humean achievement – but it is how the function is defined, not the name we give it, that might allow normative conclusions:

> I must distinguish in the imagination betwixt the principles which are permanent, irresistible, and universal; such as the customary transition from causes to effects, and from effects to causes: And the principles, which are changeable, weak, and irregular; (T. I iv 4, 225)

For the Humean faculty of "reason" we can find at least four distinct uses: (1) as the deductive faculty contrasted with experience (*"there is nothing in any object, consider'd in itself, which can afford us a reason for drawing conclusions beyond it"*;

and "*even after the observation of the frequent or constant conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience*". T. I iii 12, 139), (2) as the sound inductive faculty we share with animals but employ in a superior way (T. I iii 16, 176; Abstract, 610) – in II iii 3, this causal or inductive reason is the slave of the passions, and in III i 1 it is not the source of moral distinctions – (3) as passion of self-interest that restrains itself in the genesis of justice (T. III ii 2, 492) and finally (4) as the calm passion of moral approval: "this sense must certainly acquire new force, when reflecting on itself, it approves of those principles, from whence it is deriv'd, and finds nothing but what is great and good in its rise and origin" (T. III iii 6 "Conclusion of this book", 619)[1]

The lengthy "natural history" of our approval of the artificial and natural virtues should be seen as providing a reflective *rationale* (4) for the workings of moralized pride, love, humility and hatred – the sentiments of approval and censure. Though Hume's explanations aim to be naturalistic and empirical, they incorporate sweeping claims which nowadays would compete with a whole range of social sciences: socio-biology, anthropology, economics, psychology. Perhaps they are best seen as speculative claims; their empirical status is only potential and – as science stands – unfulfilled. What makes them good explanations is this potential and their form defined as functions.

[1] A. Baier in *A Progress of Sentiments* (1991) suggests that in the case of "reason" Hume's shift of the concept is deliberate. For "reason" as a calm passion, cf. Árdal's 1976 essay "Some Implications of the Virtue of Reasonableness in Hume's *Treatise*". Jones (1982, 6) gives the following quote from Hume's *Essays*: "What is commonly, in a popular sense, called reason, and is so much recommended in moral discourse, is nothing but a general and a calm passion, which takes a comprehensive and a distant view of its object, and actuates the will without exciting any sensible emotion."

Bibliography:

Anscombe, G.E.M. 1958. "Modern Moral Philosophy", *Philosophy* 33.

Árdal, Páll S. 1966. *Passion and Value in Hume's Treatise.* Edinburgh: Edinb. UP.

Árdal, Páll S. 1976. "Some Implications of the Virtue of Reasonableness in Hume's *Treatise*", in Livingston & King, 1976.

Aristotle (384 - 322 BC). Works. References to Bekker's pages.

Aquinas. 1267-73. *Summa Theologicae.* Vol. 19: *The Emotions* (edition 1967: London: Eyre & Spottiswoode).

Ayer, A.J. 1946 (1936). *Language, Truth and Logic* (2nd edition). London: Victor Gollancz.

Ayer, A.J. 1980. *Hume.* New York: Hill & Wang.

Baier, Annette. 1976. "Mixing Memory and Desire", *American Philosophical Quarterly* 13. (reprinted in Baier, 1985).

Baier, Annette. 1977. "The Intentionality of Intentions", *Review of Metaphysics* 30.

Baier, Annette. 1978. "Hume's Analysis of Pride", *Journal of Philosophy* 75.

Baier, Annette. 1985. Postures of the Mind: Essays on Mind and Morals. Minneapolis: University of Minnesota Press.

Baier, Annette. 1991. *A Progress of Sentiments: Reflections on Hume's Treatise.* Cambridge, Mass: Harvard UP.

Binmore, Ken. 1993. "Bargaining and Morality", in Gauthier & Sugden, 1993.

Blackburn, Simon. 1981. "Reply: Rule-Following and Moral Realism", in Holtzman & Leich, 1981.

Blackburn, Simon. 1984. *Spreading the Word.* Oxford: Oxford UP.

Blackburn, Simon. 1985. "Error and the Phenomenology of Value", in Honderich, 1985. (reprinted in Blackburn, 1993).

Blackburn, Simon. 1988. "Attitudes and Contents", *Ethics* 98. (reprinted in 1993).

Blackburn, Simon. 1992a. "Through Thick and Thin", *Arist. Soc. Supp. Vol.* 66.

Blackburn, Simon. 1992b. "Gibbard on Normative Logic", *Philosophy and Phenomenological Research* 52.

Blackburn, Simon. 1993a. *Essays in Quasi-Realism.* Oxford: OUP.

Blackburn, Simon. 1993b. "Realism, Quasi, or Queasy?", in Haldane & Wright, 1993.

Boyd, Richard. 1988. "How to Be a Moral Realist", in Sayre-McCord, 1988.

Brandt, Richard. B. 1979. *A Theory of the Good and the Right.* Oxford: Clarendon.

Brink, David O. 1989. *Moral Realism and the Foundations of Ethics.* Cambridge: CUP.

Butler, Joseph. 1749 (1726). *Fifteen Sermons* (4th edition). (excerpts in Raphael, 1969).

Campbell, R. & Sowden, L. (eds.). 1985. *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem.* Vancouver: University of British Columbia Press.

Charlton, William. 1988. *Weakness of Will.* Oxford: Blackwell.

Clarke, Samuel. 1728 (1706). *A Discourse Concerning the Unchangeable Obligation of Natural Religion, and the Truth and Certainty of the Christian Revelation* (7th edition). (excerpts in D.D. Raphael, 1969).

Cooper, John. 1993. "Aristotle's Theory of the Emotions", paper 4-5-1993, UCL.

Copp, David. 1990. "Explanation and Justification in Ethics", *Ethics* 100.

Dancy, Jonathan. 1993. *Moral Reasons.* Oxford: Blackwell.

Darwall, Stephen & Gibbard, Allan & Railton, Peter. 1992. "Toward *Fin de siècle* Ethics: Some Trends", *Philosophical Review* 101.

Davidson, Donald. 1976. "Hume's Cognitive Theory of Pride", *Journal of Philosophy* 73. (reprinted in Davidson, 1980).

Davidson, Donald. 1980. *Essays on Actions and Events.* Oxford: OUP.

De Sousa, R.B. 1980. "The Rationality of Emotions", in A.O. Rorty, 1980.

De Sousa, R.B. 1987. *The Rationality of Emotion.* Cambridge, Mass: MIT Press.

Elster, Jon. 1979. *Ulysses and the Sirens: Studies in Rationality and Irrationality.* Cambridge: Cambridge UP.

Falk, W.D. 1948. "'Ought' and Motivation", *Pro. Arist. Soc.* 48. (reprinted in Falk, 1986).

Falk, W.D. 1986. *Ought, Reasons, and Morality.* Ithaca: Cornell UP.

Foot, Philippa. 1958a. "Moral Arguments", *Mind* 67. (reprinted in Foot, 1978).

Foot, Philippa. 1958b. "Moral Beliefs", *Proc. Arist. Soc.* 59. (reprinted in Foot, 1978).

Foot, Philippa. 1961. "Goodness and Choice", *Arist. Soc. Supp. Vol.* (reprinted in Foot, 1978).

Foot, Philippa. 1972a. "Reasons for Actions and Desires", *Arist. Soc. Supp. Vol.* (reprinted in Foot, 1978).

Foot, Philippa. 1972b. "Morality as a System of Hypothetical Imperatives", Philosophical Review 81, 3. (reprinted in Foot, 1978).

Foot, Philippa. 1978. "Are Moral Considerations Overriding?", in Foot, 1978.

Foot, Philippa. 1978. *Virtues and Vices, and Other Essays.* Oxford: Blackwell.

Frankena, William K. 1958. "Obligation and Motivation in Recent Moral Philosophy", in Melden, 1958. (reprinted in Goodpaster, 1976).

Frankfurt, Harry. 1971. "Freedom of the Will and the Concept of a Person", *Journal of Philosophy* 68.

Frege, Gottlob. 1960 (1892). "On Sense and Reference" ("Über Sinn und Bedeutung"), in *Philosophical Writings: Translations*, P. Geach & M. Black. Oxford: Blackwell.

Frege, Gottlob. 1967 (1918). "The Thought: A Logical Enquiry" ("Der Gedanke. Eine logische Untersuchung"), in P.F. Strawson (ed.), *Philosophical Logic*. Oxford: OUP.

Gaita, Raimond. 1991. *Good and Evil: An Absolute Conception*. London: Macmillan.

Gardiner, P.L. 1954. "On Assenting to a Moral Principle", *Proc. Arist. Soc.* 55.

Gauthier, David. 1986. *Morals by Agreement.* Oxford: Clarendon.

Gauthier, David. 1987. "Reason to be Moral?", *Synthese* 72.

Gauthier, David. 1993. "Uniting Separate Persons", in Gauthier &Sugden, 1993.

Gauthier, David & Sugden, Robert (eds.) 1993. *Rationality, Justice and the Social Contract*. Hemel Hempstead: Harvester Wheatsheaf.

Geach, Peter. 1958. "Imperative and Deontic Logic", *Analysis* 18.

Geach, Peter. 1965. "Assertion", *Philosophical Review* 74.

Geach, Peter. 1977. *The Virtues*. Cambridge: CUP.

Gibbard, Allan. 1988. "Hare's Analysis of 'Ought' and its Implications", in Seanor & Fotion, 1988.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings*. Oxford: Oxford UP.

Gibbard, Allan. 1992. "Thick Concepts and Warrant for Feelings", *Arist. Soc. Supp. Vol.* 66.

Goodpaster, K.E. (ed.). 1976. *Perspectives on Morality: Essays by William K. Frankena*. London: University of Notre Dame Press.

Gordimer, Nadine. 1980 (1979). *Burger's Daughter*. London: Penguin.

Gordon, R.M. 1987. *The Structure of Emotions*. New York: Cambridge UP.

Gosling, Justin. 1990. *Weakness of the Will*. London & New York: Routledge.

Green, O.H. 1992. *The Emotions: A Philosophical Theory*. Doordrecht: Kluwer.

Greenspan, P.S. 1980. "A Case of Mixed Feelings: Ambivalence and the Logic of Emotion", in A.O. Rorty, 1980.

Griffiths, A.Ph. (ed.). 1992. *A.J. Ayer: Memorial Essays*. Cambridge: CUP.

Haldane, John & Wright, Crispin (eds.). 1993. *Reality, Representation, and Projection*. New York: Oxford UP.

Hare, R.M. 1952. *The Language of Morals*. Oxford: Clarendon.

Hare, R.M. 1963. *Freedom and Reason*. Oxford: Clarendon.

Hare, R.M. 1964. "The Promising Game", *Revue International de Philosophie* 70. (reprinted in Hare, 1989).

Hare, R.M. 1981. *Moral Thinking*. Oxford: Clarendon.

Hare, R.M. 1984. "Supervenience", *Proc. Arist. Soc.* 85. (reprinted in Hare, 1989).

Hare, R.M. 1985. "Ontology in Ethics", in Honderich, 1985. (reprinted in Hare, 1989).

Hare, R.M. 1986. "How to Decide Moral Questions Rationally", *Critica* 18. (reprinted in Hare, 1989).

Hare, R.M. 1988. "Comments", in Seanor & Fotion, 1988.

Hare, R.M. 1989. *Essays in Ethical Theory*. Oxford: Clarendon.

Harman, Gilbert. 1977. *The Nature of Morality*. New York: Oxford UP.

Harrison, Jonathan. 1976. *Hume's Moral Epistemology*. Oxford: Clarendon.

Heelas, John & Locke, Andrew (eds.). 1981. *Indigenous Psychologies*: The Anthropology of the Self. London: Academic Press.

Hobbes, Thomas. 1991 (1651). *Leviathan*, Cambridge: Cambridge UP.

Hollis, M. & Sugden R. 1993. "Rationality in Action", *Mind* 102.

Holtzman, S. & Leich, S. (eds.) 1981. *Wittgenstein: To Follow a Rule*. London: Routledge & Kegan Paul.

Honderich, Ted. 1975. "The Use of the Basic Proposition of a Theory of Justice", *Mind* 84.

Honderich, Ted (ed.). 1985. *Ethics and Objectivity*. London: Routledge & Keg. Paul.

Honderich, Ted. 1988. *The Consequences of Determinism (A Theory of Determinism, Vol. 2)*. Oxford: Clarendon.

Hume, David. 1978 (1739-40). *A Treatise of Human Nature*. (ed. L.A. Selby-Bigge & P.H. Nidditch). Oxford: Clarendon.

Hume, David. 1975 (1777 (1748, 1751)). *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. (ed. L.A. Selby-Bigge & P.H. Nidditch). Oxford: Clarendon.

Hume, David. 1985 (1741-77). *Essays Moral, Political, and Literary*. (ed. E.F. Miller). Indianapolis: Liberty Classics.

Hume, David. 1882-6 (1757). *The Natural History of Religion* (vol. 4 of *David Hume: The Philosophical Works*, ed. T.H. Green & T.H. Grose). London: Longman.

Hume, David. 1932. *The Letters of David Hume.* (ed. J.Y.T. Greig). Oxford: Clarendon.

Hutcheson, Francis. 1738 (1725). *An Inquiry Concerning the Original of Our Ideas of Virtue or Moral Good* (4th edition). (excerpts in Raphael, 1969).

Hutcheson, Francis. 1742 (1728). *An Essay on the Nature and Conduct of the Passions and Affections. With Illustrations on the Moral Sense* (3rd edition). (excerpts in Raphael, 1969).

James, William. 1884. "What is an Emotion?". *Mind* 9 (old series).

James, William. 1890. *The Principles of Psychology.* London: Macmillan.

Johnson, Samuel. 1927 (1759). *The History of Rasselas Prince of Abissinia.* Oxford: Clarendon.

Johnston, Mark. 1989. "Dispositional Theories of Value", *Arist. Soc. Supp. Vol.* 63.

Johnston, Mark. 1993. "*Objectivity Refigured*", in Haldane & Wright, 1993.

Jones, Peter. 1982. *Hume's Sentiments: Their Ciceronian and French Context.* Edinburgh: Edinburgh UP.

Kahneman, D. & Tversky, A. 1979. "Prospect Theory: An Analysis of Decision under Risk", *Econometrica* 47.

Kant, Immanuel. 1785. *Grundlegung zur Metaphysik der Sitten.* (trans. H.J. Paton. 1991 (1948). London: Routledge).

Kenny, Anthony. 1963. *Action, Emotion and Will.* London: Routledge & Kegan Paul.

Kölbel, Max. 1994. *The Coherence of Expressivism*, M.Phil. dissertation. University of London.

Kölbel, Max. 1995. "Expressivism and the Geach Objection", *Proceedings of Analyomen 2.* Berlin & New York: De Gruyter.

Kretschmer, Martin. 1993. "Review of A. Baier (1991) *A Progress of Sentiments* and W. Brand (1992) *Hume's Theory of Moral Judgment*", *Mind* 102.

Kretschmer, Martin. 1996. "Review of D.F. Norton (ed.), 1993 *The Cambridge Companion to Hume*", *Philosophical Books* 37.

La Rochefoucauld, Francois de. 1976 (1665). *Reflexions, ou sentences et maximes morales.* Paris: Gallimard.

Lewis, David. 1966. "An Argument for the Identity Theory", *Journal of Philosophy* 63.

Lewis, David. 1972. "Psychophysical and Theoretical Identifications", *Australasian Journal of Philosophy* 50.

Lewis, David. 1979. "Prisoner's Dilemma is a Newcomb Problem", *Philosophy and Public Affairs* 8 (reprinted in Campbell & Sowden, 1985).

Lewis, David. 1989. "Dispositional Theories of Value", *Arist. Soc. Supp. Vol.* 63.

Livingston, D.W. & King, J.T. (eds.) 1976. *Hume: A Re-evaluation*. New York: Fordham UP.

Luce, R.D. & Raiffa, H. 1985 (1957). *Games and Decisions*. New York: Dover (Wiley).

Lyons, William. 1980. *Emotion*. Cambridge: Cambridge UP.

Mackie, J.L. 1977. *Ethics: Inventing Right and Wrong*. London: Penguin.

Mackie, J.L. 1980. *Hume's Moral Theory*. London: Routledge & Kegan Paul.

Mackie, J.L. 1984. *Persons and Values* (eds. J. Mackie & P. Mackie). Oxford: Clarendon.

McDowell, John. 1978. "Are Moral Requirements Hypothetical Imperatives?", *Arist. Soc. Supp. Vol.* 52.

McDowell, John. 1979. "Virtue and Reason", *Monist* 62.

McDowell, John. 1981. "Non-Cognitivism and Rule-Following", in Holtzman & Leich, 1981.

McDowell, John. 1983. "Aesthetic Value, Objectivity and the Fabric of the World", in Schaper, 1983.

McDowell, John. 1985. "Values and Secondary Properties", in Honderich, 1985.

McDowell, John. 1987. *Projection and Truth in Ethics*. Kansas: Lindley Lectures.

McDowell, John. 1994. "Eudaimonism and Realism in Aristotle's Ethics", *Keeling-Colloquium*, paper 19/2/94, University College London.

McNaughton, David. 1988. *Moral Vision*. Oxford: Blackwell.

Melden, A.I. (ed.). 1958. *Essays in Moral Philosophy*. Seattle: University of Washington Press.

Mill, John Stuart. 1863. *Utilitarianism*. London: Parker, Son & Bourn.

Moore, G.E. 1903. *Principia Ethica*. Cambridge: CUP.

Moore, G.E. 1912. *Ethics*. Oxford: OUP.

Mortimer, Geoffrey. 1971. *Weakness of Will*. London: Macmillan & St Martin Press.

Nagel, Thomas. 1970. *The Possibility of Altruism*. Oxford: Clarendon.

Norton, D.F. 1982. *David Hume: Common-Sense Moralist, Sceptical Metaphysician*. Princeton: Princeton UP.

Norton, D.F. (ed.) 1993. *The Cambridge Companion to Hume*. Cambridge: CUP.

Nozick, Robert. 1969. "Newcomb's Problem and Two Principles of Choice", in Nicholas Rescher (ed.), *Essays in Honor of Carl G. Hempel*. Dordrecht: Reidel (abridged version reprinted in Campbell & Sowden, 1985).

Nussbaum, Martha. 1992. "Virtue Revived: Habit, passion, reflection in the Aristotelian tradition", *Times Literary Supplement*, 3-7-92.

Oakley, Justin. 1992. *Morality and the Emotions*. London: Routledge.

Pareto, V. 1972 (1927). *Manual of Political Economy*. London: Macmillan.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon.

Pettit, Philip. 1990. "The Reality of Rule-Following", *Mind* 99.

Pettit, Philip. 1991. "Realism and Response-Dependence", *Mind* 100.

Pitcher, George. 1965. "Emotion", *Mind* 74.

Plato (427 - 347 BC). Works. References to Stephanus' pages.

Platts, Mark. 1980. "Moral Reality and The End of Desire", in Platts, 1980.

Platts, Mark. (ed.). 1980. *Reference, Truth and Reality*. London: Routledge & Kegan Paul.

Platts, Mark. 1991. *Moral Realities*. London: Routledge.

Price, A.W. 1986. "Doubts about Projectivism", *Philosophy* 61.

Price, A.W. 1995. *Mental Conflict*. London: Routledge.

Price, Richard. 1787 (1758). *A Review of the Principal Questions of Morals* (3rd edition). (excerpts in Raphael, 1969).

Putnam, Hilary. 1981. *Reason, Truth and History*. Cambridge: CUP.

Railton, Peter. 1986. "Moral Realism", *Philosophical Review* 45.

Railton, Peter. 1992. "Nonfactualism about Normative Discourse", *Philosophy and Phenomenological Research* 52.

Railton, Peter. 1993. "What the Non-Cognitivist Helps Us to See the Naturalist Must Help Us to Explain", in J. Haldane & C. Wright, 1993.

Ramsey, Frank Plumpton. 1931. "Truth and Probability", in *The Foundations of Mathematics and Other Logical Essays*. London: Routledge & Kegan Paul.

Raphael, D.D. (ed.). 1969. *British Moralists 1650 - 1800: Selected Writings*. Oxford: Clarendon.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, Mass: Harvard UP.

Robertson, John. 1990. "Hume on Practical Reason", *Proc. Arist. Soc.* 90.

Rorty, A.O. 1980. "Explaining Emotions", in A.O. Rorty, 1980.

Rorty, A.O. (ed.) 1980. *Explaining Emotions*. Berkley: University of California Press.

Russell, Bertrand. 1928. *Sceptical Essays*. London: Allen & Unwin.

Sainsbury, R.M. 1988. *Paradoxes*. Cambridge: CUP.

Santas, Gerasimos. 1964. "The Socratic Paradoxes", *Philosophical Review* 73.

Savage, Leonard J. 1972 (1954). *The Foundations of Statistics*. New York: Dover (Wiley).

Sayre-McCord, G. (ed.). 1988. *Essays on Moral Realism*. Ithaca: Cornell UP.

Schank, R.L. 1932. "A Study of a Community and Its Groups and Institutions Conceived of as Behaviors of Individuals", *Psychological Monographs* 43.

Schaper, E. (ed.). 1983. *Pleasure, Preference and Value*. Cambridge: CUP.

Seanor, D. & Fotion, N. (eds.). 1988. *Hare and Critics*. Oxford: Clarendon.

Searle, John. 1962. "Meaning and Speach Acts", *Philosophical Review* 71.

Searle, John. 1964. "How to derive 'Ought' from 'Is'", *Philosophical Review* 73.

Shaftesbury, Anthony Ashley Cooper-Third Earl of. 1714 (1699). *Characteristics of Men, Manners, Opinions, Times* (2nd edition). (excerpts in Raphael, 1969).

Singer, Peter. 1973. "The Triviality of the Debate over 'Is-Ought' and the Definition of 'Moral'", *Am.Ph.Q.* 10.

Singer, Peter. 1981. *The Expanding Circle: Ethics and Sociobiology*. New York: New American Library.

Sidgwick, Henry. 1907 (1874). *The Methods of Ethics* (7th edition). London: Macmillan.

Smith, Adam. 1790 (1759). *The Theory of Moral Sentiments* (6th edition). (excerpts in Raphael, 1969).

Smith, Jean. 1981. "Self and Experience in Maori Culture", in Heelas & Locke, 1981.

Solomon, R.C. 1976. *The Passions*. Garden City, N.Y.: Doubleday & Anchor.

Stevenson, C.L. 1937. "The Emotive Theory of Ethical Terms", *Mind* 46.

Stevenson, C.L. 1944. *Ethics and Language*. New Haven: Yale UP.

Strawson, P.F. 1962. "Freedom and Resentment", *Proceedings of the British Academy*. (reprinted in Strawson, 1974).

Strawson, P.F. 1974. *Freedom and Resentment and Other Essays*. London: Methuen.

Stroud, Barry. 1977. *Hume*. London: Routledge & Kegan Paul.

Stroud, Barry, 1992. "Ayer's Hume", in Griffiths, 1992.

Sturgeon, Nicholas L. 1985. "Gibbard on Moral Judgment and Norms", *Ethics* 96.

Taylor, C.C.W. 1980. "Plato, Hare and Davidson on Akrasia", *Mind* 89.

Taylor, Gabrielle. 1985. *Pride, Shame, and Guilt.* Oxford: Clarendon.

Thalberg, I. 1977. *Perception, Emotion and Action*. Oxford: Blackwell.

Tugendhat, Ernst. 1989. "Zum Begriff und zur Begründung von Moral", in C. Bellert & N. Müller-Schöll (eds.). 1989. *Mensch und Moderne: Festschrift für Helmut Fahrenbach*. Würzburg. (reprinted in Tugendhat, 1992).

Tugendhat, Ernst. 1992. *Philosophische Aufsätze*. Frankfurt/M: Suhrkamp.

Warnock, G.J. 1971. *The Object of Morality*. London: Methuen.

Wiggins, David. 1976. "Truth, Invention, and the Meaning of Life", *Proceedings of the British Academy* 62. (reprinted in Wiggins, 1987).

Wiggins, David. 1987 (1980, 1983). "Truth, and Truth as Predicated of Moral Judgments", in Wiggins, 1987.

Wiggins, David. 1987. "A Sensible Subjectivism?", in Wiggins, 1987.

Wiggins, David. 1987. *Needs, Values, Truth*. Oxford: Blackwell.

Wiggins, David. 1990. "Moral Cognitivism, Moral Relativism and Motivating Beliefs", *Proc. Arist. Soc.* 91.

Wiggins, David. 1991. "Categorical Requirements: Kant and Hume on the Idea of Duty", *Monist* 74.

Wiggins, David. 1992. "Ayer's Ethical Theory: Emotivism or Subjectivism?", in Griffiths, 1992.

Wiggins, David. 1993. "Cognitivism, Naturalism, and Normativity: A Reply to Peter Railton", in Haldane & Wright, 1993.

Williams, Bernard. 1966. "Consistency and Realism". *Arist. Soc. Supp. Vol.* (reprinted in Williams, 1973).

Williams, Bernard. 1973. *Problems of the Self.* Cambridge: CUP.

Williams, Bernard. 1980. "Internal and External Reasons", in Harrison, R. (ed.), *Rational Action*. Cambridge: CUP. (reprinted in Williams, 1981).

Williams, Bernard. 1981. *Moral Luck*. Cambridge: CUP.

Williams, Bernard. 1985a. "Ethics and the Fabric of the World", in Honderich, 1985.

Williams, Bernard. 1985b. *Ethics and the Limits of Philosophy*. London: Fontana & Collins.

Williams, Bernard. 1993. *Shame and Necessity*. Berkley & Los Angeles: University of California Press.

Wright, Crispin. 1985. "Review of Blackburn (1984) *Spreading the Word*", *Mind* 94.

Wright, Crispin. 1987. "Realism, Antirealism, Irrealism, Quasi-Realism", *Mid West Studies in Philosophy* 12.

Wright, Crispin. 1988. "Moral Values, Projection, and Secondary Qualities", *Arist. Soc. Supp. Vol.* 62.

Wright, Crispin. 1992. *Truth and Objectivity*. Cambridge, Mass: Harvard UP.

Mind 102, 1993

task of showing that this suggestion is a reasonable one, however, will require a sustained argument of very much the same character as Aune's argument about presupposition.

Overall, Aune's book is very welcome indeed and will make an excellent framework for classroom discussions. The hardback is too expensive to assign readily, but we can hope that a reasonably priced paperback version will appear soon.

Arizona State University RICHARD CREATH
Tempe
Arizona
AZ 85287
USA

A Progress of Sentiments: Reflections on Hume's Treatise, by Annette C. Baier. Cambridge, Mass. and London: Harvard University Press, 1991. Pp. xi + 333. £31.95

Hume's Theory of Moral Judgment, by Walter Brand. Dordrecht, Boston and London: Kluwer Academic Publishers, 1992. Pp. xi + 164. £42.95.

David Hume is credited with the invention of numerous philosophical puzzles: we still write on induction, probability, and causation; we argue about personal identity and bundles of perceptions; we derive, or fail to derive, "Ought" from "Is". There is, however, little agreement on whether and how these and other puzzles fit into a systematic picture. What was David Hume up to? For most of this century the orthodox interpretation saw Hume as a kind of embryonic Logical Positivist. Hume's overall aim, it was said, was the destruction of metaphysics ("Commit it… to the flames", *First Enquiry*, Selby-Bigge edition, p. 165). Analysis was to be the new, limited task of philosophy. Typical questions attributed to Hume were of this form: "*given what characteristics of sense-impressions do we assert material-object propositions?*" (H.H. Price, *Hume's Theory of the External World*, Oxford: Clarendon Press, 1940, p. 15). The impressions or data in this picture are given. We don't know of an external world behind the veil, we can only spell out under what conditions we believe in it.

Similar reductive questions were asked concerning causal propositions, personal identity propositions, and moral propositions. In each case the outcome was essentially sceptical: No instances of causal power can be observed; we assert causal connections not of necessity but merely on the basis of regularity relations. No metaphysical self can be inspected; we don't even know the conditions under which we speak of identical persons. Moral propositions resist reductions into verifiable data; they are cognitively meaningless.

With the decline of Logical Positivism this picture became less persuasive. As we know, fashionable doctrines are more likely to be attributed to great philosophers. Still, sceptical readings have remained dominant in the Hume literature. That may change now. "Realist" interpretations of Hume are becoming influential, notably with Galen Strawson's book *The Secret Connexion: Causation, Realism, and David Hume* (Oxford: Clarendon Press, 1989). (For similar tendencies see e.g. John P. Wright, *The Sceptical Realism of David Hume*, Manchester: Manchester University Press, 1983; Edward Craig, *The Mind of God and the Works of Man*, Oxford: Clarendon Press, 1987, Ch. 2.) In interpretations of Hume's moral philosophy we may observe a related trend. D. F. Norton e.g. claims that Hume, though a metaphysical sceptic, is a moral realist (*David Hume*, Princeton, New Jersey: Princeton University Press,1982; see also David Wiggins' moral cognitivist Hume interpretations in *Needs, Values, Truth*, Oxford: Blackwell,1987, Chs. 2, 5; and in many of his essays).

The recent shift in Hume interpretation has itself generated some literature. Barry Stroud's article *Ayer's Hume* (in *A. J. Ayer: Memorial Essays*, ed. A. Ph. Griffiths, Cambridge: Cambridge University Press, 1992) gives an excellent account of the Positivists' attraction to Hume. Kenneth P. Winkler seeks to fight back against *The New Hume*, as he calls the recent wave of realist interpretations (*The Philosophical Review*, 100, No. 4, 1991, pp. 541-579).

During the present battle between the Old and the New Hume it is most welcome to receive two new books which attempt unified interpretations of Hume's philosophy as presented in the *Treatise*, and resist the temptation to put our contemporary philosophical concerns first. Initially, Baier as well as Brand might be regarded as exponents of the New Hume in that they reject the traditional sceptical picture according to which (reformulated in terms of the epistemological potence of reason) Hume supports the following claims: (1) We have no reasons to adhere to a special corpus of beliefs; in particular, we have no reasons for our predictions. (2) In the case of moral evaluations we cannot even coherently ask whether they are based on reasons; they are cognitively empty. In short, when we reason we "employ our reason only because we cannot help it" (*Abstract* of Books1 and 2 of the *Treatise* as printed in the Selby-Bigge edition of the *Treatise*, p. 657). Of course, it is hard to deny that Hume said this and other similar things but Baier and Brand argue convincingly that it is not the whole truth. They both draw partly on the same sections of the *Treatise*—e.g. the much-neglected "Rules by which to judge of causes and effects" (*Treatise*, Bk. 1, Pt. 3, Sc. 15)—which, perhaps, is not entirely coincidental, since Brand acknowledges his debts to Baier's earlier teaching at the City University of New York. Still, there is more than a difference of emphasis between the two accounts. Brand retreats in the end to a more traditional sceptical interpretation while Baier argues that Hume is willing to embrace a revised naturalized concept of reason.

First, I shall have a closer look at Brand's book. It is, rather strangely, entitled *Hume's Theory of Moral Judgment* despite its claim to be "A Study in the Unity" (p. iii) of Hume's *Treatise*. About half the volume is devoted to the role of reason in Hume's epistemology of Book 1 ("Of the Understanding"). Presumably, Brand must see it as his main achievement to have applied this account to the case of moral evaluations in Book 3 ("Of Morals").

Brand identifies two principles which are at work in Book 1's account of human belief formation. The first, imagination (1), will be familiar to most readers of Hume textbooks. Whenever we are confronted with certain relevant regularities of our experience we inevitably acquire beliefs of some kind. For example, when it has been observed that event *X* is followed regularly by event *Y* we will come to believe that *X* caused *Y*, and whenever an *X* is remembered or actually observed we will believe that a *Y* will have occurred or will occur. The possibility of event *Y* following will be felt or imagined more vividly than an alternative course of events. Hume calls this state of mind a belief. It is in this sense "*that belief is more properly an act of the sensitive, than of the cogitative part of our natures*" (*Treatise*, Selby-Bigge edition, p. 183).

One obvious charge has always been made against Hume. If belief is just a matter of being in a certain lively state of mind, there is no room to account for false beliefs. We either believe, thanks to an automatic operation of the imagination, or we don't believe when there is no sufficient regularity to activate the imagination. Hume didn't think he was committed to such a position, and Brand rightly draws our attention to a second principle: the regulative operation of the understanding (2). Hume presents us with the case of a man "who being hung out from a high tower in a cage of iron cannot forbear trembling" (*Treatise*, p.148). The man has previously experienced a constant conjunction between height and dangerous falls. He cannot help believing himself to be in danger. The man also "knows himself to be perfectly secure from falling, by his experience of the solidity of the iron, which supports him" (ibid.). The man's experience gives rise to two conflicting beliefs, both based on the principle of imagination.

As the case is presented it would be clearly wrong for the man to believe himself in danger. A piece of causal reasoning is required which would identify height as a necessary, not a sufficient feature of the previous regularity between height and dangerous falls. How is he going to do that? He should, Hume holds, revise his imaginatively acquired beliefs according to "Rules by which to judge of causes and effects". In our case Hume's rule number six might do: "The difference in the effects of two resembling objects must proceed from that particular, in which they differ" (*Treatise*, p. 174). Such rules do not identify ultimate real causes but merely take into account formerly hidden regularities. Beliefs that have been revised in this manner are potentially open to further revision.

Brand thinks, and he provides some evidence from Bk. 1, Pt. 4, Sc. 1("Of scepticism with regard to reason"), that Hume took the potentially infinite reversibility of our beliefs to be a sceptical threat. Any belief we maintain might be false, i.e. in need of further correction. Only the resilient and unreasonable imagination makes us believe at all, and saves us from total scepticism. "The solution to the problem of the justification of belief vanishes as unanswerable" (p. 64), writes Brand at the end of his chapters on Book 1. Neither the principle of imagination nor corrective reasoning should be endorsed. "Sometimes the one, sometimes the other prevails, according to the disposition and character of the person" (*Treatise*, p. 150). (Brand doesn't give much weight to the follow-up which shows little sign of scepticism: "The vulgar are commonly guided by the first [principle of imagination], and wise men by the second [principle of corrective reasoning]".)

What about moral evaluations? According to Brand's reading of Hume, moral beliefs are based on "that propensity we have to sympathize with others, and to receive by communication their inclinations and sentiments, however different from, or even contrary to our own" (*Treatise*, p. 316). Sympathy as a principle of communication results in beliefs about mental states of other people just as the principle of imagination leads to beliefs about the world. We observe regularities between expressions or words and certain patterns of behaviour, and assume that similar behaviour is caused by similar states of mind. In one of Hume's examples we see a ship

> tost by a tempest, and in danger every moment of perishing on a rock or sand-bank... Suppose the ship to be driven so near me, that I can perceive distinctly the horror, painted on the countenance of the seamen and passengers, hear their lamentable cries, see the dearest friends give their last adieu, or embrace with a resolution to perish in each others arms... . (*Treatise*, p. 594)

If we, as spectators, come to share the feelings of these persons, sympathy might explain why we feel motivated to relieve a suffering we feel ourselves. This is perfectly in line with Hume's notorious claim (in Bk. 3, Pt. 1, Sc. 1) that morality cannot be based on reason since reason alone cannot account for the felt and motivating obligation which is characteristic of morality. But can sympathy, understood in this way, generate a normative "standard of virtue" (*Treatise*, p. 591) or "a right or a wrong taste of morals" (*Treatise*, p. 547)?

There are two main objections against sympathy as the basic principle of moral evaluation. (1) Sympathy seems to be biased in its operation: we feel more for those close to us than for strangers; and (2) sympathy cannot easily explain moral beliefs entertained in the absence of sympathy-occasioning causes. To his credit, Hume voices these objections himself but his answers are rather short. To the objection of bias he replies:

> ... 'tis impossible we cou'd ever converse together on any reasonable terms, were each of us to consider characters and persons, only as they appear from his peculiar point of view. In order, therefore, to prevent those continual *contradictions*, and arrive at a more *stable* judgment of things, we fix on some *steady* and *general* points of view... . (*Treatise*, pp. 581-2)

To the objection of absent causes Hume remarks:

> Where a character is, in every respect, fitted to be beneficial to society, the imagination passes easily from the cause to the effect, without considering that there are still some circumstances wanting to render the cause a compleat one. (*Treatise*, p. 585)

It is Brand's central idea to fill in these comments with the principles which operated on belief in Book 1: Sympathy, like imagination, produces conflicting beliefs which call for corrective reasoning. Corrective reasoning, again, must rely on the imagination to complete hypothetical and absent causes. Brand asks "Should the judgments of imagination or understanding be accepted as regulative and prescriptive?" (p. 133), and replies that Hume provides no answer. Hume's theory of moral evaluation, Brand claims, concludes with "the same scepticism" (ibid.) as did his epistemology; two inconsistent principles are at work in human belief formation.

Should we accept this as the promised unifying picture of Hume's philosophy? Though I disagree with Brand on the exact nature of Hume's epistemological scepticism the main obstacle for Brand's interpretation lies, I think, in the proposed analogy between beliefs about the world and moral beliefs.

In what respects is corrective reasoning in epistemology and morals meant to be alike? The man trembling in the iron cage employs *causal* reasoning to relieve his fear. This reasoning, according to Hume, is not *demonstratively* certain but takes into account regularity-based probabilities. Where does moral reason fit into this picture?

Suppose one of the shipwrecked passengers of Hume's example is my daughter. Naturally, I feel inclined to help her first. What contradiction or conflict does Brand think arises from the sympathy-induced propositional endorsement of this action? Is the contradiction like calling one thing red (my daughter) and another green (any other passenger) when both are really green? If it is of this kind the required correction is likely to be a matter of *analytic* or *demonstrative* reasoning from the meaning of moral terms, or from constraints on the moral point of view. Even if Brand should succeed in clarifying the analogy between causal and demonstrative reason, things do not get better but worse. It was Hume's earlier claim to the practical insufficiency of reason which led him to consider explaining morality in terms of sympathy in the first place. There is an obvious danger if Brand introduces a concept of reason to solve difficulties arising from sympathetic explanations of morality.

Brand draws the need for this move almost entirely from the section "Of the origin of the natural virtues and vices" (*Treatise*, Bk. 3, Pt. 3, Sc. 1). This section fills 18 of the 167 pages of Book 3 of the *Treatise* and considers, as the heading indicates, explanations of the natural virtues (e.g. benevolence). It is here that the above mentioned problems of bias and of absent causes are briefly discussed. These problems are for Hume a concern but he believes (*Treatise*, p. 580) that sympathetic explanations are in any case less plausible for the artificial virtues (such as justice, property-rights, the keeping of promises etc.) which occupy him for most of Book 3. The artificial virtues are explained as conventional and not directly from sympathy. Primarily, they are approved of by reflection on the mutual interest that each of us has in some sort of cooperation and individual protection which seems to be a reflection of *practical reason*. Curiously, Brand maintains that this self-restraining kind of reasoning (e.g. *Treatise*, p. 492) is again on the same lines as in the correction of biased sympathy (p. 120).

I don't think there is an easy way to reconcile Hume's claims about reason, but if we are to do so we must realize that Hume's use of "reason" is—deliberately or not—ambiguous. He wavers, and sometimes distinguishes between *demonstrative, causal,* and *practical* senses as indeed does our ordinary concept of reason. Brand tends to run different senses together.

Brand's book may not establish itself as a major interpretation of the *Treatise.* It will, however, be of interest for work on sympathy. It comes with a bibliography and an index.

Baier, unlike Brand, takes note of the shift in the use of "reason". When Hume writes in Book I that "*even after the observation of the frequent and constant conjunction of objects, we have no reason to draw any inference concerning any objects beyond those of which we have had experience*" (*Treatise*, p. 139), Baier takes him *not* to be claiming that we can never have reasons for our causal inferences but that the rationalist's concept of reason as demonstration or deduction cannot do the job. Baier's view is vindicated by sections Bk. 2, Pt. 3, Sc. 3 ("Of the influencing motives of the will") and Bk. 3, Pt. 1, Sc. 1 ("Moral Distinctions not deriv'd from Reason") where Hume argues that reason *both* in its deductive (or demonstrative) and inductive (or causal) version is non-motivational and therefore insufficient to draw moral distinctions.

Again, this does *not* mean that we cannot have moral reasons but that neither demonstrative nor causal reasons are *sufficient* moral reasons. In Baier's reading Hume's final version of reason is something like the "capacity for mutually adjusted intention and agreement" (p. 278). It is here that "men are superior to beasts principally by the superiority of their reason" (*Treatise*, p. 610).

Baier's interpretation of the *Treatise* consists of two connected theses: firstly, that there is no inconsistency in Hume's use of "reason" but that the concept of reason is deliberately enlarged during the course of the investigation; secondly, that the progress from demonstrative to causal and from causal to morally approved practical reason is made twice by the same form of argument, an argument of "virtuous circularity" (p. 217) as Baier calls it, borrowing from Goodman. To appreciate the scope and ingenuity of this interpretation we have to follow it through more closely.

In Book 1 of the *Treatise*, reason is introduced in the rationalist manner as deductive or demonstrative reason. Hume argues that there cannot be a sound deductive inference to the conclusion that every event has some cause (Bk. 1, Pt. 3, Sc. 3), and that explicitly formulated generalisations along with deductive inference to particular events will not do since the underlying inductive principle is itself not evident (Bk. 1, Pt. 3, Sc. 6). "It is reason that demands non-circular justifications" (p. 68); Baier thus sums up the failure of Hume's first, rationalist, version of reason.

Still, we perform causal inferences, and we believe causes and effects to be necessarily connected. Hume offers a lengthy explanation how we can believe two events to be necessarily connected, an explanation which leads up to the already mentioned "Rules by which to judge of causes and effects". Here we are told how best to consult experience in order to arrive at reliable and stable beliefs about the world. Brand, we saw, presents these rules as reason's answer to the principle of imagination which carried Hume's explanation of our idea of necessary connection; but Brand says little about the status of the rules. For Baier their status is determined by the role they play in the explanation of causality.

Hume's first four rules reformulate his account of causality: We are to regard two events *X* and *Y* as causally connected if they are constantly conjoined, contiguous, and *X* is prior to *Y*. Rules (5) to (8) stipulate further conditions (somewhat in the spirit of Mill's inductive methods). Among them figure the following (*Treatise*, p. 174): "where several different objects produce the same effect, it must be by means of some quality, which we discover to be common amongst them" (5), or conversely: "The difference in the effects of two resembling objects proceed from that particular, in which they differ" (6). (The latter rule worked well for the man in the cage.)

Baier shows in some detail how closely Hume followed these rules for successful empirical thinking in his own account of the mental capacities and features of thought which enable us to think of two events as necessarily connected. We could also see it the other way around: "the rules articulate the norms observed in that part of the account leading up to that articulation" (p. 95). I am not entirely clear how to explicate the circularity here involved. Perhaps this is a possibility: Hume does not offer an analysis of causality in terms of constant conjunction; he tells a causal story. Whenever (R) a constant conjunction between two events *X* and *Y* has been observed (and the regularity between *X* and *Y* has been tested by rules (5) to (8), say, conforming examples have been found but no decisive counter-examples) and an *X* is remembered or actually observed, we will come to believe (B) that a *Y* must have occurred or must occur.

To the objection that it is the *justification* of this inference that is in question the reply is that on the meta-causal level we can repeat the same move. Whenever a regularity between mental occurrences (R) and (B) is observed (and tested according to rules (5) to (8)) we will come to believe that (R) *caused* (B), in other words, we will come to believe Hume's explanation. (For a similar argument see Barry Stroud, *Hume*, London and New York: Routledge, 1977, p. 92.)

This may be a way to understand how causal reasoning can be "self-verifying" (p. 91). Baier thinks that Hume's normative re-endorsement of causal reason is due to this feature.

Baier seeks to identify an argument of the same structure in Hume's account of morality. Reason in its deductive (or demonstrative) and inductive (or causal) version fails to account for our moral concerns insofar as they are practical, since reason is motivationally insufficient. Still, we draw moral distinctions, and we are sometimes influenced by what we believe to be our duty. How can this be? Hume explains our moral beliefs as sentiments. They are based on (1) a general disposition to share the feelings of others (i.e. sympathy as a principle of communication), (2) an inclination to prefer those close to us to strangers even if we can share their feelings (i.e. sympathy in the narrow sense of compassion), and (3) the ability to cooperate for mutual advantage. (It was mainly the last part of the explanation that Brand failed to appreciate.) Baier thinks that our moral beliefs as sentiments can be "validated" if they can be shown to be capable of being turned successfully on themselves. At the end of the *Treatise* reason and sentiment no longer stand in opposition. They are one and the same.

How are we to turn sentiments into reasons? How can we validate moral beliefs? Baier, and perhaps Hume, is here less explicit than in the case of causal beliefs. In Baier's reconstruction of Part 3 of Book 1 we arrived at a causal explanation of causality that left causal reasoning in need of no further justification. Can we perform a similar argument for practical reasoning in general, and for morally approved practical reasoning in particular?

What can it mean "to reason practically about practical reason"? Consider action *A*: my drinking a glass of water. One might say, that in thinking about *A* I approve of *A* because I was thirsty. I acknowledge reflexively that my drinking the glass of water was the best *means* to satisfy my desire. Or one might say (as Baier pointed out after seeing a version of this review), that the *desire itself* becomes reflexive: I desire to have and to satisfy my desire, namely to quench my thirst. Moral approval of practical reasoning seems to be modelled on the latter case. Let's say we want to form a *moral* judgment of military virtue. If we are to trust Hume this is not easy to do since the appraisal of military virtue is essentially partisan:

> When our own nation is at war with any other, we detest them under the character of cruel, perfidious, unjust and violent: But always esteem ourselves and allies equitable, moderate, and merciful. If the general of our enemies be successful, 'tis with difficulty we allow him the figure and character of a man. He is a sorcerer: He has a communication with daemons... He is bloody-minded, and takes a pleasure in death and destruction. But if the success be on our side, our commander has all the opposite good qualities, and is a pattern of virtue, as well as of courage and conduct. His treachery we call policy: His cruelty is an evil inseparable from war. (*Treatise*, p. 348)

If our standard becomes reflexive, Baier suggests, it is likely that we retreat from the appraisal of military virtue altogether since we wouldn't want to endorse military virtue in our enemies. This seems too optimistic. It is perfectly possible to approve morally of "excessive courage" (*Treatise*, p. 600); it is also perfectly possible to disapprove of it. As Baier reads it we *ought* to disapprove of it, as we ought to disapprove of the "monkish virtues" of "celibacy, fasting, penance, mortification, self-denial, humility, silence, solitude" (*Second Enquiry*, 270). On the other hand we ought to approve of benevolence and generosity (i.e. natural virtues); we ought to respect property rights and keep promises (i.e. artificial virtues)—to name but a few things from Hume's catalogue of virtues. Can it be that only Hume's catalogue finds reflexive approval?

I am much less confident than Baier that reflexivity is a sufficient "moral test" (p. 216). It may constrain moral thinking but it doesn't seem to *individuate* valid moral beliefs. Hume acknowledges that things cannot be that easy by his distinction between natural and artificial virtues. The distinction names different types of explanations or different modes of reflexive validation, if you like. The artificial virtues can be approved of by all reasonable beings through reflection on the mutual benefits of cooperation; the natural virtues are less cogent and are open to approval by reflection on natural tendencies most of us share. Both approvals are in Hume's view somehow connected. Baier gives an admirable exposition of these complicated matters but fails to say exactly how, as she holds, the natural and the artificial combine into *one* reflexive reason. Instead she settles for a position like:

> Our capacity for judgment outruns our capacity to reduce our judgments to rule. We trust our powers of judgment more than we trust our ability to generalize about what determines our judgment. (p. 281)

I am aware from elsewhere that Baier doubts whether to "seek justification for moral beliefs is helping... to become wiser" (preface to her 1985 collection *Postures of the Mind*, Minneapolis: University of Minnesota Press). Her reading of the *Treatise* tends to express this belief. Baier avoids talk of justification but stresses instead the *possibility* of adopting a reflexive and sustainable, "valid" moral perspective. To a degree I followed her terminology, but one may note that Baier quotes at some critical points predominantly from the *Second Enquiry* where, I would argue, justificatory sharpness takes a back seat to that "warmth in the cause of virtue" that Hutcheson famously missed in the *Treatise*. I think the issue is not yet settled whether a Humean framework of morality as practical, i.e. sentimental, is incompatible with moral explicitness.

Baier's interpretations operate for most parts within Hume's conceptual framework which occasionally makes her book difficult to read. With the focus firmly on the text, Baier is, on the other hand, in a good position to place controversial claims on Hume's conceptual map. Baier proposes convincing solutions to some of the longest running controversies of Humean scholarship. 28 pages of excellent footnotes and a useful index also deserve praise. They provide a comprehensive guide through the maze of recent Hume literature and make up for some lack of orientation in the main body of the text. Baier's book will become a standard.

Where does the discussion of Baier and Brand leave the question of the Old and New Hume? Baier's and Brand's interpretations suggest that Hume was mainly interested in the formation and correction of human beliefs and sentiments. Though some of his explanations of our causal and moral beliefs turn, as we saw, into normative epistemological claims, the evidence remains inconclusive on metaphysical matters such as realism. Hume might be neither Old nor New.

MARTIN KRETSCHMER

Department of Philosophy
University College London
Gower Street
London WC1E 6BT
UK

First Person Plural: Multiple Personality and the Philosophy of Mind, by Stephen E. Braude. London and New York: Routledge, 1991. Pp. xi + 283. $49.95.

In 1987, *The Diagnostic and Statistical Manual of Mental Disorders*, 3rd Edition Revised (DSM-IIIR), classified Multiple Personality Disorder (MPD) as a particular type of dissociative disorder in which there is:

> A. The existence within the person of two or more distinct personalities or personality states (each with its own relatively enduring pattern of perceiving. relating to, and thinking about the environment and self).
>
> B. At least two of these personalities or personality states recurrently take full control of the person's behavior.

Personalities and personality states are understood not to differ in kind, but only in degree of robustness, with personality states being far less complex or extensive than a full-blown personality.

These diagnostic criteria leave some important questions unanswered: What precisely constitutes a personality? How do we "count" personalities? What is it like to suffer from MPD? What is it like to be an alternate personality (i.e. an "alter")? Braude's book, the first full-length discussion by a philosopher, attempts to provide an analysis of MPD in order to shed light upon these questions.

To this effect, he introduces a number of terminological distinctions, which can be summarized as follows:

> State x is *indexical* for a subject S iff S believes x to be his own state.
>
> State x is *autobiographical* for S iff S experiences x as his own state.
>
> S is an *apperceptive centre* iff S is the subject of autobiographical states, most of which are indexical for S.
>
> A and B are *distinct, co-active apperceptive centres* iff A and B have mutually exclusive sets of indexical and autobiographical states operating simultaneously.
>
> State x is *extrareferential* for S iff S assigns x to another subject.

Since alters seem to be the subject of both indexical and autobiographical states, they thus qualify as apperceptive centres. It is also clear that alters regard various states associated with their body as being extrareferential. Combining these observations, we can say that multiples seem to have a plurality of distinct, co-active apperceptive centres associated with one body.

One curious development, yet to be given a satisfactory explanation, is that while historical cases of MPD tended to involve a pair of alters (or occasionally 3 or 4), in recent times the average number has risen to between 6 and 16, with many cases exhibiting more than 100 distinct alters. Absurdly, the "record" exceeds 4500! Clearly, "alters" in these latter uses would be described as "personality states" rather than as "personalities" by the DSM-IIIR classification, "since their functions tend to be highly circumscribed, and because they do not exhibit the more extensive range of traits and dispositions found in more personality-like alters" (Braude, p. 41). In fact, we might say that our warrant for ascribing full-blown personality status varies inversely with the number of other alters